Causal Dependence Plots

Joshua R. Loftus
Department of Statistics
London School of Economics
London, England, UK
j.r.loftus@lse.ac.uk

Lucius E. J. Bynum
Center for Data Science
New York University
New York, NY, USA
lucius@nyu.edu

Sakina Hansen
Department of Statistics
London School of Economics
London, England, UK
s.a.hansen1@lse.ac.uk

Abstract

Explaining artificial intelligence or machine learning models is increasingly important. To use such data-driven systems wisely we must understand how they interact with the world, including how they depend causally on data inputs. In this work we develop Causal Dependence Plots (CDPs) to visualize how one variable—an outcome—depends on changes in another variable—a predictor—*along with any consequent causal changes in other predictor variables*. Crucially, CDPs differ from standard methods based on holding other predictors constant or assuming they are independent. CDPs make use of an auxiliary causal model because causal conclusions require causal assumptions. With simulations and real data experiments, we show CDPs can be combined in a modular way with methods for causal learning or sensitivity analysis. Since people often think causally about input-output dependence, CDPs can be powerful tools in the xAI or interpretable machine learning toolkit and contribute to applications like scientific machine learning and algorithmic fairness.

1 Introduction

This paper develops Causal Dependence Plots (CDPs) to visualize causal relationships between predictor variables and an outcome variable. The idea is general, but we are motivated by explaining or interpreting AI or machine learning models [6, 13, 14, 27]. For simplicity we consider supervised learning where a set of features is used to predict an outcome, i.e. regression or classification. We also focus on the model-agnostic or "black-box" explanation setting, where the interpreter can query the model but not access its internal structure. Interpretation methods in this setting are more broadly applicable for distributed research, but are also functionally limited to observing how the model responds to variation in the inputs. Our general approach has extensions beyond this initial application.

Visualizations and simple explanations that focus on one input variable at a time can be powerful tools for human understanding. Two popular visualization methods, the Partial Dependence Plot (PDP) of Friedman [10] and Individual Conditional Expectation (ICE) plot from Goldstein et al. [11], show how model output depends on one input variable. However, just as with the interpretation of linear regression model coefficients, the relationships revealed by focusing on one predictor at a time can be misleading. When varying one input variable, *we must make some choice about what values to use for the other inputs*. PDPs treat other predictors as independent of the one being plotted, so they can correctly capture the model's dependence on each variable if predictors are independent and the model is additive [15]. In general, when explaining the black-box's dependence on one input, our choice of how to handle other model inputs may break or respect the existing statistical or causal dependencies between predictors. This leads us to the following:

Problem statement. If there are causal relationships between predictors in the real world, but our visualization, interpretation, or explanation method does not respect them, then the resulting model

Preprint. Under review.

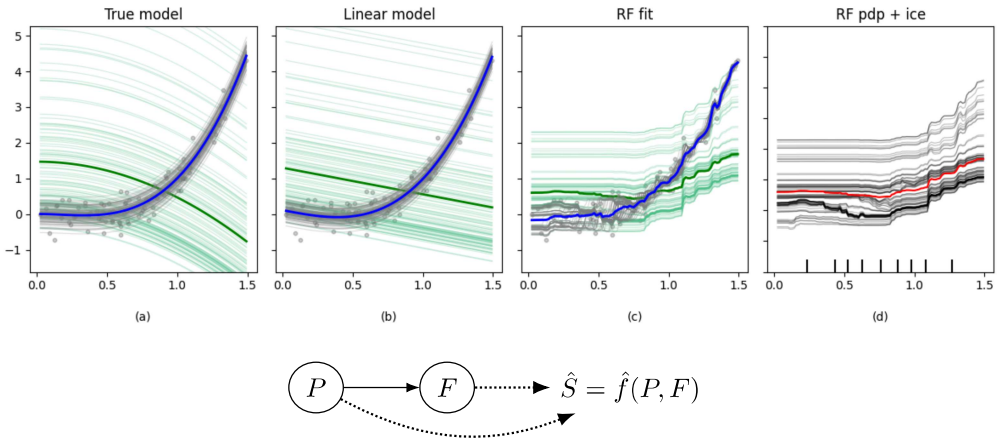

Figure 1: Motivating example. Causal Dependence Plots (top) and structural graph of the Explanatory Causal Model (bottom) for the motivating example. In the top row, panel (a) shows the relationships of the ECM. Total Dependence is shown in blue and Natural Direct Dependence in green. Counterfactual curves for individual points are shown as thin, light lines, with aggregates displayed as thick, dark lines. Panels (b-c) show our model explanation plots for a linear model and random forest model, respectively. Panel (d) shows a standard Partial Dependence Plot (red) with Individual Conditional Expectation curves, for comparison to our NDDP (green).

explanation may be irrelevant or misleading for real world purposes [28, 41]. For example, such explanations could lead to incorrect decisions about regulating or aligning algorithmic systems, sub-optimal allocations of resources based on model predictions, a breakdown between human feedback and reinforcement learning systems, or other forms of harm. In the context of scientific machine learning—where explanations can be used to generate hypotheses for follow-up investigation—a flawed interpretation may direct us toward spurious hypotheses. For these reasons, we may care about the causal validity of model explanations.

Our proposal. At a high level, we use an auxiliary Explanatory Causal Model (ECM) to interpret or explain a given machine learning model. For each input predictor that we wish to explain, we use the ECM to determine how other inputs vary when that predictor is manipulated, rather than treating them as independent or fixed. We call the resulting plots Causal Dependence Plots or CDPs.

Causal models can be designed based on the desired explanation, specified with prior domain knowledge, and/or potentially learned and estimated from data.

Motivating example. Before turning to the full details, we illustrate the idea with a simple example. Consider a mediation model for parental income P, school funding F, and graduates' average starting salary S, with structural equations $P \sim \mathcal{U}[0, 1.5]$; $F = 2P^3 + \mathcal{N}(0, 0.2^2)$; and $S = F - P^2 + \mathcal{N}(0, 0.2^2)$. The corresponding directed acyclic graph (DAG) is shown in the bottom row of Figure 1, with data plotted in the left panel of the top row, and the remaining panels show visual explanations of supervised models that predict $\hat{S} = \hat{f}(P, F)$.

Several important takeaways stand out from this display. First, the differences between blue and green curves show *there can be qualitative differences between direct (or partial) dependence and total dependence*, a fact which is highly consequential when we consider how causal interventions may change outcomes. Second, *explanations of models can be qualitatively different from the underlying causal relationships*. For example, even a flexible model like the random forest in panel (c) shows a direct dependence of $\hat{S}$ on P that is increasing when the direct dependence of S on P in the true model is decreasing. As another example, panel (b) shows that the total dependence of a linear model on a predictor can be non-linear because the mediator F depends non-linearly on P. Finally, we see that *our framework of using causal models to produce explanation plots includes, as special cases, some existing model explanation plots like ICE and PDPs*. We revisit this point later.

2 Methodology

We give background notation and definitions in Sections 2.1-2.2 and our methods in Sections 2.4-2.6.

2.1 Explaining Supervised Learning Models

For concreteness, we consider the supervised learning setting with a set of predictor features $\mathbf{X}$ and outcome variable Y. We wish to explain or interpret a predictive model represented by a function $\hat{f}$ which is typically estimated or learned using empirical risk minimization (ERM)

$$\hat{f} = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{n} \ell\left(h(\mathbf{x}_i), y_i\right)$$

with some loss function ℓ, pre-specified function class $\mathcal{H}$, and an independent and identically distributed training sample $\{(y_i, \mathbf{x}_i) : i = 1, \ldots, n\}$ with feature vectors $\mathbf{x}_i^T \in \mathbb{R}^p$. In 2.5 we focus on a specific interpretive task known as mediation analysis, again for concreteness and because it is a highly applicable example. In that section we partition the predictor variables into subsets so that X and M both notate predictors, M being a mediator.

Model-agnostic explanations. Sometimes, for practical reasons, an explanation method does not have access to the internal mathematical structure of $\hat{f}$. In this model-agnostic setting we can generate explanations based on input-output dependence by providing *synthetic inputs*

$$\tilde{\mathbf{X}} \mapsto \hat{f}\left(\tilde{\mathbf{X}}\right),$$

recording the associated predictions, and then summarizing these in some way. We denote a model-agnostic explanation generated this way as $\mathcal{E}(\hat{f}; \tilde{\mathbf{X}})$ or $\mathcal{E}(\hat{f})$ for shorthand, but note that all such explanations, including our proposed method, depend on both $\hat{f}$ and the synthetic input. We often target one feature at a time for ease of interpretation and creating low-dimensional visualizations. We denote an explanation targeting feature j as $\mathcal{E}_j(\hat{f})$, and in this case the synthetic input is

$$\tilde{\mathbf{x}}_j := (x_1, \ldots, \tilde{x}_j, \ldots, x_p) \mapsto \hat{f}\left(\tilde{\mathbf{x}}_j\right).$$

If feature j is numeric, then it is typically varied along a grid in its domain. In most existing model-agnostic explanation methods, the values for other features are held fixed at observed values in a dataset used to generate the explanation. Note that this explanatory dataset may not be the same as the training data for $\hat{f}$. We emphasize this by notating the data used to generate an explanation as $\{(y_i', \mathbf{x}_i') : i = 1, \ldots, m\}$, where y_i' may not be supplied or required depending on the type of explanation. Bar graphs can be used when the explanatory feature is categorical.

Definition 2.1 (Individual Conditional Expectation (ICE) Plot)**.** We obtain a separate curve for each observation in the explanatory dataset x_i', $i = 1, \ldots, m$, by plotting the map

$$\tilde{x}_j \mapsto \hat{f}\left(\tilde{\mathbf{x}}_{ij}'\right), \text{ where } \tilde{\mathbf{x}}_{ij}' := (x_{i1}', \ldots, \tilde{x}_j, \ldots, x_{ip}').$$

An ICE plot for feature j displays all m of these curves.

Definition 2.2 (Partial Dependence Plot (PDP))**.** The PDP for feature j can be obtained from the ICE plot for x_j by computing the empirical mean over the explanatory data at each point in the grid, that is $\tilde{x}_j \mapsto \frac{1}{m}\sum_{i=1}^{m} \hat{f}\left(\tilde{\mathbf{x}}_{ij}'\right)$.

There are a variety of other model-agnostic explanation methods, see for example [27]. But since our current proposal is a visualization, these are the main alternative methods for comparison.

Fundamental problem of univariate explanations. To create a model-agnostic explanation of model dependence on a single feature, like a plot with x_j on the horizontal axis and $\hat{f}$ on the vertical axis, *we must decide what to do with the other features* when synthetically varying $\tilde{x}_j$. Nearly all existing explanation methods use the same approach as the PDP and ICE plots: they *hold other features fixed* at values in an auxiliary, explanatory dataset. This may be unrealistic if, for example, other features depend on x_j causally. And it may not be mathematically defined or allow any interpretation if features are mutually constitutive, e.g. population, GDP, and GDP per capita.

2.2 Structural Causal Models

Our notational conventions and definitions below are influenced by [5, 32, 33]. Letting $\mathbf{U}$ be a set of exogeneous noise variables, $\mathbf{V}$ a set of $p = |\mathbf{V}|$ observable variables, and $\mathbf{G}$ a set of functions such that for each $j \in 1, \ldots, p$ we have $V_j = g_j(\mathbf{PA}_j, U_j)$, where $\mathbf{PA}_j \subseteq \mathbf{V}$ and $U_j \subseteq \mathbf{U}$ are the observable and exogeneous parents, respectively, of variable V_j. Let the directed acyclic graph (DAG) $\mathcal{G}$ have vertices given by variables and, for each $V_j \in \mathbf{V}$ and each of the parent variables in $\mathbf{PA}_j$ and U_j, a directed edge oriented from the parents to V_j. This graph can be useful for explanations by showing visually which variables are inputs in each function in $\mathbf{G}$.

Definition 2.3 (Structural Causal Model (SCM))**.** A (probabilistic) SCM $\mathcal{M}$ is a tuple $\langle \mathbf{U}, \mathbf{V}, \mathbf{G}, P_{\mathbf{U}} \rangle$ where $P_{\mathbf{U}}$ is the joint distribution of the exogeneous variables. This distribution and the functions $\mathbf{G}$ determine the joint distribution $P^{\mathcal{M}}$ over all the variables in $\mathcal{M}$. Finally, causality in this model is represented by additional assumptions that $\mathcal{M}$ admits the modeling of interventions and/or counterfactuals as defined below.

Definition 2.4 (Interventions)**.** For the SCM $\mathcal{M}$, an intervention I produces a modified SCM denoted $\mathcal{M}^{\text{do}(I)}$ which may have different structural equations $\mathbf{G}^I$. Correspondingly, some variables may have different parent sets, so the DAG representation $\mathcal{G}^{\text{do}(I)}$ may also change. We denote the new, interventional distribution as $P^{\mathcal{M};\text{do}(I)}$. A simple class of interventions involves intervening on one variable, e.g.

$$I = \text{do}\left(V_j := \tilde{g}(\tilde{\mathbf{PA}}_j, \tilde{U}_j)\right),$$

which changes how V_j and all variables on directed paths from V_j in $\mathcal{G}$ are generated. An even simpler sub-class of these are the atomic interventions setting one variable V_j to one constant value v, which we denote $I_{j,v} := \text{do}(V_j = v)$. Note that in this case V_j has no parents in the graph $\mathcal{G}^{\text{do}(I)}$; the source of the intervention itself is outside the world of the model.

Interventions are useful for modeling changes to a data generating process (DGP), for example, experiments that control a particular variable to see how its value changes other variables, or policy changes aimed at altering or removing existing causal relationships. In addition to generating new observations as a DGP, an SCM can also be used to model counterfactual values for observations that have already been determined. A counterfactual distribution is an interventional distribution defined over a specific dataset with information or constraints given by some of the observed values in that data, as we now describe. Here we slightly abuse notation by letting boldface represent the dataset, e.g., $\mathbf{V}$ are the observed variables for all observations in a previously generated dataset.

Definition 2.5 (Counterfactuals)**.** For variable V_j with observed values of its parents $\mathbf{PA}_j = v$, we may hold some or all of v fixed and vary $U_j := u$, passing these through $g_j(v, u)$ (or $\tilde{g}_j(\tilde{v}, u)$ if we also do an intervention that changes any of the values in $\mathbf{PA}_j$). The counterfactuals $V_j(\tilde{v}, u)$ are values V_j would have taken if any of its observed and/or exogeneous parents had taken the different values $(\tilde{v}, u)$. With intervention I, to define the counterfactual distribution $P^{\mathcal{M}|\mathbf{V}=\mathbf{v};\text{do}(I)}$, we use the posterior or conditional (depending on our probability model approach) distribution $P_{\mathbf{U}|\mathbf{V}=\mathbf{v}}$ to model uncertainty about $\mathbf{U}$ while computing counterfactual values of any variables for a previously generated observation in the modified SCM $\mathcal{M}^{\text{do}(I)}$.

2.3 Univariate Causal Explanations

Our proposed solution to the fundamental problem highlighted for univariate explanations is to use an auxiliary ECM $\mathcal{M}_j$ and let this causal model determine how other features vary as functions of x_j. We denote these explanations as $\mathcal{E}_j(\hat{f}; \mathcal{M}_j)$ or $\mathcal{E}_j(\hat{f})$ if the context is clear. In the deterministic or noiseless case, suppose we know functions g_{kj} such that $x_k = g_{kj}(x_j)$, with g_{jj} the identity. In this case the model $\mathcal{M}_j$ tells us

$$x_j \mapsto g(x_j) =: (g_{1j}(x_j), \ldots, x_j, \ldots, g_{pj}(x_j))$$

is a curve in $\mathbb{R}^p$ parameterized by x_j, and we generate the explanation $\mathcal{E}_j(\hat{f})$ using

$$\tilde{x}_j \mapsto \hat{f}\left(g(\tilde{x}_j)\right).$$

Next, a few more definitions will let us extend this strategy to more general, non-deterministic causal models. We propose generating various kinds of causal explanations of a supervised learning model $\hat{f}$ (potentially a black-box) using an auxiliary ECM $\mathcal{M}_{\mathbf{X}}$ that captures causal relationships among the predictor variables. With $\mathcal{G}_{\mathbf{X}}$ the associated DAG, we represent this graphically in Figure 2, where the arrow from the subgraph $\mathcal{G}_{\mathbf{X}}$ to the explanation $\mathcal{E}(\hat{f})$ is dotted to distinguish it from arrows among the features. Different explanations correspond to various causal operations, such as interventions, performed in $\mathcal{M}_{\mathbf{X}}$.

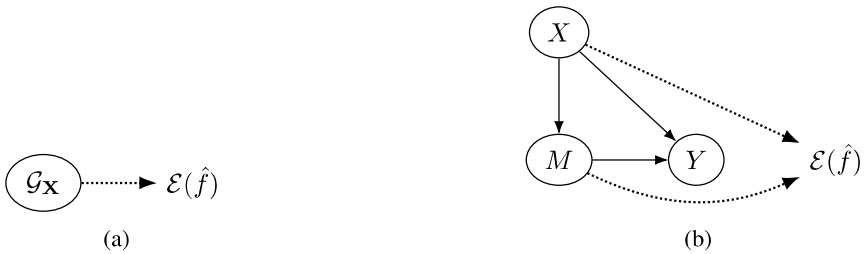

Figure 2: A structural causal model for predictors $\mathcal{G}_{\mathbf{X}}$ is used to produce an explanation $\mathcal{E}(\hat{f})$ of the predictive model $\hat{f}$. In the mediation example (b), predictor X causes Y directly and also through mediator M. Solid arrows represent causal relationships in the data generation process, and dotted arrows show the formal dependence of the model explanation on predictors.

Remark 2.6. Note that if the desired causal explanation uses counterfactuals, then we likely obtain the observed values from an auxiliary explanatory dataset. But since an SCM can generate data, we may also use it to generate the initial observed values and then re-use these when computing counterfactuals for the explanation. That is, we may generate the auxiliary explanatory dataset $\mathbf{V}$ from $P^{\mathcal{M}}$ and then, with an intervention I, generate counterfactuals from $P^{\mathcal{M}|\mathbf{V}=\mathbf{v};\text{do}(I)}$.

The generality and flexibility of SCMs allow us to pose many different interpretive questions, and since SCMs can generate synthetic data, we can use them to compute many different kinds of explanations. In the next two sections, we focus on using plots as explanations and describe several canonical types of interpretive questions.

2.4 Causal Dependence Plots

One key motivation for this work is to define various causal analogues of the PDP and ICE plots, which we are now prepared to do. For the following definitions, we assume predictor variables $\mathbf{X} \in \Omega_{\mathbf{X}}$, an outcome of interest $Y \in \Omega_Y$, and a black-box function $\hat{f}(x) : \Omega_{\mathbf{X}} \to \Omega_Y$ with outputs that we may also denote $\hat{Y}$. A structural causal model $\mathcal{M}_{\mathbf{X}}$ is either assumed or learned from data. Importantly, $\mathcal{M}_{\mathbf{X}}$ specifies the causal relationships *only for the predictors* $\mathbf{X}$ and need not involve the outcome Y.

Individual counterfactuals and expected effects. We use the shorthand $\hat{f}(P^{\mathcal{M}})$, where $\hat{f}$ takes a distribution $P^{\mathcal{M}}$ as its argument, to denote using data from that distribution as the input to the black-box function $\hat{f}$. For each of our plots, we show both a set of individual counterfactual curves $\hat{f}(P^{\mathcal{M}|\mathbf{V}=\mathbf{v};\text{do}(I)})$ and their empirical average over the explanatory dataset

$$\hat{\mathbb{E}}\left[\hat{f}(P^{\mathcal{M}|\mathbf{V}=\mathbf{v};\text{do}(I)})\right]. \tag{1}$$

For each type of causal explanation with a given *Named Effect*, we define the *Individual Counterfactual Name Effect* curves as the set of counterfactual curves $\hat{f}(P^{\mathcal{M}})$ for each individual in the explanatory dataset, the *Name Effect Function* as their expectation, and the *Name Dependence Plot* as a plot displaying all of these curves (or rather, an empirical estimate in the case of the Effect Function).

Before defining different types of CDPs, we first introduce a useful abstraction. Generating causal explanations involves performing abduction, action, and prediction with a structural causal model that is augmented to include the predictions we wish to explain.

Definition 2.7 (Explanatory Causal Model (ECM))**.** An ECM $\mathcal{M}'$ augments the original SCM $\mathcal{M}_{\mathbf{X}}$ by including the predicted outcome $\hat{Y}$ as an additional variable with $\hat{f}$ as its structural equation. We can then, for example, compute $\hat{f}(P^{\mathcal{M}_{\mathbf{X}}|\mathbf{V}=\mathbf{v};\text{do}(I)})$ using the ECM as $P^{\mathcal{M}'|\mathbf{V}=\mathbf{v};\text{do}(I)}_{\hat{Y}}$. We describe the construction of an ECM in Algorithm 1.

We use this process in Algorithms 2 through 5 to compute each of the effects we now describe. We begin with perhaps the most straightforward and important causal effect.

Definition 2.8 (Total Dependence Plot (TDP))**.** For an intervention I, the Individual Counterfactual Total Effect (ICTE) curves

$$\mathsf{TE}(I) = \hat{f}(P^{\mathcal{M}_{\mathbf{X}}|\mathbf{X}=\mathbf{x};\text{do}(I)}) \tag{2}$$

show the total effect of intervention I on black-box output for each individual observation in the explanatory dataset. The empirical average of these over the explanatory data is a (Monte Carlo) estimate of the Total Effect Function (TEF), and a plot showing the ICTE and TEF is a Total Dependence Plot (TDP). We calculate the TDP following Algorithm 2.

Remark 2.9. In the remaining definitions, we give notation only for the individual counterfactual curves and leave the other objects implicitly defined.

We often wish to decompose how much of the total effect of X on $\hat{Y}$ (or Y) is attributable to different possible pathways between the variables. This involves understanding several causal quantities in addition to the total effect described above, which we now define.

Definition 2.10 (Partially Controlled Dependence Plot (PCDP))**.** Consider intervention I affecting some subset of variables in DAG $\mathcal{G}_{\mathbf{X}}$ and atomic intervention C that holds constant any other subset of variables not intervened upon by I. The Individual Counterfactual Partially Controlled Effect curves

$$\mathsf{PCE}(I, C) = \hat{f}(P^{\mathcal{M}_{\mathbf{X}}|\mathbf{X}=\mathbf{x};\text{do}(I,C)}) \tag{3}$$

represent the effect of intervention I on black-box output $\hat{Y}$ while other variables are set (via intervention) to constant values. We compute the PCDP via Algorithm 3.

Definition 2.11 (Natural Direct Dependence Plot (NDDP))**.** Consider atomic intervention I and a corresponding intervention J that intervenes on all children of any nodes that are intervened upon by intervention I and sets them to their observed values in dataset $\mathbf{x}$. For example, if $I = \text{do}(A = a, B = b)$, then intervention J will set all children of the variables A and B to their observed values in $\mathbf{x}$. We then define the Individual Counterfactual Natural Direct Effect curves as

$$\mathsf{NDE}(I) = \hat{f}(P^{\mathcal{M}_{\mathbf{X}}|\mathbf{X}=\mathbf{x};\text{do}(I,J)}). \tag{4}$$

This quantity represents the effect of intervention I on black-box output $\hat{Y}$ while all variables not intervened upon are set to their 'natural,' i.e., pre-intervention values in the explanatory dataset. Algorithm 4 demonstrates how to compute the NDDP.

From this construction of NDDP, we see by comparing it to Definition 2.2 that it is equivalent to the PDP, confirming what we observed in Figure 1.

Proposition 2.12. *Partial dependence plots show natural direct effects. When generating plots using explanatory data $D \sim \mathcal{M}_{\mathbf{X}}$, the ICE plot curves and Individual Counterfactual Natural Dependence curves—and hence also the PDP and NDDP plots—are identical.*

Definition 2.13 (Natural Indirect Dependence Plot (NIDP))**.** Consider atomic intervention I and a corresponding intervention K that removes from DAG $\mathcal{G}_{\mathbf{X}}$ all outgoing edges from any of the nodes intervened upon by intervention I and sets those nodes to their observed values in the explanatory dataset. For example, if $I = \text{do}(A = a, B = b)$, then intervention K will remove all outgoing edges from A and B and set A and B to their original observed values. We then define Individual Counterfactual Natural Indirect Effect curves

$$\mathsf{NIE}(I) = \hat{f}(P^{\mathcal{M}^{\text{do}(I)}_{\mathbf{X}}|\mathbf{X}=\mathbf{x};\text{do}(K)}). \tag{5}$$

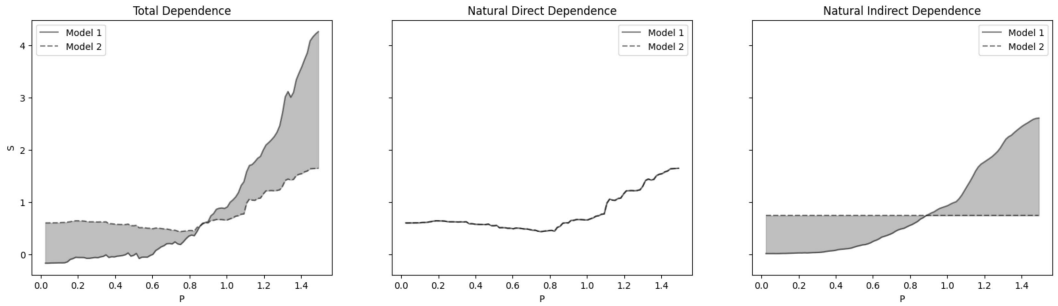

Figure 3: TDP, NDDP, and NIDP for the salary example that incorporate causal uncertainty, now visualizing a range of possible effect functions induced by two competing candidate causal models — one with mediation and one without.

Notice that intervention I is performed before intervention K. This quantity represents the effect of intervention I on black-box output $\hat{Y}$ that is due only to any indirect pathways to $\hat{Y}$. We compute the NIDP following Algorithm 5. The difference between two values of this function can be used to express the natural indirect effect as a special case.

The Supplementary Material includes full descriptions of algorithms for computing all of the above plots.

2.5 Mediation Analysis

Many applications involve a causal structure we refer to as a mediation triangle, with an example shown in Figure 2b. In mediation analysis, we often wish to decompose how much of the total effect of X on Y is attributable to the pathway through M and how much of it is direct. Our above definitions allow us to visualize frequently studied quantities of interest in this setting. For example, the difference between two values of the PCDP can be used to express the controlled direct effect as a special case — specifically, with interventions I, C defined as $I = \text{do}(X = x)$ and $C = \text{do}(M = m)$ in the mediation triangle. Although mediation analysis motivates CDPs and helps build intuition, we emphasize that the definitions in Section 2.4 can be used *with any structural causal model*. See Section 3 for other, more complex examples.

2.6 Incorporating Uncertainty in Causal Dependence

There are various ways to incorporate uncertainty about the true causal model into CDPs. We now explore a natural first extension of the CDP that shows a *range of possible effect functions* induced by a *set of auxiliary ECMs*. The set of ECMs could be pre-specified or, for example, a Markov equivalence class of DAGs output by a causal structure learning algorithm. Returning to our motivating example from Section 1, we might question whether parental income P impacts school funding F, considering instead an SCM without mediation: $P \rightarrow S \leftarrow F$. Figure 3 shows a range of possible effect functions interpolating between this ECM without the indirect effect and the original ECM in Section 1, for each of the TDP, NDDP, and NIDP. In this we have assumed the same structural equations for the edges that are common to both models. These plots show a range for how predictions might depend on one predictor P when we are unsure how the other predictor depends on P. In the Supplementary Material we show an example with real data where we use candidate ECMs discovered by the PC algorithm.

In this example, we have shown how incorporating multiple causal models into CDPs allows us to directly visualize our uncertainty about the underlying causal model and its impact on the effect we expect a predictor X to have on black-box output $\hat{Y}$. More broadly, this process allows us to characterize how X will impact Y or $\hat{Y}$ under multiple conditions, enabling a versatile sensitivity analysis.

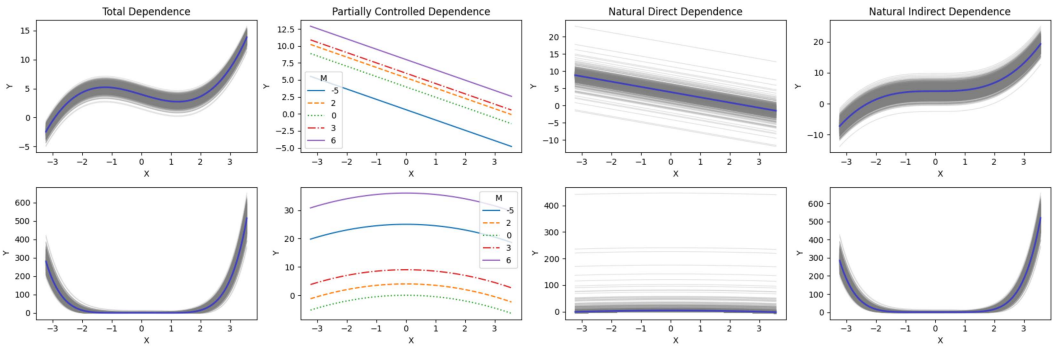

Figure 4: CDPs for the simulation example in Section 3, shown for both a 'correct' black-box model (bottom row) and 'incorrect' black-box model (top row).

3 Experiments

We now illustrate several kinds of causal explanations our method can produce.

Simulations. Consider the non-linear mediation example, governed by the following DGP.

$$X \sim \mathcal{N}(0,1), \quad M = \frac{1}{2}X^3 + \mathcal{N}(0,1), \quad Y = M^2 - \frac{1}{2}X^2 + \mathcal{N}(0,1).$$

We use this DGP to fit two different black-box models: one model that assumes the correct functional form (i.e., the relationship for Y shown in the DGP above), and an 'incorrect' model that predicts Y via linear regression. We use the Python `DoWhy` package [3, 40] in our experiments to sample counterfactual data in the construction of our plots. Figure 4 shows the CDPs for each of these models using the black-box training data as the explanatory data. We can glean a couple insights from Figure 4. First, the CDPs are all sensitive to whether the functional form assumptions of the black-box model fit the ground truth data generating process. The second is that the different effects on the outcome Y that we may want to investigate will show up visually across the different plots. For example, the TDP for the incorrect model shows dependence on X that looks cubic, while the true relationship is quadratic and sextuplic. By looking at the NIDP and NDDP, we can see that for the incorrect model the cubic relationship is due to an indirect effect through mediator M. The PCDP for the correct vs. incorrect model shows us directly how the quadratic term — the controlled direct effect on Y — is either present or not in the black-box model.

Real data with causal discovery. The Breast Cancer Wisconsin (Original) dataset [44] is a publicly available dataset often used to test algorithms on medical data. The dataset contains 9 ordinal variables, which represent attributes of the cells within a breast mass: Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli and Mitoses. The outcome variable is the class of the breast tumor, benign or malignant.

We use a causal structural learning algorithm, specifically the PC algorithm [42] implemented in Julia `CausalInference`, to learn a DAG for this dataset, on a smaller subset of predictor variables for simplicity. Figure 5 shows the resulting DAG and CDPs for a random forest model to classify the Class variable.

This shows CDPs can be combined with other causal methods like structural learning algorithms. The PC algorithm output had an undirected edge between Cell Size and Cell Shape, so we investigate sensitivity to the different graph structures consistent with this output in the Supplementary Material.

4 Applications, Extensions, Related Work

Explanations under covariate shift: Often a model that has already been trained will be used for predictions on data that may not follow the training DGP. If knowledge about causal relationships in

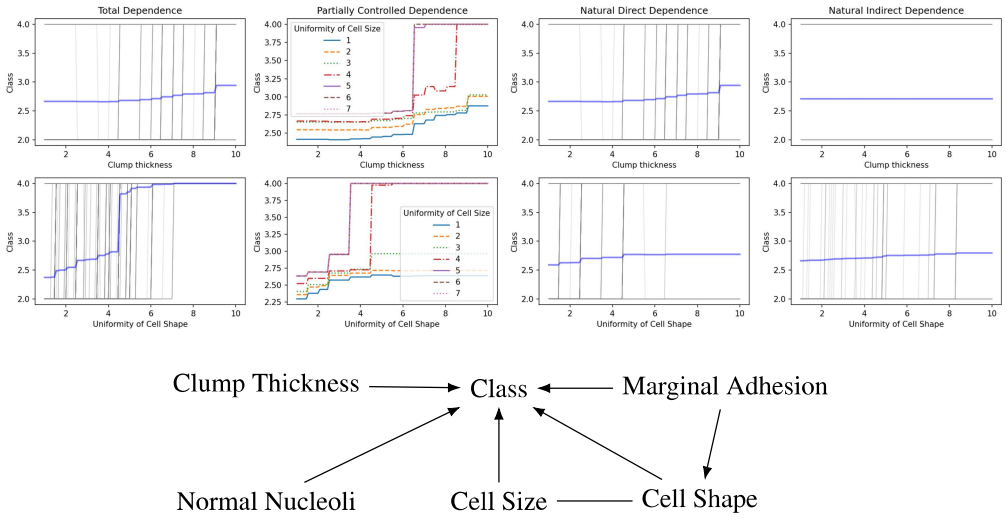

Figure 5: Breast cancer data example. CDPs for a random forest classifier and predictors Clump Thickness (first row) and Uniformity of Cell Shape (second row). Structural graph $\mathcal{G}_B$ for the ECM learned by the PC algorithm (last row). The outcome Class is binary: 2 for benign, 4 for malignant.

the shifted covariate distribution is available, we can leverage that to choose an ECM, and CDPs can visualize how the model will behave under covariate shift.

Causal semi-supervised learning: Given knowledge of causal structure among predictors and a supervised learning model to predict an outcome, our method can be used to attempt inference of causal relationships from the predictors to the outcome, similar in spirit to [39] and [49].

Auditing for fairness or other desiderata: Previous work has applied causal methodology to fairness [4, 8, 20–22, 25, 26, 29, 36, 46, 48], recourse [19, 34, 45], and other desiderata. CDPs can be used to probe a black-box for such properties.

Scientific theory development: Large and complex machine learning models may be fit to massive datasets where underlying structure is largely unknown. In such settings, relatively simple ECMs can be used to formulate hypotheses relating multiple predictors and plot various causal dependencies as a way of generating new hypotheses or checking the assumed ones for plausibility.

Uncertainty and sensitivity analysis: Future work can develop methods for visualizing uncertainty. For example, sensitivity analysis based on conformal prediction [7, 16, 24, 47]. If there is uncertainty between several candidate ECMs, our method in Section 2.6 or one analogous to that in [36] could produce composite CDPs.

Related work: Recent work motivated by recourse [18] uses contrastive or counterfactual explanations [43]. Some of this work is not based on causality despite using the term "counterfactual," but some does focus on causal dependence [38]. Blöbaum and Shimizu [2] produce causal explanations by identifying the predictor with the largest total effect, which is most applicable when assuming linearity. Zhao and Hastie [49] investigated causal interpretations of PDP. That paper aimed to use such plots for causal inference about the underlying DGP rather than as explanations of the black-box, and showed that when the DGP satisfies the backdoor criterion [30] then a PDP visualizes the total effect (TE) of a predictor. Cox Jr [9] observed an association between partial dependence plots and NDE, an equivalence we formally establish in Proposition 2.12, to our knowledge the first such result. Lazzari et al. [23] proposed weighting observations when computing PDPs, which could potentially be used for confounder adjustment. Our unified framework shows how disparate model interpretation plots originally proposed without causal motivation or justification can be related to causal interpretations. Aside from the coincidental cases of PDPs and M-plots [27] with NDE and TE, we are not aware of any previous work providing explanations or interpretive plots for the other kinds of causal explanations that fit in our framework.

5 Discussion

Limitations. Causal modeling in general involves some limitations that we do not repeat here, but see for example [12]. Model-agnostic explanation methods are also subject to limitations [1, 27], a few of which we will highlight because we believe they are important to keep in mind when using our method.

Mismatch between the black-box and the true DGP: if the predictive model fails to fit the DGP, then practically any model explanation will also fail if our interpretive goal is to learn about the DGP [49]. This issue is particularly troubling in the model-agnostic setting where we cannot conduct model diagnostics and probe the internals of the black-box.

Mismatch between the explanatory SCM and true/target DGP: CDPs may be misleading if the true DGP differs in important ways from the ECM. However, standard PDPs and similar explanation methods also require auxiliary explanatory data, and that data may also differ from the target DGP. So this is not much of an additional limitation specific to our method.

Availability of the ECM: if we consider models as tools and are not concerned about whether there is a "true" causal model for the DGP, we still need to choose which tools to use when producing an explanation. In this sense full knowledge of an SCM can be a strong assumption. However, in Sections 2.6 and 4 we discussed some ways this can be improved. In general, *if a causal explanation is desired or necessary, then we cannot avoid making causal assumptions*.

Conclusion. In this paper we proposed Causal Dependence Plots, a method that uses a given (potentially learned) explanatory causal model to create various plots with causal interpretations. This allows us to use the powerful language of structural causal models to pose and answer a variety of causally meaningful questions. Our framework unites previously disparate, non-causally motivated interpretive tools like partial dependence plots, and reveals some new kinds of causal interpretations we have not seen previously explored in the literature. Additional future work in this direction could explore other relatively small canonical causal structures for useful applications, or interface with other kinds of models, for example extending to non-tabular data by applying causal representation learning. Relating explanation methods to Pearl's ladder of causation [31], most previous interpretable machine learning and explainable AI methods—like PDPs—concern associations and hence are confined to the first rung of the ladder. With CDPs we ascend the ladder, creating model interpretations intended to change the world. While interpretability provided the initial motivation for CDPs, we believe plotting such causal relationships will be useful in other settings as well.

References

[1] T Altmann, J Bodensteiner, C Dankers, T Dassen, N Fritz, S Gruber, et al. Limitations of interpretable machine learning methods. 2020.

[2] Patrick Blöbaum and Shohei Shimizu. Estimation of interventional effects of features on prediction. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2017.

[3] 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. *arXiv preprint arXiv:2206.06821*, 2022.

[4] Lucius Bynum, Joshua Loftus, and Julia Stoyanovich. Disaggregated Interventions to Reduce Inequality. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–13. Association for Computing Machinery, New York, NY, USA, October 2021. ISBN 978-1-4503-8553-4. URL `https://doi.org/10.1145/3465416.3483286`.

[5] Lucius Bynum, Joshua Loftus, and Julia Stoyanovich. Counterfactuals for the future. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[6] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, 8(8):832, August 2019. ISSN 2079-9292.

doi: 10.3390/electronics8080832. URL https://www.mdpi.com/2079-9292/8/8/832. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

[7] Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*, 116(536):1849–1864, 2021.

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

[9] Louis Anthony Cox Jr. Modernizing the bradford hill criteria for assessing causal relationships in observational data. *Critical reviews in toxicology*, 48(8):682–712, 2018.

[10] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[11] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

[12] Sander Greenland and Mohammad Ali Mansournia. Limitations of individual causal models, causal graphs, and ignorability assumptions, as illustrated by random confounding and design unfaithfulness. *European journal of epidemiology*, 30:1101–1110, 2015.

[13] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51(5):93:1–93:42, August 2018. ISSN 0360-0300. doi: 10.1145/3236009. URL https://doi.org/10.1145/3236009.

[14] David Gunning, Eric Vorm, Jennifer Yunyan Wang, and Matt Turek. DARPA's explainable AI (XAI) program: A retrospective. *Applied AI Letters*, 2(4):e61, 2021. ISSN 2689-5595. doi: 10.1002/ail2.61. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ail2.61. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ail2.61.

[15] Trevor Hastie and Robert Tibshirani. Generalized Additive Models. *Statistical Science*, 1 (3):297 – 310, 1986. doi: 10.1214/ss/1177013604. URL https://doi.org/10.1214/ss/1177013604.

[16] Ying Jin, Zhimei Ren, and Emmanuel J Candès. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences*, 120(6):e2214889120, 2023.

[17] Diviyan Kalainathan, Olivier Goudet, and Ritik Dutta. Causal discovery toolbox: Uncovering causal relationships in python. *J. Mach. Learn. Res.*, 21:37:1–37:5, 2019.

[18] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv:2010.04050 [cs, stat]*, March 2021. URL http://arxiv.org/abs/2010.04050. arXiv: 2010.04050.

[19] Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Towards Causal Algorithmic Recourse. In Andreas Holzinger, Randy Goebel, Ruth Fong, Taesup Moon, Klaus-Robert Müller, and Wojciech Samek, editors, *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, Lecture Notes in Computer Science, pages 139–166. Springer International Publishing, Cham, 2022. ISBN 978-3-031-04083-2. doi: 10.1007/978-3-031-04083-2_8. URL https://doi.org/10.1007/978-3-031-04083-2_8.

[20] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30, 2017.

[21] Matt Kusner, Chris Russell, Joshua Loftus, and Ricardo Silva. Making Decisions that Reduce Discriminatory Impacts. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3591–3600. PMLR, May 2019. URL https://proceedings.mlr.press/v97/kusner19a.html. ISSN: 2640-3498.

[22] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual Fairness. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

[23] Matilde Lazzari, Jose M Alvarez, and Salvatore Ruggieri. Predicting and explaining employee turnover intention. *International Journal of Data Science and Analytics*, 14(3):279–292, 2022.

[24] Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B*, 83(5):911–938, 2021.

[25] Joshua R. Loftus, Chris Russell, Matt J. Kusner, and Ricardo Silva. Causal Reasoning for Algorithmic Fairness. *arXiv:1805.05859 [cs]*, May 2018. URL http://arxiv.org/abs/1805.05859. arXiv: 1805.05859.

[26] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553*, 2020.

[27] Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2022. URL https://christophm.github.io/interpretable-ml-book/.

[28] Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter*, 22(1):18–33, 2020.

[29] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[30] Judea Pearl. [bayesian analysis in expert systems]: Comment: Graphical models, causality and intervention. *Statistical Science*, 8(3):266–269, 1993. ISSN 08834237. URL http://www.jstor.org/stable/2245965.

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

[32] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2), 2000.

[33] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

[34] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. FACE: Feasible and Actionable Counterfactual Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350. Association for Computing Machinery, New York, NY, USA, February 2020. ISBN 978-1-4503-7110-0. URL https://doi.org/10.1145/3375627.3375850.

[35] Joseph Ramsey and Bryan Andrews. Fask with interventional knowledge recovers edges from the sachs model. *ArXiv*, abs/1805.03108, 2018.

[36] Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/1271a7029c9df08643b631b02cf9e116-Abstract.html.

[37] 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.

[38] Numair Sani, Daniel Malinsky, and Ilya Shpitser. Explaining the behavior of black-box prediction algorithms with causal learning. *arXiv preprint arXiv:2006.02482*, 2020.

[39] B Schölkopf, D Janzing, J Peters, E Sgouritsa, K Zhang, and J Mooij. On causal and anticausal learning. In *29th International Conference on Machine Learning (ICML 2012)*, pages 1255–1262. International Machine Learning Society, 2012.

[40] Amit Sharma and Emre Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020.

[41] Donghee Shin. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International Journal of Human-Computer Studies*, 146:102551, 2021.

[42] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[43] Ilia Stepin, Jose M. Alonso, Alejandro Catala, and Martín Pereira-Fariña. A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. *IEEE Access*, 9:11974–12001, 2021. ISSN 2169-3536. doi: 10.1109/ACCESS.2021.3051315. Conference Name: IEEE Access.

[44] W Nick Street, William H Wolberg, and Olvi L Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *Biomedical image processing and biomedical visualization*, volume 1905, pages 861–870. SPIE, 1993.

[45] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 10–19, New York, NY, USA, January 2019. Association for Computing Machinery. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287566. URL `https://doi.org/10.1145/3287560.3287566`.

[46] Ke Yang, Joshua R. Loftus, and Julia Stoyanovich. Causal Intersectionality and Fair Ranking. In Katrina Ligett and Swati Gupta, editors, *2nd Symposium on Foundations of Responsible Computing (FORC 2021)*, volume 192 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 7:1–7:20, Dagstuhl, Germany, 2021. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-187-0. doi: 10.4230/LIPIcs.FORC.2021.7. URL `https://drops.dagstuhl.de/opus/volltexte/2021/13875`. ISSN: 1868-8969.

[47] Mingzhang Yin, Claudia Shi, Yixin Wang, and David M Blei. Conformal sensitivity analysis for individual treatment effects. *Journal of the American Statistical Association*, pages 1–14, 2022.

[48] Junzhe Zhang and Elias Bareinboim. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[49] Qingyuan Zhao and Trevor Hastie. Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, 39(1):272–281, 2021. doi: 10.1080/07350015.2019.1624293. URL `https://doi.org/10.1080/07350015.2019.1624293`.

Supplemental Material: Causal Dependence Plots

6 Algorithms for Causal Dependence Plots

In this section, we provide algorithms to compute each of the causal dependence plots defined in the main text.

Algorithm 1 Explanatory Causal Model (ECM)
Inputs: $\mathcal{M}_{\mathbf{X}}$ (SCM), $\hat{f}$ (black-box predictor), $\mathbf{S} \subseteq \mathbf{X}$ (covariates used by black-box)

Make copy $\mathcal{M}'$ of SCM $\mathcal{M}_{\mathbf{X}}$
Add node for $\hat{Y}$ to causal graph $\mathcal{G}$ of SCM $\mathcal{M}'$
for x in $\mathbf{S}$ **do**
    Add edge in $\mathcal{G}$ from x to $\hat{Y}$
end for
Set structural equation for node $\hat{Y}$ to $\hat{f}$
Set exogenous variable $U_{\hat{Y}} \leftarrow 0$
return $\mathcal{M}'$

Algorithm 2 Total Dependence Plot (TDP)
Inputs: $\mathcal{M}_{\mathbf{X}}$ (SCM), $\hat{f}$ (black-box predictor), D (explanatory dataset), X_s (covariate of interest)

Get ECM $\mathcal{M}$ via Algorithm 1
Get the possible values of X_S and set to X
Set N to the number of observations in D
Initialize $N \times |X|$ matrix of estimates $\hat{Y}$
for x in X **do**
    Define intervention $I = \text{do}(X_S = x)$
    Sample counterfactual dataset D_c entailed by $P^{\mathcal{M}|D;\text{do}(I)}$
    Set $\hat{Y}[:, x] \leftarrow D_c[:, y]$ for index y corresponding to node $\hat{Y}$
end for
Plot N lines $(X, \hat{Y}[i, :])$ ▷ (Individual Counterfactuals)
Plot average $(X, \sum_i \hat{Y}[i, :]/N)$ ▷ (Causal Dependence)

Algorithm 3 Partially Controlled Dependence Plot (PCDP)
Inputs: $\mathcal{M}_{\mathbf{X}}$ (SCM), $\hat{f}$ (black-box predictor), D (explanatory dataset), X_s (covariate of interest), C (intervention controlling other variables in $\mathcal{M}_{\mathbf{X}}$)

Get ECM $\mathcal{M}$ via Algorithm 1
Get the possible values of X_S and set to X
Set N to the number of observations in D
Initialize $N \times |X|$ matrix of estimates $\hat{Y}$
for x in X **do**
    Define intervention $I = \text{do}(X_S = x, C)$
    Sample counterfactual dataset D_c entailed by $P^{\mathcal{M}|D;\text{do}(I)}$
    Set $\hat{Y}[:, x] \leftarrow D_c[:, y]$ for index y corresponding to node $\hat{Y}$
end for
Plot N lines $(X, \hat{Y}[i, :])$ ▷ (Individual Counterfactuals)
Plot average $(X, \sum_i \hat{Y}[i, :]/N)$ ▷ (Causal Dependence)

Algorithm 4 Natural Direct Dependence Plot (NDDP)

Inputs: $\mathcal{M}_{\mathbf{X}}$ (SCM), $\hat{f}$ (black-box predictor), D (explanatory dataset), X_s (covariate of interest)

Get ECM $\mathcal{M}$ via Algorithm 1
Get the possible values of X_S and set to X
Set N to the number of observations in D
Initialize $N \times |X|$ matrix of estimates $\hat{Y}$
Get all children of X_S in $\mathcal{M}$, excluding $\hat{Y}$, and store in $\mathbf{C}$
Make copy $\mathcal{M}'$ of SCM $\mathcal{M}$
for x in $\mathbf{C}$ **do**
    Remove all incoming edges to x from $\mathcal{M}'$
end for
for x in X **do**
    for i in N **do**
        Get observed values of all variables in $\mathbf{C}$ for unit i and store in $\mathbf{c}_i$
        Define intervention $I = \text{do}(X_S = x, \mathbf{C} = \mathbf{c}_i)$
        Sample counterfactual observation d_c for unit i entailed by $P^{\mathcal{M}'|D[i];\text{do}(I)}$
        Set $\hat{Y}[i, x] \leftarrow d_c[y]$ for index y corresponding to node $\hat{Y}$
    end for
end for
Plot N lines $(X, \hat{Y}[i, :])$ ▷ (Individual Counterfactuals)
Plot average $(X, \sum_i \hat{Y}[i, :]/N)$ ▷ (Causal Dependence)

Algorithm 5 Natural Indirect Dependence Plot (NIDP)

Inputs: $\mathcal{M}_{\mathbf{X}}$ (SCM), $\hat{f}$ (black-box predictor), D (explanatory dataset), X_s (covariate of interest)

Get ECM $\mathcal{M}$ via Algorithm 1
Get the possible values of X_S and set to X
Set N to the number of observations in D
Initialize $N \times |X|$ matrix of estimates $\hat{Y}$
Get all children of X_S in $\mathcal{M}$, excluding $\hat{Y}$, and store in $\mathbf{C}$
Make copy $\mathcal{M}'$ of SCM $\mathcal{M}$
for x in $\mathbf{C}$ **do**
    Remove all incoming edges to x from $\mathcal{M}'$
end for
Define intervention $I = \text{do}(X_S = x)$
for x in X **do**
    for i in N **do**
        Sample counterfactual observation d_c for unit i entailed by $P^{\mathcal{M}|D[i];\text{do}(I)}$
        Get counterfactual values of all variables in $\mathbf{C}$ from observation d_c and store in $\mathbf{c}_i$
        Define intervention $J = \text{do}(X_S = x, \mathbf{C} = \mathbf{c}_i)$
        Sample counterfactual observation d'_c for unit i entailed by $P^{\mathcal{M}'|D[i];\text{do}(J)}$
        Set $\hat{Y}[i, x] \leftarrow d'_c[y]$ for index y corresponding to node $\hat{Y}$
    end for
end for
Plot N lines $(X, \hat{Y}[i, :])$ ▷ (Individual Counterfactuals)
Plot average $(X, \sum_i \hat{Y}[i, :]/N)$ ▷ (Causal Dependence)

7 Real data with causal discovery

In this section, we explore the graph structures consistent with the uncertain edge in the DAG $\mathcal{G}_B$ in the main text for the Breast Cancer Wisconsin dataset.

Figure S1 shows the TDP, NDDP, and NIDP for a learned additive noise model (ANM) with three different structures consistent with $\mathcal{G}_B$: (1) an ANM with the edge Cell Shape $\rightarrow$ Cell Size, (2) an ANM with the edge Cell Size $\rightarrow$ Cell Shape, and (3) an ANM with no edge between Cell Size and

Cell Shape. This figure shows that the takeaway about cell shape impacting tumor class is indeed sensitive to our choice of what to assume about the uncertain edge, particularly in the case of total dependence.

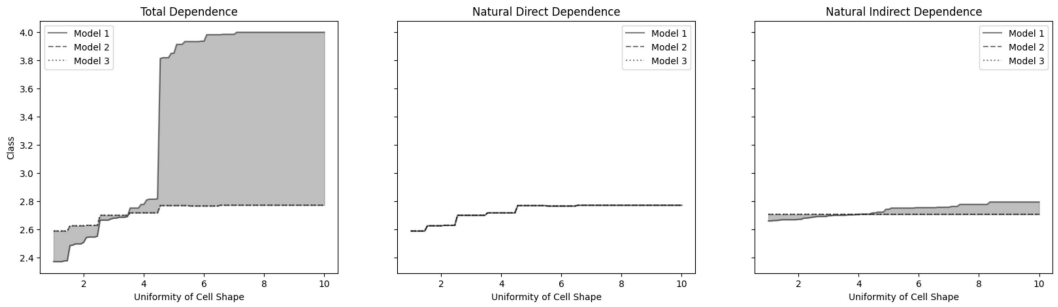

Figure S1: Total Dependence Plots, Natural Direct Dependence Plots and Natural Indirect Dependence Plots for the Breast Cancer Wisconsin dataset under three possible DAGs found by the PC algorithm: (1) $\mathcal{G}_B$ with the edge Cell Shape $\rightarrow$ Cell Size, (2) $\mathcal{G}_B$ with the edge Cell Size $\rightarrow$ Cell Shape, and (3) $\mathcal{G}_B$ with no edge between Cell Size and Cell Shape.

8 Real data with domain expertise

As an example that makes use of domain expertise for the underlying causal model, Figure S2 shows a DAG, TDP, and NDDP for the Sachs et al. [37] dataset, for which data and a ground-truth DAG[1] are publicly available in the Causal Discovery Toolbox [17]. While the actual biology of the problem is not our focus here, there are meaningful implications from Figure S2. From this model, the TDP shows a relationship between the two features that is importantly different from the NDDP. Recall that the NDDP captures the same relationship we would see from a PDP or an ICE plot. If we don't consider how other variables in the graph will change in response to changes in PKA, the trend we uncover will be practically reversed.

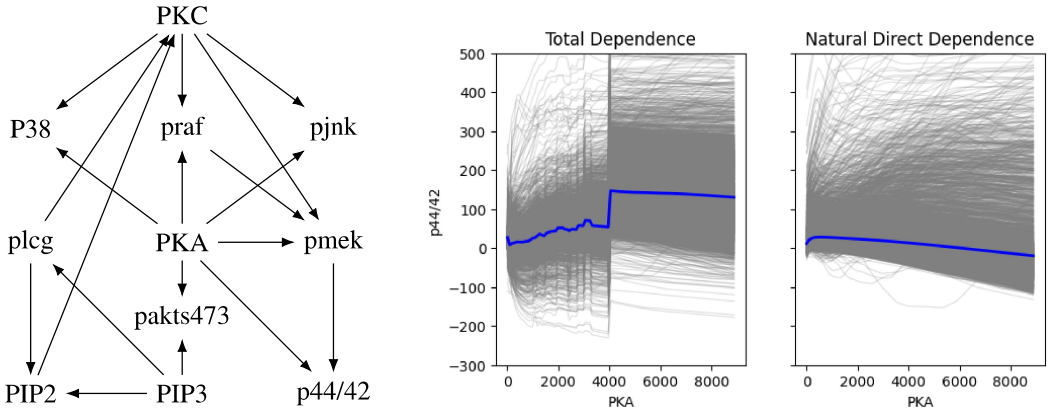

Figure S2: DAG $\mathcal{G}_S$ for the Sachs et al. [37] dataset as well as a corresponding Total Dependence Plot and Natural Direct Dependence Plot for the effect of PKA on p44/42.

[1]Following the discussion in Ramsey and Andrews [35] and follow-up ground truth DAG for the Sachs et al. [37] dataset in Ramsey and Andrews's Figure 5, we choose the edge PIP3 $\rightarrow$ PIP2 in order to eliminate a would-be cycle, and otherwise leave the released DAG unchanged.