

DiConStruct:

Causal Concept-based Explanations through Black-Box Distillation

Ricardo Moreira*

RICARDO.MOREIRA@FEEDZAI.COM

Jacopo Bono*

JACOPO.BONO@FEEDZAI.COM

Mário Cardoso*

MARIO.CARDOSO@FEEDZAI.COM

Pedro Saleiro*

PEDRO.SALEIRO@FEEDZAI.COM

Mário Figueiredo†

MARIO.FIGUEIREDO@TECNICO.ULISBOA.PT

Pedro Bizarro*

PEDRO.BIZARRO@FEEDZAI.COM

*Feedzai

†Instituto de Telecomunicações, ELLIS Unit Lisbon, Instituto Superior Técnico, ULisboa, Portugal

Editors: Francesco Locatello and Vanessa Didelez

Abstract

Model interpretability plays a central role in human-AI decision-making systems. Ideally, explanations should be expressed using human-interpretable semantic concepts. Moreover, the causal relations between these concepts should be captured by the explainer to allow for reasoning about the explanations. Lastly, explanation methods should be efficient and not compromise the predictive task performance. Despite the recent rapid advances in AI explainability, as far as we know, no method yet fulfills these three desiderata. Indeed, mainstream methods for local concept explainability do not yield causal explanations and incur a trade-off between explainability and prediction accuracy. We present DiConStruct, an explanation method that is both concept-based and causal, which produces more interpretable local explanations in the form of structural causal models and concept attributions. Our explainer works as a distillation model to any black-box machine learning model by approximating its predictions while producing the respective explanations. Consequently, DiConStruct generates explanations efficiently while not impacting the black-box prediction task. We validate our method on an image dataset and a tabular dataset, showing that DiConStruct approximates the black-box models with higher fidelity than other concept explainability baselines, while providing explanations that include the causal relations between the concepts.

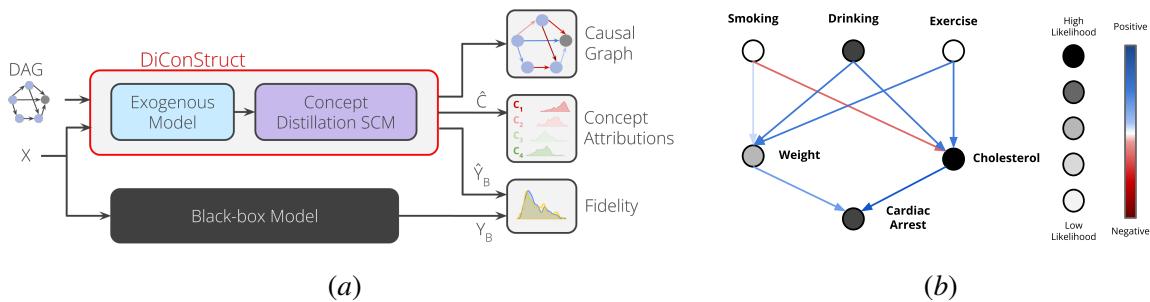


Figure 1: **DiConStruct is a surrogate explainer for obtaining explanations in the form of (1) causal graphs relating concepts to black-box scores and (2) concept attributions.** (a) DiConStruct overview. (b) Toy example of a learned causal graph explanation for predicting cardiac arrest. For example, a low likelihood of smoking (white node) has negative impact on cholesterol (red edge) but positive impact on weight (blue edge).

1. Introduction

A considerable amount of work in explainable AI (XAI) is based on feature attribution (Burkart and Huber, 2021). Those methods generate explanations for given machine learning (ML) model predictions as the set of contributions of each of the model inputs (features). In practice, feature-based explanations are often difficult to interpret (Kim et al., 2018) due to the absence of knowledge about the features' semantics and their connection to higher level concepts. This lack of interpretability makes feature attributions unsuitable for tasks where human-AI collaboration is vital, such as real-time human validation or correction of model decisions. Moreover, without incorporating the causal principles of interventions and counterfactuals, purely statistics-based ML systems are incapable of reasoning about explanations (Pearl, 2018). The understanding of such limitations has motivated a considerable amount of work applying causality to XAI. Although promising, most of these works still focus on feature-level explainability, such as feature attributions (Chattopadhyay et al., 2019; Frye et al., 2020; Wang et al., 2021; Heskes et al., 2020; Janzing et al., 2020; Khademi and Honavar, 2020; Zhao and Hastie, 2021; Jung et al., 2022), feature-based counterfactual explanations (Karimi et al., 2020, 2021; Bertossi, 2020), and causally induced model structure (Geiger et al., 2022).

To improve the interpretability of explanations, there has been an increasing effort to develop methods that produce explanations understandable by non-AI experts. Towards this effort, leveraging higher-level concepts and producing concept-based explanations have become a popular method (Kim et al., 2018). These concepts are human-defined meaningful characteristics that should describe well the most relevant patterns in the data. For example, in a dataset where the task is to classify birds, the wing color could be chosen as a concept. In this way, global or local explanations in the form of concept attributions are produced. Self-explaining methods (Koh et al., 2020; Alvarez-Melis and Jaakkola, 2018) jointly learn the prediction and the explanation tasks with a single model, but have a potential disadvantage of creating a trade-off between the two tasks (Bahadori and Heckerman, 2021). Such trade-off is especially undesirable in contexts where keeping the performance of the main prediction task is critical. Finally, apart from work by Sani et al. (2020), which is limited to producing global explanations, concept explanation methods are still not incorporating causal principles, limiting their interpretability by ignoring the underlying causal relationships.

The problems mentioned above motivate us to develop an explanation method that incorporates the principles of causality and provides concept-based explanations, as depicted in Figure 1. We aim to improve the quality of explanations produced by understanding the causal relationships between the concepts and predictions, in the form of a learned causal graph. Furthermore, we aim at creating a surrogate explainer instead of a self-explaining method, ensuring that the main prediction task is never compromised. In the following section, we provide further arguments for why we believe XAI research should be focusing on bringing together concept explanations and causal modeling.

In summary, the main contribution of this work is a novel explainer, with the following key characteristics:

- it is concept-based and causal, yielding more interpretable local explanations in the form of structural causal models (SCM) and concept attributions;
- it is a surrogate model, hence not affecting the predictive performance of the ML model, while simultaneously producing explanations efficiently in a single forward pass.

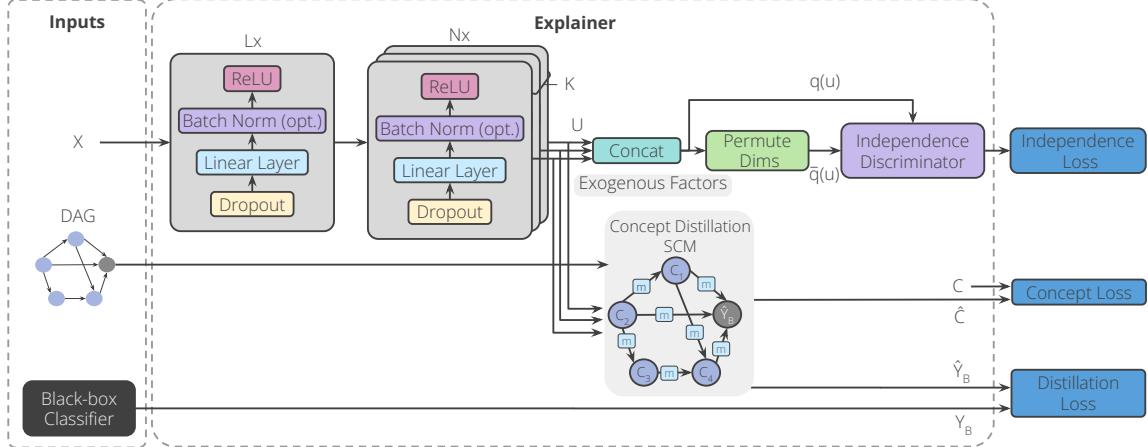


Figure 2: DiConSTRUCT architecture. On the left side are the inputs to our explainer: (1) an instance to explain x ; (2) a DAG relating the concepts (Section 2.3 discusses approaches to obtain this DAG); (3) a black-box classifier.

Moreover, we validate our explainer in two different domains (image data and tabular data), showing that it approximates the black-box models with higher fidelity than other concept explainability baselines, while providing explanations that include the causal relations between the concepts.

The code of our experiments is publicly available ¹.

2. DiConSTRUCT

2.1. Preliminaries

In the same vein of what was proposed by Sani et al. (2020), we model the causal mechanism $C \rightarrow \hat{Y}$, where C are concepts and \hat{Y} the black-box model predictions (see Appendix C). Our goal is to gain insights into what information the ML model learned during training (global explanations) as well as explaining the individual decision processes (local explanations). We frame our explainability problem as a specific knowledge distillation task, where the explainer acts as a surrogate model learning to mimic a black-box binary classifier f_B , trained to produce the score of a binary target, $f_B(x) = P(y = 1|x) = y_B$. The task of our explainer is, given some instance x , to produce high-fidelity estimates of y_B , while also correctly predicting the K semantic concepts, $\{c_k\}_{k=1}^K$, associated with that instance. All these predictions are produced in a structural causal model (SCM) that models the joint distribution $P(C, Y_B)$. For each instance, the explainer produces a local SCM through corresponding local concept attributions.

We define this explainer more formally as $\Phi : \mathbb{X} \rightarrow \mathbb{Y}_B \times \mathbb{Y}_C$, where \mathbb{X} is the space of the model features, $\mathbb{Y}_B = [0, 1]$ is the space of the classifier estimates, y_B , corresponding to the distillation targets, and $\mathbb{Y}_C = \{0, 1\}^K$ is the space of binary labels for the K semantic concepts. Function Φ is learned on both objectives mentioned in the previous paragraph using a dataset

1. <https://anonymous.4open.science/r/causal-diconstruct-4F18/>

$\mathcal{D}_\phi = \{(\mathbf{x}^{(i)}, y_B^{(i)}, \{c_k^{(i)}\}_{k=1}^K)\}_{i=1}^M$ with M instances and a pre-defined directed acyclic graph (DAG): \mathcal{G}_{C,Y_B} . The DAG is assumed to be an input to our explainer (see Section 2.3). The explainer then produces estimates $\Phi(\mathbf{x}) = (\hat{y}_B, \{\hat{c}_k\}_{k=1}^K)$ and explanations in the form of an SCM. We refer to Appendix B for a formal definition of an SCM and its relation with a DAG. Finally, one can obtain instance-wise concept attributions by performing counterfactual analysis on this causal graph (details in Section 2.6).

We define 2 necessary conditions for the obtained explanations to be causally valid.

- **Exogenous independence:** the external variables of the SCM are so-called "exogenous" variables (definition in Appendix B), which are learned for each instance. For interpretability purposes, the exogenous variables should be mutually independent. If this condition is not satisfied, it implies that some other unmeasured variable is causing their dependence, meaning that not all relationships between the concept variables are being represented in the explainer's structural equations. This would make interventions and counterfactual analysis unfeasible, since not all required information would be available in the SCM. In other words, we require *causal sufficiency* (Pearl, 2009) in order to perform counterfactual analysis.
- **Concept completeness:** for our explainer to approximate the black-box model, we require the causal parents of Y_B in \mathcal{G}_{C,Y_B} to fully explain the black-box behavior. For this to be possible, we need (1) the pre-defined concept variables to verify the *completeness* criteria (Yeh et al., 2020) and that (2) all causal mechanisms that are responsible for correctly approximating Y_B to be in place in the DAG. Because our explainer is a surrogate model, the targets are not Y , but the predictions of a black-box model Y_B . In this scenario, the *completeness score* proposed by Yeh et al. (2020) reduces to the accuracy of the concept model $g(\mathbf{c}) = \hat{Y}_B$ (i.e. the model predicting the label just given the concept scores) at predicting the targets Y_B ².

We name our explainer DiConStruct (from *distillation*, *concept-based* and *structural* causal model); and it can be divided into two components (see Figure 2):

- **Exogenous model:** learns to predict independent exogenous variables $\mathbf{u} = (u_1 \dots u_K)$ from X ;
- **Concept distillation SCM:** learns the joint distribution $P(C, Y_B)$ in an SCM with partial structural assignments $m_{j,k}$ for each edge (j, k) .

In the following subsections, we define our exogenous model, propose how the causal DAG \mathcal{G}_{C,Y_B} can be obtained for modelling $P(C, Y_B)$, define our concept distillation SCM, and, finally, show how we obtain *concept attributions*.

2.2. Exogenous Model with an independence objective

To guarantee the exogenous independence condition (Section 2.1), we learn our exogenous model, defined as $f_E : \mathbb{X} \rightarrow \mathbb{Y}_C$, imposing an independence objective on its outputs. The exogenous model is designed as a multi-output feedforward neural network consisting of L common layer blocks and N concept-specific layer blocks (see Figure 2). Each layer block is composed of a dropout layer, a linear layer, an optional batch normalization layer (Ioffe and Szegedy, 2015), and a ReLU activation function. The outputs are subject to an independence objective, for which we followed prior art by Kim and Mnih (2018). Essentially, we want to make the joint distribution of exogenous variables, $q(\mathbf{u}) = p(u_1, \dots, u_K)$, indistinguishable from $\bar{q}(\mathbf{u}) = \prod_{k=1}^K p(u_k)$. As proposed by Arcones and Gine (1992), we sample from $\bar{q}(\mathbf{u})$ by taking samples $\mathbf{u} \sim q(\mathbf{u})$ and

2. The term in the denominator of the original *completeness score* definition will always be 1 (maximum) since we can always find a model that approximates $f_B(x)$ perfectly, like the black-box itself.

randomly shuffling the components of \mathbf{u} (i.e. exogenous variables) across a batch of instances. For each component, a different seed is used for shuffling, breaking all existing dependencies in $q(\mathbf{u})$. We use an independence discriminator model $f_I : \mathbb{Y}_C \rightarrow [0, 1]$, trained to distinguish $q(\mathbf{u})$ from $\bar{q}(\mathbf{u})$, as a means to approximate the Kullback–Leibler divergence between $q(\mathbf{u})$ and $\bar{q}(\mathbf{u})$ by (Arcones and Gine, 1992; Nguyen et al., 2010):

$$D_{KL}(q(\mathbf{u}) || \bar{q}(\mathbf{u})) = \mathbb{E}_{q(\mathbf{u})} \left[\log \frac{q(\mathbf{u})}{\bar{q}(\mathbf{u})} \right] \approx \mathbb{E}_{q(\mathbf{u})} \left[\log \frac{f_I(\mathbf{u})}{1 - f_I(\mathbf{u})} \right] := \mathcal{L}_E. \quad (1)$$

2.3. Obtaining a Causal DAG for $P(C, Y_B)$

As mentioned previously, to achieve high-fidelity approximations of the black-box score, we require a close to *complete* concept definition. The task of defining which concepts are employed for explaining the black-box model is commonly performed by domain experts. Given the concepts, we then propose 2 different ways of obtaining \mathcal{G}_{C, Y_B} : (1) from expert knowledge and (2) using causal discovery. While option (2) is more attainable since it depends solely on data, we stress that it still requires concept definition to be made beforehand. Furthermore, causal discovery will not correct for any possible lack of *completeness* of the concept definition, as it only performs structure learning, not discovery of new concepts. It is important to highlight that the concepts and causal DAG are given as inputs to our method; their discovery is not part of our method.

2.4. Global and Local Structural Causal Concept Distillation

Given the exogenous variables (Section 2.2) and a causal graph (Section 2.3), we now define our second component of the DiConStruct explainer: the concept distillation SCM. This model takes any causal graph \mathcal{G}_{C, Y_B} and creates a set of partial structural assignments $m_{j,k}$ for each graph edge (j, k) . These partial structural assignments encode the strengths of the direct effects of the causal mechanisms $C_j \rightarrow C_k$ and $C_j \rightarrow Y_B$. We model those causal mechanisms in two ways. Firstly, in a global way, by assuming a linear dependency on *logits* of the different variables in the graph,

$$m_{j,k}(\hat{c}_j) = \sigma^{-1}(\hat{c}_j)w_{j,k}, \quad (2)$$

where $k = 1, \dots, K$, σ is the sigmoid function, \hat{c}_j is the structural assignment of concept j and will be defined below in equation 4, and $w_{j,k}$ can be thought of as global edge weights. Secondly, in a local way, by introducing the non-linear weighting functions $W_{j,k} : [0, 1]^K \times \mathbb{Y}_B \rightarrow \mathbb{R}$, which receive as inputs the set of exogenous variables U , and the true black-box score,

$$m_{j,k}(\hat{c}_j, \mathbf{u}, y_B) = \sigma^{-1}(\hat{c}_j)W_{j,k}(\mathbf{u}, y_B), \quad (3)$$

where the $W_{j,k}(\mathbf{u}, y_B)$ define the edge weights for each instance. The black-box model score y_B is included as input to the weighting functions as a means to improve the distillation fidelity. We create the partial structural assignments following the same formulation both for edges of type $C_j \rightarrow C_k$ and type $C_j \rightarrow Y_B$. Finally, we obtain the structural assignments for each concept in the graph as

$$\hat{c}_k := f_k(\mathbf{PA}_k, u_k) = \sigma \left(b_k + \sigma^{-1}(u_k) + \sum_{j \in \mathbf{PA}_k} m_{j,k} \right), \quad (4)$$

where \mathbf{PA}_k is the index set of causal parents of C_k in $\mathcal{G}_{C, \hat{Y}_B}$. Importantly, these structural assignments are computed by traversing the DAG in topological order, starting from the nodes without

parents. Depending on the instantiation of the explainer, the partial $m_{j,k}$ can take either the global or the local formulation. We use an analogous structural assignment for \hat{Y}_B , without the logit term since no exogenous variable is learned for Y_B .

The structural assignments presented above depend on static weights $w_{j,k}$ (or $w_{j,B}$) or weighting functions $W_{j,k}$ (or $W_{j,B}$) and biases b_k (or b_B). All these components are learned during training and fixed at inference time. To learn these components, we use two binary cross-entropy (BCE) losses between the predictions \hat{c}_k and \hat{y}_B and the targets c_k and y_B . We name the losses \mathcal{L}_C (explainability loss) and \mathcal{L}_D (distillation loss), respectively:

$$\mathcal{L}_C = \sum_{i=1}^M \sum_{k=1}^K c_k^{(i)} \log_2(\hat{c}_k^{(i)}), \quad (\text{explainability}) \quad (5)$$

$$\mathcal{L}_D = \sum_{i=1}^M y_B^{(i)} \log_2(\hat{y}_B^{(i)}), \quad (\text{distillation}) \quad (6)$$

with M denoting the instances in the evaluation set.

2.5. Learning the DiConStruct explainer

All components of the DiConStruct explainer are learned jointly through gradient descent (GD) and backpropagation using the combined losses from equations 1, 5 and 6:

$$\mathcal{L} = \gamma \mathcal{L}_E + \beta \mathcal{L}_C + \mathcal{L}_D, \quad (7)$$

where β and γ are tunable hyperparameters. We point out that the gradients of \mathcal{L}_E only affect the exogenous component. Because \mathcal{L}_E consists of an expectation over a sample of instances, we compute it in batches, therefore restricting our method to learn through batch or mini-batch GD.

To train the independence discriminator f_I to distinguish $q(\mathbf{u})$ from $\bar{q}(\mathbf{u})$, we also use mini-batch GD with a BCE loss. The GD updates performed on the discriminator are interleaved with the updates of the other model components. Optionally, the updates of f_I might occur at a configurable frequency, set by a parameter θ .

2.6. Concept Attributions from Counterfactual Statements

For a given instance $\mathbf{x}^{(i)}$, to measure the total causal effect of a given concept C_k on Y_B , we can use the following counterfactual statements from our SCM $\mathcal{C}^{(i)}$:

$$P_{do(C_k:=a)}^{\mathcal{C}^{(i)}}(Y_B^{(i)} = 1), \quad (8)$$

which, depending on the value of $a \in \{0, 1\}$, aims to answer different questions. If $a = 1$, the counterfactual question is: *what would have been the probability of a positive prediction of the black-box, had the concept C_k been present in the instance $\mathbf{x}^{(i)}$* . Conversely, if $a = 0$ the counterfactual question becomes: *what would have been the probability of a positive prediction of the black-box, had the concept C_k not been present in the instance $\mathbf{x}^{(i)}$* . These two statements are computed following the three steps for computing counterfactuals suggested by Pearl (2009):

- **Abduction:** we use the predicted $\mathbf{u}^{(i)}$ as the necessary exogenous variables for this instance;

- **Action:** for the assignment $do(C_k := 1/0)$, we directly modify $\mathcal{C}^{(i)}$ by removing all causal parents from C_k and setting its value to either 0 or 1;
- **Prediction:** leaving all the remaining concept predictions fixed, we compute the new value of \hat{y}_B , which corresponds to the counterfactual quantity.

We then obtain the concept attributions as total concept effects (TCE):

$$\text{TCE}^{(i)}(C_k := a) = P_{do(C_k:=a)}(Y_B^{(i)} = 1) - P(Y_B^{(i)} = 1), \quad a \in \{0, 1\}, \quad (9)$$

where $P(Y_B^{(i)} = 1)$ corresponds to the *factual* probability of the black-box classifier prediction to be positive for the instance $x^{(i)}$.

3. Experimental Setup

3.1. Datasets and black-box models

CUB-200-2011. We validate our method using the Caltech-UCSD Birds-200-2011 (Wah et al., 2011) dataset, consisting of 11,788 bird images in 200 subcategories (with 316 annotated concepts). We split this into train, validation, and test, with 4818, 1176, and 5794 images, respectively. The binary classification task is detecting the *warbler* bird category, the most prevalent one (12.63% prevalence). Additionally, since concepts have multiple values, we binarize the concepts by assigning 1 to the most prevalent attribute value and 0 otherwise (see Appendix F). To train a black-box model, we use a pre-trained ResNet-34 (He et al., 2016) model, removing the last 8 layers and attaching a fully connected network with varying layer dimensions.

Merchant Fraud. We use a privately held online retailer fraud detection dataset, with approximately 5.5M transactions, 2% of which represent fraudulent behavior. Concept labels are assigned using a weak supervision method starting from a small sample of human-labeled data (Appendix F). Three sequential subsets form the train, validation and test sets, composed of 4.8M, 200k, and 1M instances, respectively. We trained 2 different black-box models on this dataset, on the task of predicting the binary outcome: *fraud/legit*. The first black-box model is a LightGBM (Ke et al., 2017) and the second a feedforward neural network (NN). Because our goal is model explainability, we create a sample of the transactions that would most likely need to be explained, containing the transactions of highest uncertainty for the black-box fraud detection system. This resulted in 460k instances for training, 17k for validation, and 107k for testing. The fraud rate in this sample is 21.6%. Our DiConStruct explainers and baselines were trained and tested using these sampled datasets.

3.2. Evaluation Metrics

Main Task Performance. Given the class imbalance, we chose to use the metric true positive rate (TPR) evaluated at a fixed false positive rate (FPR), which we set to be 5%. In other words, this translates to maximizing the amount of correctly classified minority class events (label positives) while incorrectly classifying at most 5% of the majority class (label negatives).

Fidelity. We measure how close the prediction of our distillation model is to the black-box model prediction. We use the 1 - MAE (mean absolute error) to quantify the distillation quality.

Concept Performance. We measure this performance by calculating the prediction accuracy for each concept, corresponding to the error rate for binary decisions using 0.5 as probability threshold, and then averaging that over the K concepts.

3.3. Full Setup

We distinguish four DiConStruct variants corresponding to the four combinations of two configurations: exogenous model *with or without the independence* objective; *local or global* concept distillation SCM configurations.

To obtain the causal DAG, we resort to causal discovery methods for each of the three black-box models: the PC algorithm (Spirtes et al., 2000); the ICA-LiNGAM method (Shimizu et al., 2006); the NO TEARS algorithm (Zheng et al., 2018). We restricted the causal discovery to DAGs in which the black-box score variable has no children, since the score of the black-box model is never a cause for any of the concepts. Additionally, we added a *trivial* DAG baseline in which all concepts are direct parents of \hat{Y}_B without edges between any of the concepts.

Combining these four variants, with the four causal graphs obtained for each of the three black-box models (CUB NN, Merchant Fraud NN, Merchant Fraud LGBM), yields a total of 48 experiments. In each experiment, we perform a random hyperparameter search of 50 iterations, resulting in a total of 2400 trained DiConStruct explainers. Additional details regarding hyperparameters are provided in Appendix D.

4. Results

4.1. Baselines

We compare our DiConStruct method with the following baselines:

Single task – task performance: the best black-box models on the main classification task.

Single task – concept prediction: we train a model with the same structure as the exogenous model (excluding the independence objective components). The outputs of the K concept-specific layers are directly compared with the true labels using the concept loss \mathcal{L}_C .

Single task – knowledge distillation: for the knowledge distillation task, we train a standalone feedforward neural network with a BCE loss.

Concept bottleneck models (CBMs): we reproduced the concept bottleneck model framework (Koh et al., 2020). The CBM is a self-explaining model given by $f(g(x))$, where $g : \mathbb{X} \rightarrow \mathbb{Y}_C$ and $f : \mathbb{Y}_C \rightarrow \mathbb{Y}_B$. Following the *joint* approach proposed by Koh et al. (2020), we set the λ parameter to 1, giving equal weight to the main task and concept prediction losses. We varied the same parameters on g and f as we did for our DiConStruct explainers. Since DiConStruct is a post hoc explainer, we also trained a CBM as a surrogate explainer by passing the black-box scores as targets for f . We therefore have two types of CBM baselines: (1) the *joint CBM* as an intrinsically explainable method; and (2) the *distillation joint CBM* as a post hoc explainer.

For each combination of baseline, black-box model (if applicable), and dataset we perform a separate random hyperparameter search for 50 iterations for the single task baselines and 50 for the concept bottleneck baselines. Information about the hyperparameters is provided in Appendix D.

4.2. Fidelity and concept accuracy performances

We estimate the variability of hyperparameter search by bootstrapping over the trained explainers. Each bootstrap trial consists of 20 random draws from the pool of 50 trained models per experiment. The best performing model (maximizing the sum of the validation fidelity and concept accuracy) of each trial is then separately selected. This process was repeated for 1000 trials to obtain a distribution of results for each metric. Table 1 shows the results for the DAGs attaining the highest

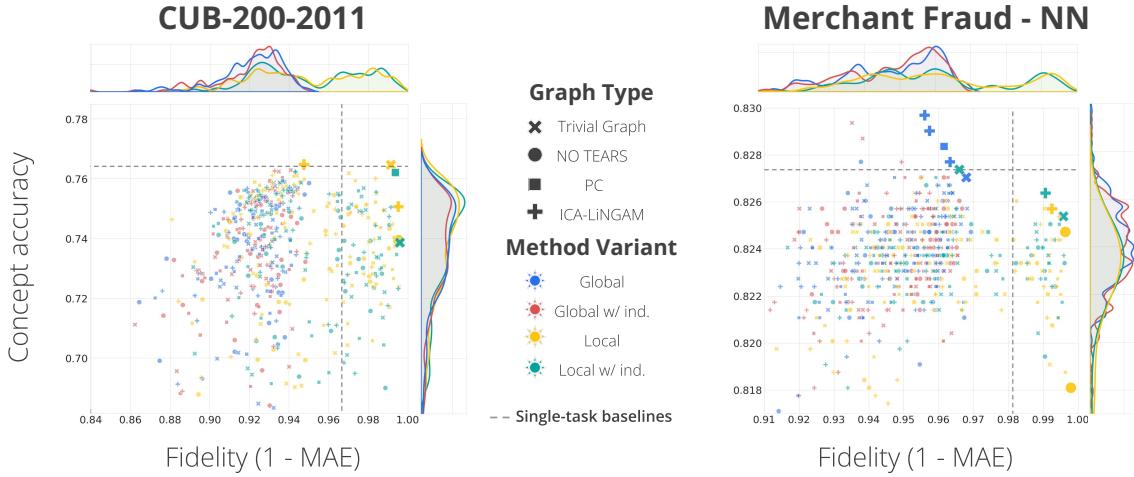


Figure 3: Distribution of fidelity and concept accuracy for the CUB-200-2011 (left) and Merchant Fraud (right) black-boxes. Each plot shows the result of 16 experiments of 50 trials, and each marker is a trained DiConStruct explainer. Models on the Pareto efficiency frontier are highlighted with bigger markers.

sum of mean fidelity and mean concept accuracy on each method variant. For comparison, we add the results of the CBMs and the single task baselines, for which we performed the same bootstrap analysis. In terms of concept accuracy, DiConStruct variants are comparable with both the CBMs and the single task baselines. Being a surrogate explainer, the main task performance of our DiConStruct variants is the same as the best black-box model. With the original CBMs, we observe a significantly lower task performance resulting from the tradeoff caused by the joint training of both tasks. This drop in performance can be mitigated by training the CBM as a surrogate explainer (distillation CBM). Comparing the global variants of our method with the baselines on the fidelity metric, we observe the performances are comparable. Local variants of DiConStruct, on the other hand, outperform the baselines in terms of fidelity for the two dataset and black-box models. As mentioned previously, this is due to the use of the black-box score as an input to the local concept distillation SCM. Similar conclusions can be drawn from Figure 3, where we see that some configurations outperform the single task baselines individually, particularly notable for the local variants. The full set of results considering each variant and DAG is provided in Appendix E.2.

4.3. Local Explanation Example

Figure 4 shows an example of a single-instance explanation for the Merchant Fraud dataset in the form of a learned SCM (left) and corresponding concept attributions (right). Concept attributions are obtained by intervening on each concept individually $do(C_k := a)$, $a \in \{0, 1\}$, measuring the total concept effect (TCE) (Equation 9) for both interventions and summing their absolute values. Since interventions are performed in the logit space, the assigned values should be in-distribution and associated with the concept being present/not present. Following what is proposed in the CBM paper by Koh et al. (2020), we set the values to be the 5th or 95th percentile over the training distribution

	Model	Variant	Validation			Test		
			Task Perf. (%)	Concept Perf. (%)	Fidelity (%)	Task Perf. (%)	Concept Perf. (%)	Fidelity (%)
CUB-200-2011	Ours	Global	75.58 ± 0.65	93.52 ± 0.77	-	75.44 ± 0.44	94.3 ± 0.8	-
		Global w/ Ind.	77.85	75.61 ± 0.69	93.16 ± 0.75	79.05	75.43 ± 0.65	93.83 ± 0.64
		Local	-	75.25 ± 0.76	98.72 ± 0.79	-	75.11 ± 0.63	98.79 ± 0.74
		Local w/ Ind.	-	75.05 ± 1.08	98.78 ± 0.86	-	74.89 ± 1.19	98.83 ± 0.8
Merchant Fraud - NN	Baselines	Joint CBM ($\lambda = 1$)	79.25 ± 0.98	75.57 ± 0.46	-	67.33 ± 2.13	75.76 ± 0.55	-
		Distill. Joint CBM ($\lambda = 1$)	77.85	75.48 ± 0.53	93.1 ± 0.52	79.05	75.52 ± 0.59	93.9 ± 0.56
		Single task - Task Perf.	77.85	-	-	79.05	-	-
		Single task - Concept Perf.	-	76.11 ± 0.21	-	-	76.07 ± 0.26	-
		Single task - Fidelity	-	-	96.07 ± 0.49	-	-	96.33 ± 0.26
Ours	Ours	Global	-	82.64 ± 0.14	97.12 ± 0.29	-	82.58 ± 0.12	96.62 ± 0.28
		Global w/ Ind.	74.67	82.6 ± 0.11	96.96 ± 0.13	63.35	82.55 ± 0.09	96.45 ± 0.24
		Local	-	82.5 ± 0.14	99.39 ± 0.37	-	82.45 ± 0.13	99.27 ± 0.42
		Local w/ Ind.	-	82.47 ± 0.13	99.34 ± 0.41	-	82.42 ± 0.12	99.23 ± 0.49
Baselines	Baselines	Joint CBM ($\lambda = 1$)	48.42 ± 0.31	82.49 ± 0.14	-	47.47 ± 3.64	82.34 ± 0.08	-
		Distill. Joint CBM ($\lambda = 1$)	74.67	82.62 ± 0.13	96.87 ± 0.18	63.35	82.57 ± 0.12	96.19 ± 0.29
		Single task - Task Perf.	74.67	-	-	63.35	-	-
		Single task - Concept Perf.	-	82.25 ± 0.19	-	-	82.25 ± 0.19	-
		Single task - Fidelity	-	-	98.13 ± 0.22	-	-	97.86 ± 0.23

Table 1: Mean and standard deviation of results for the CUB-200-2011 (top) and the Merchant Fraud (bottom) NN black-boxes. Results shown for the task performance (Task Perf.), concept accuracy (Concept Perf.) and Fidelity metrics.

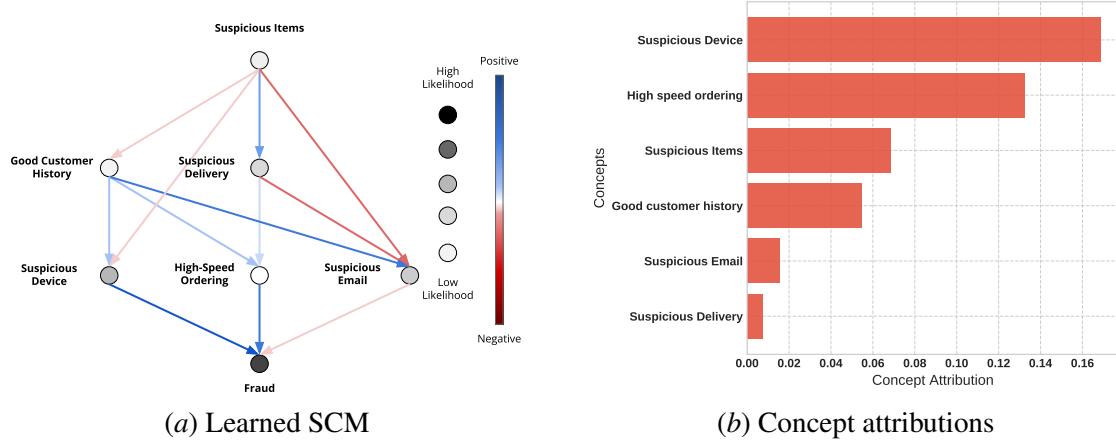


Figure 4: (a) Learned SCM for an instance of the Merchant Fraud dataset. Blue edge color denotes positive interactions, red denotes negative interactions, and the intensity represents the interaction strength. A positive/negative interaction increases/decreases the value of the destination node, respectively. Concept likelihood (nodes) is encoded from white (low) to black (high). (b) Concept attribution plot for the same instance.

of each corresponding concept, respectively simulating the absence ($a = 0$) or presence ($a = 1$) of a chosen concept. We can observe that *Suspicious Device* is the concept that contributes the most for predicting *Fraud* given its strong positive interaction and high concept attribution. Similarly, the absence of *Good Customer History*, a concept we semantically associate with legitimate events, is

increasing the likelihood of the children fraudulent concepts. Through these explanations, users can obtain a clearer understanding of not just which concepts are the most relevant for a prediction, but also how they causally affect each other.

5. Related Work

Previous works have approached explainability as the task of assessing the causal effects of a model’s inputs (X) on their outputs (\hat{Y}). Most methods focus on the raw input features, providing explanations in the form of feature attributions, but others are closer to our method and focus on concept explanations through causal graphs or by providing concept attributions.

Causal feature attribution: Most methods provide explanations as a set of contributions of each of the model inputs (features). Some works interpret ML models as causal mechanisms, where the input features are the causes and the model predictions are the effects (Janzing et al., 2020; Chattopadhyay et al., 2019; Geiger et al., 2022; Schwab and Karlen, 2019; Wang et al., 2021; Khademi and Honavar, 2020; Frye et al., 2020; Heskes et al., 2020; Jung et al., 2022). With this interpretation, Frye et al. (2020); Wang et al. (2021); Jung et al. (2022); Heskes et al. (2020) propose extensions of SHAP (Lundberg and Lee, 2017) that consider the causal structure in the data when calculating the Shapley values. Other methods seek to align neural networks with SCMs, enabling training schemes that produce explainable models (Geiger et al., 2022) or calculating feature attributions through the use of *do*-calculus (Chattopadhyay et al., 2019). Finally, Schwab and Karlen (2019) propose a surrogate explainer trained to approximate a black-box model using the Granger-causal objective. Although easily applicable, feature-attribution methods can suffer from a lack of interpretability due to high-dimensionality and lack of knowledge about the feature semantics.

Concept explainability: One of the seminal works in this area is due to Kim et al. (2018), who developed TCAV. Using TCAV, Ghorbani et al. (2019) automate the search for concept definitions in images. The CAV method was recently used to mitigate learning of spurious correlations in the data (Wu et al., 2023). Following a different approach, Alvarez-Melis and Jaakkola (2018) propose to learn an explainability task together with the classification task, allowing for explanations to be obtained together with predictions at test time. Concept bottleneck models (CBM) (Koh et al., 2020), work similarly but allow for test-time interventions. Subsequent works have criticized CBMs, claiming the learned concepts lack semantically meaningful input patterns (Margelioiu et al., 2021), and that using two objectives results in a trade-off between explainability and performance (Yüksekgün̄l et al., 2022). The latter work also extends CBMs with a residual predictor that seeks to recover the original main classification performance. Finally, Yeh et al. (2020) put forward the definition of *completeness* for assessing how much information the used concepts share with the main classification task, and then propose a methodology to discover completeness-aware concepts. Extending concept-based explanations to tabular data, particularly in fraud detection, Balayan et al. (2020) and Belém et al. (2021) have proposed a CBM-like approach. This line of work is further expanded by Sousa et al. (2022), who introduced a surrogate explainer approach.

Causality for concept explainability: Sani et al. (2020) apply causal discovery to learn a DAG between *concepts* and predictions. Bahadori and Heckerman (2021) propose a causal solution for the lack of *completeness* in CBMs involving adjusting the concept targets as being predicted by the main classification targets. Goyal et al. (2019) propose a framework that seeks to estimate the causal effects of concepts through interventions; a conditional VAE generates instances where the existing concepts can be intervened on its bottleneck, used to estimate the average causal effect (ACE) of

each concept in the black-box model outputs. Similarly, O’Shaughnessy et al. (2020) propose a VAE with a causal influence objective, imposing that some latent space variables influence the outputs of the downstream black-box model, while others are just useful to guarantee a low reconstruction loss. The former variables can be interpreted as concepts learned from the data.

6. Limitations and future work

The assumption of a complete set of concepts is strong and it may not be satisfied in a wide set of tasks. In the same vein as what was proposed by Havasi et al. (2022), it would be interesting to try future variations of DiConStruct where latent concepts are introduced and allowed to cover any incomplete information. Other possible avenues to relax the need for human labeling are recent methods that exploit language models (Oikarinen et al., 2023; Yang et al., 2023). Additionally, one limitation of our current setup is that we model each concept and label as a binary. In the future, it would be important to extend the framework to the multi-class case, for example, by using vector-valued nodes in the SCM, changing the sigmoid to softmax activations, and using a multi-class cross-entropy loss instead of the binary cross-entropy.

Recently, it was shown that models with soft concepts can leak information that is detrimental to the interpretability and trustworthiness of the model (Havasi et al., 2022). Since DiConStruct uses an SCM label predictor, this type of leakage is expected to be somewhat mitigated. Nonetheless, it would be interesting to quantify how much leakage would still be present. Moreover, since we assume a complete set of concepts as well as an SCM capturing the dependencies between concepts, it is argued by Havasi et al. (2022) that a hard concept version should not compromise task performance while avoiding concept leakage. It would therefore be interesting to compare a hard concept DiConStruct model to the current results.

Since our proposed explainer distills the black-box model, it is prone to copy any learned bias or shortcuts used by that model. Our goal is precisely to create a causal explainer for the black-box predictions: biases should then be surfaced by analyzing the explanations. Different methods could subsequently be applied to mitigate the detected sources of bias.

Recently, Ghosh et al. (2023a,b) proposed MoIE, a post hoc method providing explanations of a black-box model by iteratively extracting interpretable models. These interpretable models in turn provide explanations using first-order logic on the concepts. In future experiments, it would be interesting to compare how the causal explanations from our DiConStruct method relate to the first-order logic relations extracted by MoIE.

7. Conclusion

Effective human-AI collaboration can benefit from model explanations expressed through semantic concepts and their causal relations. To this end, we proposed DiConStruct, a surrogate explainer designed to produce explanations that are both causally grounded and concept-based.

DiConStruct produces explanations in the form of a structural causal model (SCM) describing the interactions between the concepts and the black-box model output. We also obtain concept attributions through counterfactual analysis on the predicted SCM. On two diverse datasets, we show that DiConStruct does not sacrifice performance on the main classification task while simultaneously enriching explainability by facilitating an understanding of the causal relationships between the concepts and the predictions. This is in contrast with state-of-the-art concept explainability

methods such as CBM, for which the two objectives of concept prediction and main task performance compete with each other.

References

- David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 7786–7795, 2018. URL <https://proceedings.neurips.cc/paper/2018/hash/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html>.
- Miguel A Arcones and Evarist Gine. On the bootstrap of u and v statistics. *The Annals of Statistics*, pages 655–674, 1992.
- Mohammad Taha Bahadori and David Heckerman. Debiasing concept-based explanations with causal analysis. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL <https://openreview.net/forum?id=6puUoArESGp>.
- Vladimir Balayan, Pedro Saleiro, Catarina Belém, Ludwig Krippahl, and Pedro Bizarro. Teaching the machine to explain itself using domain knowledge. *arXiv preprint arXiv:2012.01932*, 2020.
- Catarina Belém, Vladimir Balayan, Pedro Saleiro, and Pedro Bizarro. Weakly supervised multi-task learning for concept-based explainability. *arXiv preprint arXiv:2104.12459*, 2021.
- James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.
- Leopoldo Bertossi. An asp-based approach to counterfactual explanations for classification. In *Rules and Reasoning: 4th International Joint Conference, RuleML+ RR 2020, Oslo, Norway, June 29–July 1, 2020, Proceedings 4*, pages 70–81. Springer, 2020.
- Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- Nadia Burkart and Marco F Huber. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317, 2021.
- Aditya Chatopadhyay, Piyushi Manupriya, Anirban Sarkar, and Vineeth N Balasubramanian. Neural network attributions: A causal perspective. In *International Conference on Machine Learning*, pages 981–990. PMLR, 2019.
- Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems*, 33:1229–1239, 2020.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *International Conference on Machine Learning*, pages 7324–7338. PMLR, 2022.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32, 2019.

Shantanu Ghosh, Ke Yu, Forough Arabshahi, and Kayhan Batmanghelich. Dividing and conquering a blackbox to a mixture of interpretable models: Route, interpret, repeat. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 11360–11397. PMLR, 2023a. URL <https://proceedings.mlr.press/v202/ghosh23c.html>.

Shantanu Ghosh, Ke Yu, Forough Arabshahi, and kayhan Batmanghelich. Bridging the gap: From post hoc explanations to inherently interpretable models for medical imaging. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*, 2023b. URL <https://openreview.net/forum?id=sbOVNiSNY8>.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34, 2021.

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

Marton Havasi, Sonali Parbhoo, and Finale Doshi-Velez. Addressing leakage in concept bottleneck models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 23386–23397. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/944ecf65a46feb578a43abfd5cddd960-Paper-Conference.pdf.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on artificial intelligence and statistics*, pages 2907–2916. PMLR, 2020.

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

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

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 353–362, 2021.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

Aria Khademi and Vasant G. Honavar. A causal lens for peeking into black box predictive models: Predictive model interpretation via causal attribution. *CoRR*, abs/2008.00357, 2020. URL <https://arxiv.org/abs/2008.00357>.

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

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020.

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

Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do concept bottleneck models learn as intended? *CoRR*, abs/2105.04289, 2021. URL <https://arxiv.org/abs/2105.04289>.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Tuomas P. Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL <https://openreview.net/pdf?id=F1Cg47MNvBA>.

Matthew O’Shaughnessy, Gregory Canal, Marissa Connor, Christopher Rozell, and Mark Davenport. Generative causal explanations of black-box classifiers. *Advances in Neural Information Processing Systems*, 33:5453–5467, 2020.

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

Judea Pearl. Theoretical impediments to machine learning with seven sparks from the causal revolution. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 3–3, 2018.

Sergei Popov, Stanislav Morozov, and Artem Babenko. Neural oblivious decision ensembles for deep learning on tabular data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL <https://openreview.net/forum?id=rleiuv2Vtwh>.

Numair Sani, Daniel Malinsky, and Ilya Shpitser. Explaining the behavior of black-box prediction algorithms with causal learning. *CoRR*, abs/2006.02482, 2020. URL <https://arxiv.org/abs/2006.02482>.

Patrick Schwab and Walter Karlen. Cexplain: Causal explanations for model interpretation under uncertainty. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *8th ICML Workshop on Automated Machine Learning (AutoML)*, 2021.

João Pedro Bento Sousa, Ricardo Miguel de Oliveira Moreira, Vladimir Balayan, Pedro Saleiro, and Pedro Bizarro. Conceptdistil: Model-agnostic distillation of concept explanations. In *ICLR 2022 Workshop on PAIR \textasciicircum{} 2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*, 2022.

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

Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like dags? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4):1–36, 2022.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

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

Shirley Wu, Mert Yüksekgönül, Linjun Zhang, and James Zou. Discover and cure: Concept-aware mitigation of spurious correlation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 37765–37786. PMLR, 2023. URL <https://proceedings.mlr.press/v202/wu23w.html>.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 19187–19197. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01839. URL <https://doi.org/10.1109/CVPR52729.2023.01839>.

Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in Neural Information Processing Systems*, 33:20554–20565, 2020.

Mert Yüksekgönül, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *CoRR*, abs/2205.15480, 2022. doi: 10.48550/arXiv.2205.15480. URL <https://doi.org/10.48550/arXiv.2205.15480>.

Qingyuan Zhao and Trevor Hastie. Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, 39(1):272–281, 2021.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Appendix A. Reproducibility Notes

This section provides further details regarding the experiments developed for DiConStruct. Readers can find the code of the experiments of the DiConStruct variants and single task baselines in the folder `experiments` of the supplementary materials³. The results for causal discovery and the code can also be found in the folder `experiments`. Finally, the results of the experiments as well as the statistical analysis can be found in the `results` folder of the supplementary materials³.

Appendix B. Causality Definitions

Following Pearl (2009), a structural causal model (SCM) is defined by a set of $\{x_k\}_{k=1}^K$ endogenous variables, a set of $\{u_k\}_{k=1}^K$ exogeneous variables and a set of deterministic structural functions $\{f_k\}_{k=1}^K$, such that $x_k = f_k(\mathbf{PA}_k, u_k)$. Under this formulation, the endogenous variables represent the observable random variables (the concept definitions in our system) and the exogeneous variables represent stochastic unobserved or noise variables. The functions f_k produce structural assignments that depend exclusively on \mathbf{PA}_k (the causal parents of x_k) and on the exogeneous variable u_k . For causal sufficiency (Pearl, 2009), the set of exogeneous variables $\{u_k\}_{k=1}^K$ are assumed to be jointly independent, since otherwise there would unobserved variables not captured by the SCM that are causing their dependence, making it unfeasible to perform accurate causal inferences.

A SCM is associated with a directed acyclic graph (DAG) $G = (V, E)$, with the endogeneous variables as the vertices $V = \{x_k\}_{k=1}^K$, and directed edges E denoting direct causation, defined through the functions f_k . The causal parents \mathbf{PA}_k of a given endogeneous variable x_k are connected in the graph through a directed edge $\mathbf{PA}_k \rightarrow x_k$.

The operator $do(X := a)$ simulates physical interventions by replacing certain functions from the SCM \mathcal{C} with constant values, resulting in a new model \mathcal{C}_a , such that $P_{do(X:=x)}^{\mathcal{C}}(Y) = P^{\mathcal{C}_a}(Y)$. Such interventions are relevant to identify causal effects and perform counterfactual analysis.

Appendix C. A Causal view of Concept explainability

The goal of any explainability system is to provide insights into what information the ML model learned during training. In the case of local explanations, this goal extends to also explaining the decision process followed by the ML model. Similarly to previous works (Sani et al., 2020; Janzing et al., 2020; Chattopadhyay et al., 2019; Wang et al., 2021; Jung et al., 2022; Zhao and Hastie, 2021), we view that decision process as a complex cause-effect mechanism, $X \rightarrow \hat{Y}$. This interpretation detaches the problem of explainability from the problem of modeling the *real* posterior probability of the label Y given the features X , $P(Y|X)$. We argue that the former is a more well-behaved problem for two reasons: (1) we have a solid causal direction, since we know for a fact that X is a cause for \hat{Y} and (2) X is also the *only* possible cause for \hat{Y} , excluding the possibility of unobserved external factors that could act as confounders.

However, modeling $X \rightarrow \hat{Y}$ as a causal mechanism can still be challenging, considering that X is typically of high dimensionality. Manual or automated causal discovery in such a scenario has a low chance of obtaining a correct causal graph (Vowels et al., 2022). By recentering explainability around concepts C instead of the input features X , such a problem is naturally sidestepped, since

3. <https://anonymous.4open.science/r/causal-diconstruct-4F18/>

the dimensionality of C is normally one or two orders of magnitude lower than the dimensionality⁴ of X . Causal discovery in this scenario becomes more tractable, increasing the chance of obtaining a causally correct structural graph. In the same vein of what is proposed in (Sani et al., 2020), we model the causal mechanism $C \rightarrow \hat{Y}$ using a causal graph. The possible interactions between the different concepts in C are also described in that causal graph, allowing for the direct and indirect causal effects to be used as concept attributions.

Appendix D. Hyperparameter Details

For each of the 48 experiments of the DiConStruct, and also for the concept bottleneck experiments, we performed a random hyperparameter search of 50 iterations. For the remaining models, we perform a separate random hyperparameter search of 200 iterations for each available combination of baseline, black-box model and dataset. For all models, the parameters varied were: (1) number of layers and layer dimensions for all feedforward blocks existing in the architecture⁵, (2) dropout probabilities of the dropout layers (3) learning rate, (4) L2 weight decay, (5) use of batch normalization layers, (6) batch size, and (7) categorical feature embedding dimensions (Merchant Fraud dataset only). For the CUB image dataset, we used as the inputs a feature representation of dimension 512 given by the last 8-th layer of a pre-trained ResNet-34 (He et al., 2016) model.

Additionally, for the experiments with the DiConStruct explainer, we also vary: (8) learning rate of the independence discriminator, (9) the number of epochs to update the independence discriminator, (10) use of the black-box score for the concept distillation SCM, and (11) the existence of biases on the concept distillation SCM

Appendix E. Additional Results

E.1. Concept Attribution Diversity Results

To assess the diversity of the concept attributions given an evaluation set of M instances, $\mathcal{S} = \{\{\text{TCE}^{(i)}(C_k := 0/1)\}_{k=1}^K\}_{i=1}^M \in [-1, 1]^K\}_{i=1}^M$, we compute the following metric:

$$\mathfrak{D}(a) = \frac{1}{KM} \sum_{k=1}^K \sum_{i=1}^M (\text{TCE}^{(i)}(C_k := a) - \mu_k^{\text{TCE}}(a))^2, \quad (10)$$

where $\mu_k^{\text{TCE}}(a) = \frac{1}{M} \sum_{i=1}^M \text{TCE}^{(i)}(C_k := a)$ corresponds to the average TCE of concept C_k over the evaluation dataset.

To understand how different the local concept attribution can be between explainers, we measured both diversity metrics (Equation 10) $\mathfrak{D}(0)$ and $\mathfrak{D}(1)$, on the concept attributions generated for the validation and test datasets. Figure 5 shows the distributions of $\text{TCE}^{(i)}(C_k := 0)$ attributions on the test set for two high performing⁶ explainers of the Merchant Fraud feedforward NN black-box. The explainer on the left side has low concept attribution diversity ($\mathfrak{D}(0) = 0.04$) while the one on the right has high diversity ($\mathfrak{D}(0) = 0.23$). For the explainer with low concept attribution diversity we can observe that, for each individual concept, the attributions are distributed around one central

4. When this is not the case, the humans can always reduce the concept definition (by removing or grouping concepts) to a more digestible number of concepts.
 5. Common (L), concept-specific (N), independence discriminator, and local weighting functions.
 6. We considered explainers with concept accuracy higher than 73% and \mathfrak{F}_{MAE} higher than 90%.

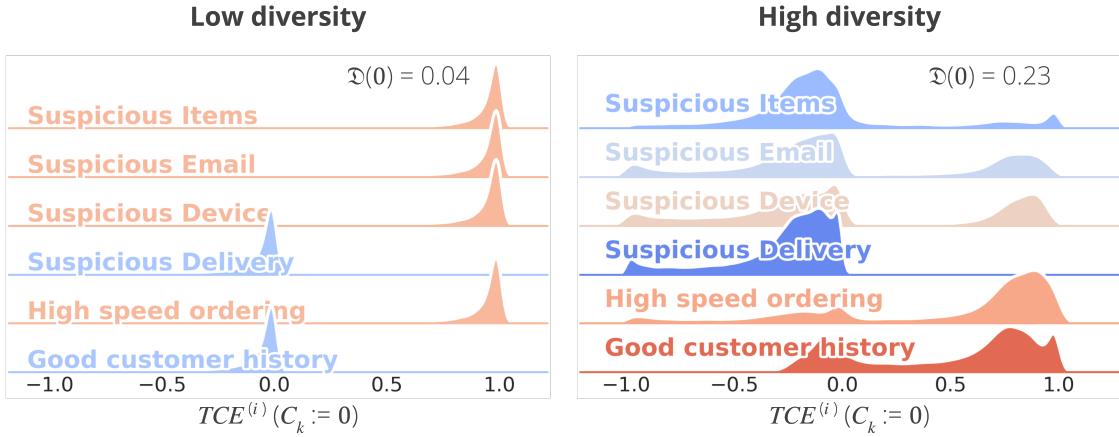


Figure 5: Concept attribution densities for two DiConStruct explainers trained with the Trivial DAG and evaluated on the Merchant Fraud test set.

Table 2: Concept attribution diversities for the selection of best results per DAG for the CUB-200-2011 feedforward NN black-box and the Merchant Fraud feedforward NN black-box. Validation and test results are shown for both $\mathfrak{D}(0)$ and $\mathfrak{D}(1)$ diversity metrics. Joint Concept Bottlenecks Models are added for comparison.

Variant	Best DAG	Validation		Test	
		$\mathfrak{D}(0)$ (%)	$\mathfrak{D}(1)$ (%)	$\mathfrak{D}(0)$ (%)	$\mathfrak{D}(1)$ (%)
CUB-200-2011					
Global	NO TEARS	4.94 ± 0.63	4.94 ± 0.63	4.36 ± 0.61	4.36 ± 0.61
Global w/ Ind.	ICA-LiNGAM	4.46 ± 0.64	4.45 ± 0.64	3.84 ± 0.61	3.84 ± 0.61
Local	ICA-LiNGAM	14.52 ± 3.79	8.77 ± 3.07	12.14 ± 3.66	7.62 ± 2.97
Local w/ Ind.	PC	17.05 ± 5.38	16.24 ± 3.37	15.12 ± 5.19	14.62 ± 3.42
Joint CBM ($\lambda = 1$)		4.26 ± 0.67	4.26 ± 0.67	3.74 ± 0.63	3.74 ± 0.63
Merchant Fraud - NN					
Global	NO TEARS	5.77 ± 0.58	5.77 ± 0.58	6.2 ± 0.64	6.2 ± 0.64
Global w/ Ind.	ICA-LiNGAM	6.01 ± 0.26	6.01 ± 0.26	6.5 ± 0.28	6.5 ± 0.28
Local	PC	7.89 ± 2.21	30.81 ± 4.55	8.66 ± 2.44	33.08 ± 4.88
Local w/ Ind.	PC	7.02 ± 1.9	32.06 ± 5.16	7.69 ± 1.87	34.68 ± 5.79
Joint CBM ($\lambda = 1$)		5.91 ± 0.2	5.91 ± 0.2	6.39 ± 0.32	6.39 ± 0.32

value with low variance, while for the high diversity explainer we have attributions distributed over $[-1, 1]$. This result indicates that some explainers may provide very rigid attributions that are close to global explanations. This justifies the need for measuring the concept attribution diversity in our results, we must assess the true locality of the different explainers.

Table 2 shows the results for the diversity metrics on the same selection of variants shown in table 1⁷. For computing the local concept attributions $TCE(C_k := 0/1)$ on CBMs, we take the outputs of $g(x)$, intervene on each concept individually ($do(\hat{c}_k := 0/1)$), and calculate the change on the output $f(g(x)_{do(\hat{c}_k:=0/1)}) - f(g(x))$. We can see that the local DiConStruct variant consistently achieves higher concept attribution diversity than the global variants and the CBMs baselines. The complete set of results of each variant and DAG combination is provided in Appendix E.2

7. This selection considers the best explainers in terms of performance, not attribution diversity.

E.2. Complete Set of Results

Table 3: Results for CUB-200-2011 and Merchant Fraud feedforward NN black-box models. The DAGs with the best results, selected for table 1, are underlined.

Variant	DAG	Validation		Test	
		\mathfrak{F}_{MAE} (%)	Concept Acc. (%)	\mathfrak{F}_{MAE} (%)	Concept Acc. (%)
CUB-200-2011					
Global	ICA-LiNGAM	92.96 ± 0.54	75.44 ± 0.46	93.88 ± 0.59	75.21 ± 0.48
	<u>NO TEARS</u>	93.52 ± 0.77	75.58 ± 0.65	94.3 ± 0.8	75.44 ± 0.44
	PC	92.1 ± 0.48	75.17 ± 0.66	92.84 ± 0.51	74.99 ± 0.61
	Trivial	93.05 ± 0.51	75.61 ± 0.52	93.76 ± 0.43	75.66 ± 0.48
Global w/ Ind.	<u>ICA-LiNGAM</u>	93.16 ± 0.75	75.61 ± 0.69	93.83 ± 0.64	75.43 ± 0.65
	NO TEARS	93.14 ± 0.57	75.21 ± 0.59	93.96 ± 0.56	75.03 ± 0.73
	PC	92.45 ± 0.34	75.51 ± 0.66	93.44 ± 0.34	75.42 ± 0.67
	Trivial	92.62 ± 0.46	75.91 ± 0.53	93.44 ± 0.59	75.86 ± 0.57
Local	<u>ICA-LiNGAM</u>	98.72 ± 0.79	75.25 ± 0.76	98.79 ± 0.74	75.11 ± 0.63
	NO TEARS	98.65 ± 0.98	74.55 ± 0.38	98.73 ± 0.92	74.22 ± 0.48
	PC	98.31 ± 0.88	74.8 ± 0.85	98.37 ± 0.85	74.71 ± 0.92
	Trivial	98.53 ± 0.71	75.43 ± 1.08	98.58 ± 0.7	75.33 ± 1.16
Local w/ Ind.	ICA-LiNGAM	98.61 ± 0.73	74.69 ± 0.73	98.66 ± 0.69	74.79 ± 0.89
	NO TEARS	97.92 ± 0.99	74.27 ± 0.75	97.99 ± 1.02	74.15 ± 0.92
	PC	98.78 ± 0.86	75.05 ± 1.08	98.83 ± 0.8	74.89 ± 1.19
	Trivial	98.4 ± 0.7	75.22 ± 0.89	98.44 ± 0.7	74.91 ± 0.67
Joint CBM ($\lambda = 1$)		-	75.48 ± 0.53	-	75.52 ± 0.59
Single task - Fidelity		96.07 ± 0.49	-	96.33 ± 0.26	-
Single task - Concept Perf.		-	76.11 ± 0.21	-	76.07 ± 0.26
Merchant Fraud - NN					
Global	ICA-LiNGAM	97.09 ± 0.22	82.58 ± 0.13	96.41 ± 0.44	82.53 ± 0.11
	<u>NO TEARS</u>	97.12 ± 0.29	82.64 ± 0.14	96.62 ± 0.28	82.58 ± 0.12
	PC	96.65 ± 0.2	82.73 ± 0.17	96.06 ± 0.23	82.68 ± 0.16
	Trivial	96.91 ± 0.29	82.48 ± 0.24	96.37 ± 0.27	82.43 ± 0.23
Global w/ Ind.	<u>ICA-LiNGAM</u>	96.96 ± 0.13	82.6 ± 0.11	96.45 ± 0.24	82.55 ± 0.09
	NO TEARS	96.88 ± 0.12	82.52 ± 0.11	96.4 ± 0.21	82.47 ± 0.09
	PC	96.73 ± 0.27	82.6 ± 0.12	95.93 ± 0.29	82.55 ± 0.09
	Trivial	96.75 ± 0.23	82.58 ± 0.13	96.25 ± 0.22	82.53 ± 0.11
Local	ICA-LiNGAM	99.48 ± 0.38	82.27 ± 0.26	99.4 ± 0.44	82.21 ± 0.25
	NO TEARS	99.29 ± 0.41	82.47 ± 0.18	99.08 ± 0.47	82.41 ± 0.17
	PC	99.39 ± 0.37	82.5 ± 0.14	99.27 ± 0.42	82.45 ± 0.13
	Trivial	99.48 ± 0.42	82.36 ± 0.16	99.37 ± 0.48	82.31 ± 0.14
Local w/ Ind.	ICA-LiNGAM	99.34 ± 0.45	82.41 ± 0.13	99.22 ± 0.51	82.36 ± 0.12
	NO TEARS	99.33 ± 0.42	82.5 ± 0.13	99.1 ± 0.52	82.45 ± 0.11
	PC	99.34 ± 0.41	82.47 ± 0.13	99.23 ± 0.49	82.42 ± 0.12
	Trivial	99.42 ± 0.42	82.4 ± 0.17	99.27 ± 0.45	82.35 ± 0.15
Joint CBM ($\lambda = 1$)		-	82.62 ± 0.13	-	82.57 ± 0.12
Single task - Fidelity		98.13 ± 0.22	-	97.86 ± 0.23	-
Single task - Concept Perf.		-	82.25 ± 0.19	-	82.25 ± 0.19

Table 4: Concept attribution diversities for the CUB-200-2011 feedforward NN black-box and the Merchant Fraud feedforward NN black-box. The DAGs with the best results, selected for table 1, are underlined.

Variant	DAG	Validation		Test	
		$\mathfrak{D}(0)$ (%)	$\mathfrak{D}(1)$ (%)	$\mathfrak{D}(0)$ (%)	$\mathfrak{D}(1)$ (%)
CUB-200-2011					
Global	ICA-LiNGAM	4.34 ± 0.57	4.34 ± 0.57	3.75 ± 0.48	3.75 ± 0.48
	<u>NO TEARS</u>	4.94 ± 0.63	4.94 ± 0.63	4.36 ± 0.61	4.36 ± 0.61
	PC	4.03 ± 0.58	4.03 ± 0.58	3.55 ± 0.55	3.55 ± 0.55
	Trivial	4.41 ± 0.64	4.41 ± 0.64	3.85 ± 0.59	3.85 ± 0.59
Global w/ Ind.	<u>ICA-LiNGAM</u>	4.46 ± 0.64	4.45 ± 0.64	3.84 ± 0.61	3.84 ± 0.61
	NO TEARS	4.97 ± 0.5	4.97 ± 0.5	4.31 ± 0.43	4.31 ± 0.43
	PC	4.22 ± 0.26	4.22 ± 0.26	3.73 ± 0.25	3.73 ± 0.25
	Trivial	4.55 ± 0.69	4.54 ± 0.69	3.95 ± 0.66	3.94 ± 0.66
Local	<u>ICA-LiNGAM</u>	14.52 ± 3.79	8.77 ± 3.07	12.14 ± 3.66	7.62 ± 2.97
	NO TEARS	15.12 ± 9.94	16.17 ± 3.97	13.5 ± 8.94	14.67 ± 3.99
	PC	20.16 ± 5.97	22.27 ± 5.81	18.21 ± 5.28	20.17 ± 4.96
	Trivial	15.6 ± 5.13	17.07 ± 3.0	14.06 ± 5.06	15.41 ± 3.14
Local w/ Ind.	ICA-LiNGAM	16.74 ± 4.57	10.18 ± 4.41	14.46 ± 4.38	8.96 ± 4.11
	NO TEARS	20.88 ± 11.18	18.39 ± 5.58	19.03 ± 10.65	17.5 ± 6.28
	<u>PC</u>	17.05 ± 5.38	16.24 ± 3.37	15.12 ± 5.19	14.62 ± 3.42
	Trivial	19.42 ± 6.11	19.76 ± 3.41	17.85 ± 5.8	18.35 ± 3.74
Joint CBM ($\lambda = 1$)		4.26 ± 0.67	4.26 ± 0.67	3.74 ± 0.63	3.74 ± 0.63
Merchant Fraud - NN					
Global	ICA-LiNGAM	6.07 ± 0.33	6.07 ± 0.34	6.48 ± 0.36	6.48 ± 0.36
	<u>NO TEARS</u>	5.77 ± 0.58	5.77 ± 0.57	6.19 ± 0.64	6.2 ± 0.64
	PC	6.01 ± 0.16	6.01 ± 0.16	6.41 ± 0.21	6.41 ± 0.21
	Trivial	6.15 ± 0.2	6.15 ± 0.2	6.69 ± 0.32	6.69 ± 0.32
Global w/ Ind.	<u>ICA-LiNGAM</u>	6.01 ± 0.26	6.01 ± 0.26	6.5 ± 0.28	6.5 ± 0.28
	NO TEARS	5.78 ± 0.58	5.78 ± 0.59	6.21 ± 0.62	6.2 ± 0.62
	PC	6.17 ± 0.22	6.17 ± 0.22	6.65 ± 0.28	6.65 ± 0.28
	Trivial	6.08 ± 0.15	6.08 ± 0.15	6.68 ± 0.18	6.68 ± 0.18
Local	ICA-LiNGAM	9.33 ± 2.02	31.62 ± 2.75	10.3 ± 2.23	33.8 ± 2.7
	NO TEARS	12.04 ± 4.86	30.91 ± 8.49	12.87 ± 5.28	32.81 ± 8.49
	<u>PC</u>	7.89 ± 2.21	30.81 ± 4.55	8.66 ± 2.44	33.08 ± 4.88
	Trivial	7.15 ± 1.59	28.38 ± 5.99	7.71 ± 1.58	30.6 ± 6.2
Local w/ Ind.	ICA-LiNGAM	8.64 ± 2.45	30.38 ± 3.22	9.47 ± 2.46	32.47 ± 3.75
	NO TEARS	11.57 ± 3.09	24.95 ± 6.44	13.06 ± 3.31	26.7 ± 6.85
	<u>PC</u>	7.02 ± 1.9	32.06 ± 5.16	7.69 ± 1.87	34.68 ± 5.79
	Trivial	7.67 ± 1.76	31.21 ± 10.76	8.17 ± 1.87	33.51 ± 11.23
Joint CBM ($\lambda = 1$)		5.91 ± 0.2	5.91 ± 0.2	6.39 ± 0.32	6.39 ± 0.32

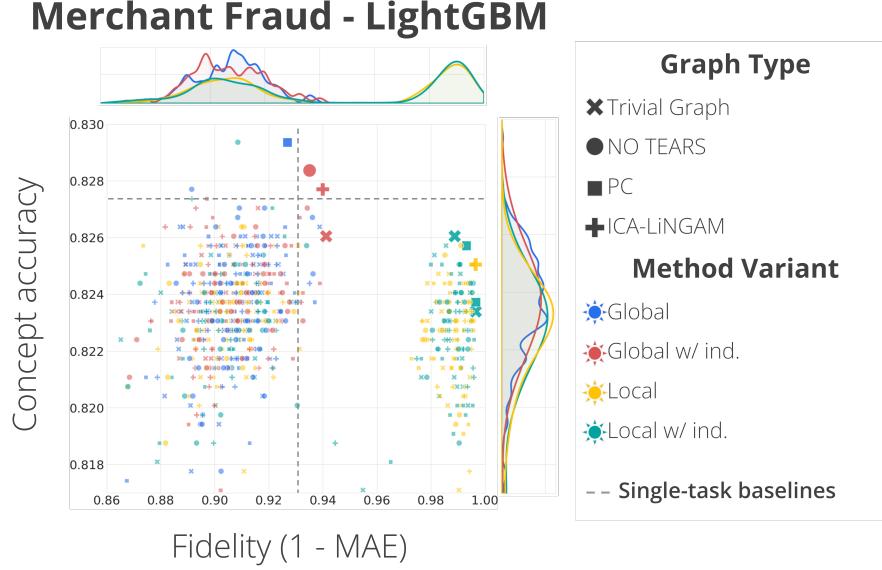


Figure 6: Distribution of fidelity and concept accuracy for the Merchant Fraud LightGBM hyper-parameter experiments. The plot shows the result of 16 experiments of 50 trials, and each marker is a trained DiConStruct explainer. Models on the Pareto efficiency frontier are highlighted with bigger markers.

E.3. Results for Merchant Fraud LightGBM black-box

The state of the art for classification tasks on tabular data still belongs to the gradient boosted decision tree models (Borisov et al., 2022; Shwartz-Ziv and Armon, 2021; Gorishniy et al., 2021; Popov et al., 2020), like LightGBM (Ke et al., 2017). For that reason, we included that as a second black-box classifier in our Merchant Fraud dataset experiments. Table 5 reports the results of those experiments. The fidelity decreases for Global methods (and single task baseline), when compared to the feedforward NN black-box experiments. We attribute this fidelity decrease in the difference in inductive bias between LightGBM and NN-based models, making the former perform better at the main classification task⁸. Intuitively, this means that LightGBM black-box learns more information from X than the NN. In terms of concept prediction performance, we can see that the DiConStruct explainer keeps similar performance to the single task baseline.

Like for the other black-box models, Figure 6 shows that the DiConStruct explainer outperforms both the knowledge distillation and the explainability single task baselines. For specific configurations, we see that the explainer is better in both tasks than the single task baselines. There is a cluster of local explainers that clearly stands out on the fidelity x-axis. Like with the other black-box experiments, this happens to be a set of explainers configured to use the score of the black-box as input for the weighting functions, $W_{j,k}(\cdot)$, resulting in a much better approximation of the black-box model score.

8. LightGBM performance was 68.31% recall@5%FPR against 63.35% recall@5%FRP of the fraud NN.

Table 5: Main task classification, fidelity and concept prediction performance of DiConStruct explainer for the Merchant Fraud LightGBM black-box. Joint Concept Bottlenecks Models and single task baselines are added for comparison.

Variant	DAG	Validation			Test		
		Task Perf. (%)	Concept Perf. (%)	\mathfrak{F}_{MAE} (%)	Task Perf. (%)	Concept Perf. (%)	\mathfrak{F}_{MAE} (%)
Merchant Fraud - LightGBM							
Global	ICA-LiNGAM	78.36	82.6 ± 0.12	93.42 ± 0.18	68.31	82.55 ± 0.1	91.98 ± 1.45
	NO TEARS	78.36	82.7 ± 0.23	93.54 ± 0.18	68.31	82.65 ± 0.22	92.21 ± 0.73
	PC	78.36	82.65 ± 0.14	93.57 ± 0.32	68.31	82.6 ± 0.12	92.48 ± 1.34
	Trivial	78.36	82.52 ± 0.11	93.62 ± 0.2	68.31	82.46 ± 0.08	92.17 ± 1.17
Global w/ Ind.	ICA-LiNGAM	78.36	82.57 ± 0.12	93.53 ± 0.34	68.31	82.52 ± 0.1	92.56 ± 1.1
	NO TEARS	78.36	82.56 ± 0.16	93.67 ± 0.27	68.31	82.51 ± 0.14	93.01 ± 0.62
	PC	78.36	82.66 ± 0.21	93.53 ± 0.17	68.31	82.61 ± 0.2	92.92 ± 0.79
	Trivial	78.36	82.62 ± 0.21	93.38 ± 0.33	68.31	82.57 ± 0.2	91.71 ± 2.19
Local	ICA-LiNGAM	78.36	82.43 ± 0.16	99.44 ± 0.18	68.31	82.38 ± 0.15	99.42 ± 0.19
	NO TEARS	78.36	82.33 ± 0.12	99.45 ± 0.18	68.31	82.28 ± 0.1	99.42 ± 0.2
	PC	78.36	82.4 ± 0.12	99.47 ± 0.16	68.31	82.34 ± 0.1	99.4 ± 0.18
	Trivial	78.36	82.43 ± 0.13	99.48 ± 0.12	68.31	82.38 ± 0.11	99.47 ± 0.15
Local w/ Ind.	ICA-LiNGAM	78.36	82.42 ± 0.13	99.41 ± 0.25	68.31	82.36 ± 0.12	99.37 ± 0.26
	NO TEARS	78.36	82.41 ± 0.17	99.48 ± 0.21	68.31	82.36 ± 0.15	99.45 ± 0.22
	PC	78.36	82.5 ± 0.09	99.52 ± 0.15	68.31	82.45 ± 0.07	99.46 ± 0.15
	Trivial	78.36	82.35 ± 0.11	99.5 ± 0.17	68.31	82.3 ± 0.08	99.31 ± 0.22
Joint CBM ($\lambda = 1$)		48.42 ± 0.31	82.62 ± 0.13	-	47.47 ± 3.64	82.57 ± 0.12	-
Single task - Fidelity		-	-	93.65 ± 0.31	-	-	90.75 ± 1.19
Single task - Task Perf.		78.36	-	-	68.31	-	-
Single task - Concept Perf.		-	82.25 ± 0.19	-	-	82.25 ± 0.19	-

Appendix F. Concept Labels

F.1. CUB-200-2011

We used the 28 bird attribute groupings as our concept labels. As described in subsection 3.1, since the original dataset contains multiple possible values per attribute group, they were binarized by assigning 1 to the most common value in the group and 0 otherwise. This resulted in 28 different concepts. To maximize the predictability of our concepts to the black-box score, we selected the top 7 concepts that correlated the most with the black-box model score. We used ROC AUC score between the binary attributes and the continuous score to measure this correlation. The selected list of concepts and corresponding prevalence are shown in table 6.

Table 6: Concepts for the CUB-200-2011 dataset and corresponding prevalence obtained globally and when conditioning on the classification labels.

Concept	Prevalences (%)		
	Global	<i>Warbler</i>	$\neg\text{Warbler}$
Shape: perching-like	48.68	76.85	44.60
Bill Length: shorter than head	56.11	75.44	53.31
Bill Shape: all-purpose	39.18	73.22	34.25
Eye Color: black	83.75	91.54	82.62
Nape Color: black	26.48	18.86	27.58
Primary Color: black	35.54	28.52	36.56
Back Color: black	28.60	23.02	29.40

F.2. Merchant Fraud Dataset

To train our method to predict concepts on the full and sampled datasets, we employed a weak supervision methodology. This consisted of creating a set of *Concept Teachers* of the form $f_{CT} : \mathbb{X} \rightarrow \mathbb{Y}_C$ trained using a small golden set of 1934 transactions labeled with concept labels in a procedure described in appendix section F.3. The *Concept Teachers* were then used to infer the probabilistic labels $p(\mathbf{c}^{(i)} | \mathbf{x}^{(i)}) \in [0, 1]$. Our explainability models were trained using these probabilistic labels as concept targets. Training with probabilistic concept targets requires no change of implementation because the BCE loss is equally differentiable for continuous targets $\in [0, 1]$.

F.3. Merchant Fraud Dataset Golden Concept Labels

To evaluate our explainability task, we randomly chose a small set from the sampled dataset, and presented those transactions to a group of human experts (fraud analysts with the knowledge of the most common fraud patterns latent on this online retailer dataset). We then asked the fraud analysts to annotate the existing concepts on the collection of 2643 transactions. These were further split into 1934 for training the *Concept Teachers*, 203 for the optimization of the teachers' parameters, and 506 for test. This last set was the one used to evaluate the explainability performance of the *Concept Teachers* and all DiConStruct models and baselines presented. Despite the requirement for having concept labels being a limitation in this work, we show how this can be mitigated by

using a weak supervision strategy. Indeed, we reduced the labelling effort by 3 orders of magnitude, when compared to the effort of fully labelling the sampled dataset of around 500k transactions. Additionally, we consider that concept explainability can benefit from using ground truth concepts obtained from expert knowledge instead of unsupervised methodologies, since the latter can more easily result in meaningless concept definitions.

To provide more context about the concepts considered for the Merchant Fraud, we show in Table 7 the list of concepts as well as their prevalence on the golden dataset (Concept Teachers training data plus the golden test set).

Appendix G. Causal discovery options

G.1. From Expert Knowledge

As a first option, we propose that an iterative work is done together with domain experts. The domain experts should be the same personas involved in concept definition so that there is as familiarity with the problem and an agreement that those are the necessary concepts. Before the procedure starts, a benchmark for the maximum attainable concept prediction accuracy should be created. The procedure is as follows:

- An informed guess of \mathcal{G}_{C,Y_B} is drawn by the domain experts, then the DiConStruct explainer is trained using that graph, and its quality is evaluated;
- If the distillation fidelity is close to 100%, the concept prediction accuracy is similar or better than the benchmark and the exogenous factors are jointly independent, the procedure can finish, and the graph structure is saved;
- If the distillation fidelity is low, but the concept prediction accuracy is good, the domain experts can try to add more direct parents to Y_B , or even remove the causal links $C_k \rightarrow Y_B$ with minimal weights. If after a few iterations on the graph structure, the distillation fidelity keeps low, the problem may be that the concept definition is not *complete*. In such case, the domain experts should assess which new concepts could be added to the concept definition.
- In case that both the distillation fidelity and the concept prediction accuracy (compared to the benchmark) are low, the graph structure might be fundamentally incorrect and/or the concept definition is very incomplete. In this scenario, the domain experts can try to redefine \mathcal{G}_{C,Y_B} from scratch or remove the causal links with lower weights and add more direct parents to both the concepts with less prediction performance and to Y_B ;
- If the exogenous independence is not being achieved, and both the distillation fidelity and the concept prediction accuracy are high, it might indicate that some confounding factors are being disregarded in the causal graph. In this scenario, despite the *completeness* and concept prediction accuracy being satisfied, the causal story is incomplete, resulting in less reliable explanations and concept attributions. Therefore, domain experts should assess which new concepts could be added to the concept definition;

G.2. Causal Discovery

In case that domain experts are not available, we propose to use a data-driven causal discovery method for obtaining \mathcal{G}_{C,Y_B} . This discovery should be constraint structures where Y_B is not causing

any of C . The dataset used for causal discovery can be the same that will then be used for training the DiConStruct explainer.

Table 7: Concepts for the Merchant Fraud dataset and corresponding prevalence on the golden dataset. The table shows the prevalence obtained globally and when conditioning on the classification labels.

Concept	Prevalences (%)		
	Global	Legit	Fraud
Good Customer History	24.45	27.85	14.29
High-Speed Ordering	11.33	8.49	19.84
Suspicious Delivery	22.86	20.16	30.95
Suspicious Device	11.73	8.75	20.63
Suspicious Email	21.07	18.93	30.16
Suspicious Items	18.49	17.24	22.22

G.3. Merchant Fraud Concept Teachers

To train the *Concept Teachers* we used the golden dataset enriched not only with the fraud features available at run-time but also with some additional information, such as fraud rule triggers and manual decisions, which is only generated during or after the historical fraud decision. The algorithm used for the *Concept Teachers* was a Random Forest classifier and the hyperparameters were optimized using a Tree-structured Parzen Estimator (TPE) ([Bergstra et al., 2011](#)) algorithm with 200 trials from which 30 trails were used for a random initialization. As the optimization metric, we used ROC AUC on the golden validation set. We then evaluated the *Concept Teachers* in the test set of our golden dataset, obtaining a mean ROC AUC of 75.72% across all concepts. Despite not being central to our work, the *Concept Teachers* performance shows an interesting result when compared to the 78.49% performance of our concept explainability baseline. This 2.8 pp improvement originates solely from the using more data in \mathbb{X} (the concept explainability baseline trained in 460k instances labeled by our *Concept Teachers*) which illustrates that our weak supervision approach is successful to some degree.

