

Faithful Model Explanations through Energy-Constrained Conformal Counterfactuals

Anonymous submission

Abstract

Counterfactual explanations offer an intuitive and straightforward way to explain black-box models and offer algorithmic recourse to individuals. To address the need for plausible explanations, existing work has primarily relied on surrogate models to learn how the input data is distributed. This effectively reallocates the task of learning realistic explanations for the data from the model itself to the surrogate. Consequently, the generated explanations may seem plausible to humans but need not necessarily describe the behaviour of the black-box model faithfully. We formalise this notion of faithfulness through the introduction of a tailored evaluation metric and propose a novel algorithmic framework for generating **Energy-Constrained Conformal Counterfactuals** that are only as plausible as the model permits. Through extensive empirical studies, we demonstrate that *ECCCo* reconciles the need for faithfulness and plausibility. In particular, we show that for models with gradient access, it is possible to achieve state-of-the-art performance without the need for surrogate models. To do so, our framework relies solely on properties defining the black-box model itself by leveraging recent advances in energy-based modelling and conformal prediction. To our knowledge, this is the first venture in this direction for generating faithful counterfactual explanations. Thus, we anticipate that *ECCCo* can serve as a baseline for future research. We believe that our work opens avenues for researchers and practitioners seeking tools to better distinguish trustworthy from unreliable models.

1 Introduction

Counterfactual explanations provide a powerful, flexible and intuitive way to not only explain black-box models but also offer the possibility of algorithmic recourse to affected individuals. Instead of opening the black box, counterfactual explanations work under the premise of strategically perturbing model inputs to understand model behaviour (Wachter, Mittelstadt, and Russell 2017). Intuitively speaking, we generate explanations in this context by asking what-if questions of the following nature: ‘Our credit risk model currently predicts that this individual is not credit-worthy. What if they reduced their monthly expenditures by 10%?’

This is typically implemented by defining a target outcome $\mathbf{y}^+ \in \mathcal{Y}$ for some individual $\mathbf{x} \in \mathcal{X} = \mathbb{R}^D$ described by D attributes, for which the model $M_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ initially predicts a different outcome: $M_\theta(\mathbf{x}) \neq \mathbf{y}^+$. Counterfactuals are then searched by minimizing a loss function that

compares the predicted model output to the target outcome: $\text{yloss}(M_\theta(\mathbf{x}), \mathbf{y}^+)$. Since counterfactual explanations work directly with the black-box model, valid counterfactuals always have full local fidelity by construction where fidelity is defined as the degree to which explanations approximate the predictions of a black-box model (Molnar 2022).

In situations where full fidelity is a requirement, counterfactual explanations offer a more appropriate solution to Explainable Artificial Intelligence (XAI) than other popular approaches like LIME (Ribeiro, Singh, and Guestrin 2016) and SHAP (Lundberg and Lee 2017), which involve local surrogate models. But even full fidelity is not a sufficient condition for ensuring that an explanation *faithfully* describes the behaviour of a model. That is because multiple distinct explanations can lead to the same model prediction, especially when dealing with heavily parameterized models like deep neural networks, which are underspecified by the data (Wilson 2020). In the context of counterfactuals, the idea that no two explanations are the same arises almost naturally. A key focus in the literature has therefore been to identify those explanations that are most appropriate based on a myriad of desiderata such as closeness (Wachter, Mittelstadt, and Russell 2017), sparsity (Schut et al. 2021), actionability (Ustun, Spangher, and Liu 2019) and plausibility (Joshi et al. 2019).

In this work, we draw closer attention to model faithfulness rather than fidelity as a desideratum for counterfactuals. We define faithfulness as the degree to which counterfactuals are consistent with what the model has learned about the data. Our key contributions are as follows: first, we show that fidelity is an insufficient evaluation metric for counterfactuals (Section 3) and propose a definition of faithfulness that gives rise to more suitable metrics (Section 4). Next, we introduce a *ECCCo*: a novel algorithmic approach aimed at generating energy-constrained conformal counterfactuals that faithfully explain model behaviour in Section 5. Finally, we provide extensive empirical evidence demonstrating that *ECCCo* faithfully explains model behaviour and attains plausibility only when appropriate (Section 6).

To our knowledge, this is the first venture in this direction for generating faithful counterfactuals. Thus, we anticipate that *ECCCo* can serve as a baseline for future research. We believe that our work opens avenues for researchers and practitioners seeking tools to better distinguish trustworthy from unreliable models.

2 Background

While counterfactual explanations (CE) can also be generated for arbitrary regression models (Spooner et al. 2021), existing work has primarily focused on classification problems. Let $\mathcal{Y} = (0, 1)^K$ denote the one-hot-encoded output domain with K classes. Then most counterfactual generators rely on gradient descent to optimize different flavours of the following counterfactual search objective:

$$\mathbf{Z}' = \arg \min_{\mathbf{Z}' \in \mathcal{Z}^L} \{ \text{yloss}(M_\theta(f(\mathbf{Z}')), \mathbf{y}^+) + \lambda \text{cost}(f(\mathbf{Z}')) \} \quad (1)$$

Here $\text{yloss}(\cdot)$ denotes the primary loss function, $f(\cdot)$ is a function that maps from the counterfactual state space to the feature space and $\text{cost}(\cdot)$ is either a single penalty or a collection of penalties that are used to impose constraints through regularization. Equation 1 restates the baseline approach to gradient-based counterfactual search proposed by Wachter, Mittelstadt, and Russell (2017) in general form as introduced by Altmeyer et al. (2023). To explicitly account for the multiplicity of explanations, $\mathbf{Z}' = \{\mathbf{z}_l\}_L$ denotes an L -dimensional array of counterfactual states.

The baseline approach, which we will simply refer to as *Wachter*, searches a single counterfactual directly in the feature space and penalises its distance to the original factual. In this case, $f(\cdot)$ is simply the identity function and \mathcal{Z} corresponds to the feature space itself. Many derivative works of Wachter, Mittelstadt, and Russell (2017) have proposed new flavours of Equation 1, each of them designed to address specific *desiderata* that counterfactuals ought to meet in order to properly serve both AI practitioners and individuals affected by algorithmic decision-making systems. The list of desiderata includes but is not limited to the following: sparsity, closeness (Wachter, Mittelstadt, and Russell 2017), actionability (Ustun, Spangher, and Liu 2019), diversity (Mothilal, Sharma, and Tan 2020), plausibility (Joshi et al. 2019; Poyiadzi et al. 2020; Schut et al. 2021), robustness (Upadhyay, Joshi, and Lakkaraju 2021; Pawelczyk et al. 2022; Altmeyer et al. 2023) and causality (Karimi, Schölkopf, and Valera 2021). Different counterfactual generators addressing these needs have been extensively surveyed and evaluated in various studies (Verma, Dickerson, and Hines 2020; Karimi et al. 2020; Pawelczyk et al. 2021; Artelt et al. 2021; Guidotti 2022).

The notion of plausibility is central to all of the desiderata. For example, Artelt et al. (2021) find that plausibility typically also leads to improved robustness. Similarly, plausibility has also been connected to causality in the sense that plausible counterfactuals respect causal relationships (Mahajan, Tan, and Sharma 2019). Consequently, the plausibility of counterfactuals has been among the primary concerns for researchers. Achieving plausibility is equivalent to ensuring that the generated counterfactuals comply with the true and unobserved data-generating process (DGP). We define plausibility formally in this work as follows:

Definition 2.1 (Plausible Counterfactuals). *Let $\mathcal{X}|\mathbf{y}^+ = p(\mathbf{x}|\mathbf{y}^+)$ denote the true conditional distribution of samples in the target class \mathbf{y}^+ . Then for \mathbf{x}' to be considered a plausible counterfactual, we need: $\mathbf{x}' \sim \mathcal{X}|\mathbf{y}^+$.*

To generate plausible counterfactuals, we first need to quantify the conditional distribution of samples in the target class ($\mathcal{X}|\mathbf{y}^+$). We can then ensure that we generate counterfactuals that comply with that distribution.

One straightforward way to do this is to use surrogate models for the task. Joshi et al. (2019), for example, suggest that instead of searching counterfactuals in the feature space \mathcal{X} , we can instead traverse a latent embedding \mathcal{Z} (Equation 1) that implicitly codifies the DGP. To learn the latent embedding, they propose using a generative model such as a Variational Autoencoder (VAE). Provided the surrogate model is well-specified, their proposed approach *REVISE* can yield plausible explanations. Others have proposed similar approaches: Dombrowski, Gerken, and Kessel (2021) traverse the base space of a normalizing flow to solve Equation 1; Poyiadzi et al. (2020) use density estimators ($\hat{p} : \mathcal{X} \mapsto [0, 1]$) to constrain the counterfactuals to dense regions in the feature space; and, finally, Karimi, Schölkopf, and Valera (2021) assume knowledge about the structural causal model that generates the data.

A competing approach towards plausibility that is also closely related to this work instead relies on the black-box model itself. Schut et al. (2021) show that to meet the plausibility objective we need not explicitly model the input distribution. Pointing to the undesirable engineering overhead induced by surrogate models, they propose that we rely on the implicit minimisation of predictive uncertainty instead. Their proposed methodology, which we will refer to as *Schut*, solves Equation 1 by greedily applying Jacobian-Based Saliency Map Attacks (JSMA) in the feature space with cross-entropy loss and no penalty at all. The authors demonstrate theoretically and empirically that their approach yields counterfactuals for which the model M_θ predicts the target label \mathbf{y}^+ with high confidence. Provided the model is well-specified, these counterfactuals are plausible. This idea hinges on the assumption that the black-box model provides well-calibrated predictive uncertainty estimates.

3 Why Fidelity is not Enough: A Motivational Example

As discussed in the introduction, any valid counterfactual also has full fidelity by construction: solutions to Equation 1 are considered valid as soon as the label predicted by the model matches the target class. So while fidelity always applies, counterfactuals that address the various desiderata introduced above can look vastly different from each other.

To demonstrate this with an example, we have trained a simple image classifier M_θ on the well-known *MNIST* dataset (LeCun 1998): a Multi-Layer Perceptron (MLP) with test set accuracy > 0.9 . No measures have been taken to improve the model's adversarial robustness or its capacity for predictive uncertainty quantification. The far left panel of Figure 1 shows a random sample drawn from the dataset. The underlying classifier correctly predicts the label ‘nine’ for this image. For the given factual image and model, we have used *Wachter*, *Schut* and *REVISE* to generate one counterfactual each in the target class ‘seven’. The perturbed images are shown next to the factual image from left to right

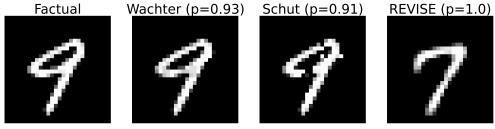


Figure 1: Counterfactuals for turning a 9 (nine) into a 7 (seven): original image (left), then the counterfactuals generated using *Wachter*, *Schut* and *REVISE*.

in Figure 1. Captions on top of the images indicate the generator along with the predicted probability that the image belongs to the target class. In all cases, that probability is very high, while the counterfactuals look very different.

Since *Wachter* is only concerned with closeness, the generated counterfactual is almost indistinguishable from the factual. The approach by *Schut* et al. (2021) expects a well-calibrated model that can generate predictive uncertainty estimates. Since this is not the case, the generated counterfactual looks like an adversarial example. Finally, the counterfactual generated by *REVISE* looks much more plausible than the other two. But is it also more faithful to the behaviour of our *MNIST* classifier? That is much less clear because the surrogate used by *REVISE* introduces friction: the generated explanations no longer depend exclusively on the black-box model itself.

So which of the counterfactuals most faithfully explains the behaviour of our image classifier? Fidelity cannot help us to make that judgement, because all of these counterfactuals have full fidelity. Thus, fidelity is an insufficient evaluation metric to assess the faithfulness of CE.

4 Faithful first, Plausible second

Considering the limitations of fidelity as demonstrated in the previous section, analogous to Definition 2.1, we introduce a new notion of faithfulness in the context of CE:

Definition 4.1 (Faithful Counterfactuals). *Let $\mathcal{X}_\theta|y^+ = p_\theta(x|y^+)$ denote the conditional distribution of x in the target class y^+ , where θ denotes the parameters of model M_θ . Then for x' to be considered a faithful counterfactual, we need: $x' \sim \mathcal{X}_\theta|y^+$.*

In doing this, we merge in and nuance the concept of plausibility (Definition 2.1) where the notion of ‘consistent with the data’ becomes ‘consistent with what the model has learned about the data’.

4.1 Quantifying the Model’s Generative Property

To assess counterfactuals with respect to Definition 4.1, we need a way to quantify the posterior conditional distribution $p_\theta(x|y^+)$. To this end, we draw on ideas from energy-based modelling (EBM), a subdomain of machine learning that is concerned with generative or hybrid modelling (Grathwohl et al. 2020; Du and Mordatch 2019). In particular, note that if we fix y to our target value y^+ , we can conditionally draw from $p_\theta(x|y^+)$ by randomly initializing x_0 and then using Stochastic Gradient Langevin Dynamics (SGLD) as follows,

$$\mathbf{x}_{j+1} \leftarrow \mathbf{x}_j - \frac{\epsilon_j^2}{2} \mathcal{E}_\theta(\mathbf{x}_j | \mathbf{y}^+) + \epsilon_j \mathbf{r}_j, \quad j = 1, \dots, J \quad (2)$$

where $\mathbf{r}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the stochastic term and the step-size ϵ_j is typically polynomially decayed (Welling and Teh 2011). The term $\mathcal{E}_\theta(\mathbf{x}_j | \mathbf{y}^+)$ denotes the model energy conditioned on the target class label \mathbf{y}^+ which we specify as the negative logit corresponding to the target class label \mathbf{y}^+ . To allow for faster sampling, we follow the common practice of choosing the step-size ϵ_j and the standard deviation of \mathbf{r}_j separately. While \mathbf{x}_J is only guaranteed to distribute as $p_\theta(x|\mathbf{y}^+)$ if $\epsilon \rightarrow 0$ and $J \rightarrow \infty$, the bias introduced for a small finite ϵ is negligible in practice (Murphy 2023).

Generating multiple samples using SGLD thus yields an empirical distribution $\hat{\mathcal{X}}_{\theta, y^+}$ that approximates what the model has learned about the input data. While in the context of EBM, this is usually done during training, we propose to repurpose this approach during inference in order to evaluate the faithfulness of model explanations. The technical appendix provides additional implementation details for any tasks related to energy-based modelling.

4.2 Quantifying the Model’s Predictive Uncertainty

Faithful counterfactuals can be expected to also be plausible if the learned conditional distribution $\mathcal{X}_\theta|y^+$ (Defintion 4.1) is close to the true conditional distribution $\mathcal{X}|y^+$ (Definition 2.1). We can further improve the plausibility of counterfactuals without the need for surrogate models that may interfere with faithfulness by minimizing predictive uncertainty (*Schut* et al. 2021). Unfortunately, this idea relies on the assumption that the model itself provides predictive uncertainty estimates, which may be too restrictive in practice.

To relax this assumption, we use conformal prediction (CP), an approach to predictive uncertainty quantification that has recently gained popularity (Angelopoulos and Bates 2021; Manokhin 2022). Crucially for our intended application, CP is model-agnostic and can be applied during inference without placing any restrictions on model training. It works under the premise of turning heuristic notions of uncertainty into rigorous estimates by repeatedly sifting through the training data or a dedicated calibration dataset.

Conformal classifiers produce prediction sets for individual inputs that include all output labels that can be reasonably attributed to the input. These sets are formed as follows,

$$C_\theta(\mathbf{x}_i; \alpha) = \{\mathbf{y} : s(\mathbf{x}_i, \mathbf{y}) \leq \hat{q}\} \quad (3)$$

where \hat{q} denotes the $(1 - \alpha)$ -quantile of \mathcal{S} and α is a pre-determined error rate. These sets tend to be larger for inputs that do not conform with the training data and are characterized by high predictive uncertainty. To leverage this notion of predictive uncertainty in the context of gradient-based counterfactual search, we use a smooth set size penalty introduced by Stutz et al. (2022):

$$\Omega(C_\theta(\mathbf{x}; \alpha)) = \max \left(0, \sum_{\mathbf{y} \in \mathcal{Y}} C_{\theta, \mathbf{y}}(\mathbf{x}; \alpha) - \kappa \right) \quad (4)$$

Here, $\kappa \in \{0, 1\}$ is a hyper-parameter and $C_{\theta, y}(\mathbf{x}_i; \alpha)$ can be interpreted as the probability of label y being included in the prediction set (see appendix for details). In order to compute this penalty for any black-box model, we merely need to perform a single calibration pass through a holdout set \mathcal{D}_{cal} . Arguably, data is typically abundant and in most applications, practitioners tend to hold out a test data set anyway. Consequently, CP removes the restriction on the family of predictive models, at the small cost of reserving a subset of the available data for calibration. This particular case of conformal prediction is referred to as *split conformal prediction* (SCP) as it involves splitting the training data into a proper training dataset and a calibration dataset.

4.3 Evaluating Plausibility and Faithfulness

The parallels between our definitions of plausibility and faithfulness imply that we can also use similar evaluation metrics in both cases. Since existing work has focused heavily on plausibility, it offers a useful starting point. In particular, Guidotti (2022) have proposed an implausibility metric that measures the distance of the counterfactual from its nearest neighbour in the target class. As this distance is reduced, counterfactuals get more plausible under the assumption that the nearest neighbour itself is plausible in the sense of Definition 2.1. In this work, we use the following adapted implausibility metric,

$$\text{impl}(\mathbf{x}', \mathbf{X}_{y^+}) = \frac{1}{|\mathbf{X}_{y^+}|} \sum_{\mathbf{x} \in \mathbf{X}_{y^+}} \text{dist}(\mathbf{x}', \mathbf{x}) \quad (5)$$

where \mathbf{x}' denotes the counterfactual and \mathbf{X}_{y^+} is a subsample of the training data in the target class y^+ . By averaging over multiple samples in this manner, we avoid the risk that the nearest neighbour of \mathbf{x}' itself is not plausible according to Definition 2.1 (e.g an outlier).

Equation 5 gives rise to a similar evaluation metric for unfaithfulness. We swap out the subsample of observed individuals in the target class for the set of samples generated through SGLD ($\widehat{\mathbf{X}}_{y^+}$):

$$\text{unfaith}(\mathbf{x}', \widehat{\mathbf{X}}_{\theta, y^+}) = \frac{1}{|\widehat{\mathbf{X}}_{\theta, y^+}|} \sum_{\mathbf{x} \in \widehat{\mathbf{X}}_{\theta, y^+}} \text{dist}(\mathbf{x}', \mathbf{x}) \quad (6)$$

Our default choice for the $\text{dist}(\cdot)$ function in both cases is the Euclidean Norm. Depending on the type of input data other choices may be more adequate, which we discuss further in Section 6.1.

5 Energy-Constrained Conformal Counterfactuals

Given our proposed notion of faithfulness, we now describe *ECCCo*, our proposed framework for generating Energy-Constrained Conformal Counterfactuals. It is based on the premise that counterfactuals should first and foremost be faithful. Plausibility, as a secondary concern, is then still attainable to the degree that the black-box model itself has learned plausible explanations for the underlying data.

We begin by substituting the loss function in Equation 1,

$$\mathbf{Z}' = \arg \min_{\mathbf{Z}' \in \mathcal{Z}^L} \{L_{\text{JEM}}(f(\mathbf{Z}'); M_\theta, \mathbf{y}^+) + \lambda \text{cost}(f(\mathbf{Z}'))\} \quad (7)$$

where $L_{\text{JEM}}(f(\mathbf{Z}'); M_\theta, \mathbf{y}^+)$ is a hybrid loss function used in joint-energy modelling evaluated at a given counterfactual state for a given model and target outcome:

$$L_{\text{JEM}}(f(\mathbf{Z}'); \cdot) = L_{\text{clf}}(f(\mathbf{Z}'); \cdot) + L_{\text{gen}}(f(\mathbf{Z}'); \cdot) \quad (8)$$

The first term, L_{clf} , is any standard classification loss function such as cross-entropy loss. The second term, L_{gen} , is used to measure loss with respect to the generative task¹. In the context of joint-energy training, L_{gen} induces changes in model parameters θ that decrease the energy of observed samples and increase the energy of samples generated through SGLD (Du and Mordatch 2019).

The key observation in our context is that we can rely solely on decreasing the energy of the counterfactual itself. This is sufficient to capture the generative property of the underlying model since it is implicitly captured by its parameters θ . Importantly, this means that we do not need to generate conditional samples through SGLD during our counterfactual search at all as we explain in the technical appendix.

This observation leads to the following simple objective function for *ECCCo*:

$$\begin{aligned} \mathbf{Z}' = \arg \min_{\mathbf{Z}' \in \mathcal{Z}^L} & \{L_{\text{clf}}(f(\mathbf{Z}'); M_\theta, \mathbf{y}^+) + \lambda_1 \text{cost}(f(\mathbf{Z}')) \\ & + \lambda_2 \mathcal{E}_\theta(f(\mathbf{Z}') | \mathbf{y}^+) + \lambda_3 \Omega(C_\theta(f(\mathbf{Z}'); \alpha))\} \end{aligned} \quad (9)$$

The first penalty term involving λ_1 induces closeness like in Wachter, Mittelstadt, and Russell (2017). The second penalty term involving λ_2 induces faithfulness by constraining the energy of the generated counterfactual. The third and final penalty term involving λ_3 ensures that the generated counterfactual is associated with low predictive uncertainty. To tune these hyperparameters we have relied on grid search.

Concerning feature autoencoding ($f : \mathcal{Z} \mapsto \mathcal{X}$), *ECCCo* does not rely on latent space search to achieve its primary objective of faithfulness. By default, we choose $f(\cdot)$ to be the identity function as in *Wachter*. This is generally also enough to achieve plausibility, provided the model has learned plausible explanations for the data. In some cases, plausibility can be improved further by mapping counterfactuals to a lower-dimensional latent space. In the following, we refer to this approach as *ECCCo+*: that is, *ECCCo* plus dimensionality reduction.

Figure 2 illustrates how the different components in Equation 9 affect the counterfactual search for a synthetic dataset. The underlying classifier is a Joint Energy Model (JEM) that was trained to predict the output class (blue or orange) and generate class-conditional samples (Grathwohl et al. 2020). We have used four different generator flavours to produce

¹In practice, regularization loss is typically also added. We follow this convention but have omitted the term here for simplicity.

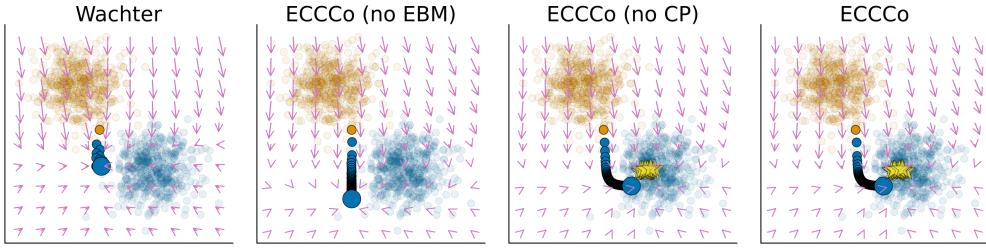


Figure 2: Gradient fields and counterfactual paths for different generators. The objective is to generate a counterfactual in the blue class for a sample from the orange class. Bright yellow stars indicate conditional samples generated through SGLD. The underlying classifier is a Joint Energy Model.

a counterfactual in the blue class for a sample from the orange class: *Wachter*, which only uses the first penalty ($\lambda_2 = \lambda_3 = 0$); *ECCCo (no EBM)*, which does not constrain energy ($\lambda_2 = 0$); *ECCCo (no CP)*, which involves no set size penalty ($\lambda_3 = 0$); and, finally, *ECCCo*, which involves all penalties defined in Equation 9. Arrows indicate (negative) gradients with respect to the objective function at different points in the feature space.

While *Wachter* generates a valid counterfactual, it ends up close to the original starting point consistent with its objective. *ECCCo (no EBM)* pushes the counterfactual further into the target domain to minimize predictive uncertainty, but the outcome is still not plausible. The counterfactual produced by *ECCCo (no CP)* is energy-constrained. Since the *JEM* has learned the conditional input distribution reasonably well in this case, the counterfactuals are both faithful and plausible. Finally, the outcome for *ECCCo* looks similar, but the additional smooth set size penalty leads to somewhat faster convergence.

6 Empirical Analysis

Our goal in this section is to shed light on the following research questions:

Research Question 6.1 (Faithfulness). *To what extent are counterfactuals generated by *ECCCo* more faithful than those produced by state-of-the-art generators?*

Research Question 6.2 (Balancing Desiderata). *Compared to state-of-the-art generators, how does *ECCCo* balance the two key objectives of faithfulness and plausibility?*

The second question is motivated by the intuition that faithfulness and plausibility should coincide for models that have learned plausible explanations of the data.

6.1 Experimental Setup

To assess and benchmark the performance of our proposed generator against the state of the art, we generate multiple counterfactuals for different models and datasets. In particular, we compare *ECCCo* and its variants to the following counterfactual generators that were introduced above: firstly; *Schut*, which works under the premise of minimizing predictive uncertainty; secondly, *REVISE*, which is state-of-the-art (SOTA) with respect to plausibility; and, finally,

Wachter, which serves as our baseline. In the case of *ECCCo+*, we use principal component analysis (PCA) for dimensionality reduction: the latent space \mathcal{Z} is spanned by the first n_z principal components where we choose n_z to be equal to the latent dimension of the VAE used by *REVISE*.

For the predictive modelling tasks, we use multi-layer perceptrons (*MLP*), deep ensembles, joint energy models (*JEM*) and convolutional neural networks (LeNet-5 *CNN* (LeCun et al. 1998)). Both joint-energy modelling and ensembling have been associated with improved generative properties and adversarial robustness (Grathwohl et al. 2020; Lakshminarayanan, Pritzel, and Blundell 2016), so we expect this to be positively correlated with the plausibility of *ECCCo*. To account for stochasticity, we generate multiple counterfactuals for each target class, generator, model and dataset. Full details concerning our parameter choices, training procedures and model performance can be found in the appendix.

We perform benchmarks on eight datasets from different domains. From the credit and finance domain we include three tabular datasets: *Give Me Some Credit (GMSC)* (Kaggle 2011), *German Credit* Hoffman (1994) and *California Housing* Pace and Barry (1997). All of these are commonly used in the related literature (Karimi et al. 2020; Altmeyer et al. 2023; Pawelczyk et al. 2021). Following related literature (Schut et al. 2021; Dhurandhar et al. 2018) we also include two image datasets: *MNIST* (LeCun 1998) and *Fashion MNIST* (Xiao, Rasul, and Vollgraf 2017). Detailed descriptions and results for all datasets can be found in the appendix.

In the following, we will focus on the most relevant results highlighted in Tables 1 and 2. The tables show sample averages along with standard deviations for our key evaluation metrics for the *California Housing* and *GMSC* datasets (Table 1) and the *MNIST* dataset (Table 2). For each metric, the best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*). For the tabular datasets, we use the default Euclidian distance to measure unfaithfulness and implausibility as defined in Equations 6 and 5, respectively. The third metric presented 1 in Table quantifies the predictive uncertainty of the counterfactual as measured by Equation 4. For the vision datasets, we rely on measuring the structural dissimilarity between images for

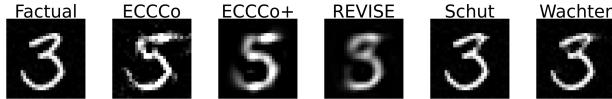


Figure 3: Counterfactuals for turning a 3 into a 5: factual (left), then the counterfactuals generated by *ECCCo*, *ECCCo+*, *REVISE*, *Schut* and *Wachter*.

our unfaithfulness and implausibility metrics (Wang, Simoncelli, and Bovik 2003).

6.2 Faithfulness

Overall, we find strong empirical evidence suggesting that *ECCCo* consistently achieves state-of-the-art faithfulness. Across all models and datasets highlighted here, all variations of *ECCCo* consistently outperform all other generators with respect to faithfulness, in many cases substantially. This pattern is mostly robust across all other benchmark datasets (Tables 7 to 14 in the technical appendix).

In particular, we note that the best results are generally obtained when using the full *ECCCo* objective (Equation 9). In other words, constraining both energy and predictive uncertainty typically yields the most faithful counterfactuals. We expected the former to play a more significant role in this context and that is typically what we find across all datasets. For example, the results for *GMSC* in Table 1 indicate that faithfulness can be improved substantially by relying solely on the energy constraint (*ECCCo (no CP)*). In some cases though, as for the *California Housing* dataset, *ECCCo (no EBM)* actually outperforms *ECCCo (no CP)*. This indicates that predictive uncertainty minimization plays an important role in achieving faithfulness.

We also generally find that the highest degree of faithfulness is obtained when the counterfactual search is performed directly in the feature space \mathcal{X} . While *ECCCo+* typically attains high levels of faithfulness compared to most other generators, it is consistently outperformed by *ECCCo*. The case is even stronger for *REVISE*, which performs worst out of all generators for faithfulness on the *GMSC* dataset and better only than *Wachter* on *California Housing*.

These findings are consistent with the notion that surrogate models may inhibit faithfulness. Even though dimensionality reduction through PCA in the case of *ECCCo+* can be considered a relatively mild form of intervention, the first n_z principal components fail to capture some of the variation in the data, that the underlying model itself may be sensitive to. This notion is illustrated nicely in Figure 3, where the counterfactual produced by *ECCCo* is somewhat noisier and grainier than the one produced by *ECCCo+*.

In conclusion, we recommend in light of the findings here to use the full *ECCCo* search objective whenever model faithfulness is a key priority.

6.3 Balancing Desiderata

Overall, we find strong empirical evidence suggesting that *ECCCo* can achieve near state-of-the-art plausibility without sacrificing faithfulness. Figure 3 shows one such exam-

ple taken from the *MNIST* benchmark where the objective is to turn the factual three (far left) into a five. The underlying model is a LeNet-5 *CNN*. The different images show the counterfactuals produced by the generators, of which all but the one produced by *Schut* are valid. Both variations of *ECCCo* produce plausible counterfactuals.

Looking at the benchmark results presented in Tables 1 and 2 we firstly note that although *REVISE* generally performs best, *ECCCo* and in particular *ECCCo+* often approach SOTA performance. Upon visual inspection of the generated images we actually find that *ECCCo+* performs much better than *REVISE* (see appendix). Zooming in on the details we observe that *ECCCo* and its variations do particularly well, whenever the underlying model has been explicitly trained to learn plausible representations of the data. For both tabular datasets in Table 1, *ECCCo* improves plausibility substantially compared to the baseline. This broad pattern is mostly consistent for all other datasets, although there are notable exceptions for which *ECCCo* takes the lead on both plausibility and faithfulness (see, for example, Tables 9 and 12 in the appendix).

While we maintain that generally speaking plausibility should hinge on the quality of the model, our results also indicate that it is possible to trade off some degree of faithfulness for plausibility if needed: *ECCCo+* generally outperforms other variants of *ECCCo* in this context at the small cost of slightly reduced faithfulness. For the vision datasets especially, we find that *ECCCo+* is consistently second only to *REVISE* for all models and regularly substantially better than the baseline. Looking at the *California Housing* data, latent space search markedly improves plausibility without sacrificing faithfulness: for the *JEM Ensemble*, *ECCCo+* performs substantially better than the baseline and only marginally worse than *REVISE*. Importantly, *ECCCo+* does not attain plausibility at all costs: for the *MLP*, plausibility is still very low but this seems to faithfully represent what the model has learned.

We conclude that *ECCCo* offers us a way to balance the objectives of faithfulness and plausibility. *ECCCo+* can be used to tilt the scale in favour of plausibility if needed.

6.4 Additional Desiderata

While we have deliberately focused on our key metrics of interest so far, it is worth briefly considering other common desiderata for counterfactuals. With reference to the rightmost columns for each dataset in Table 1, we firstly note that *ECCCo* typically reduces predictive uncertainty as intended. Consistent with its design, *Schut* performs well on this metric even though it does not explicitly address uncertainty as measured by conformal prediction set sizes.

Another commonly discussed desideratum is closeness (Wachter, Mittelstadt, and Russell 2017): counterfactuals that are closer to their factuals are associated with smaller costs to individuals in the context of algorithmic recourse. As evident from the additional tables in the appendix, the closeness desideratum tends to be negatively correlated with plausibility and faithfulness. Consequently, both *REVISE* and *ECCCo* generally yield more costly counterfactuals than the baseline. Nonetheless, *ECCCo* does not seem to stretch

Table 1: Results for tabular datasets: sample averages +/- one standard deviation across valid counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	California Housing			GMSC		
		Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓
MLP Ensemble	ECCCo	3.69 ± 0.08**	1.94 ± 0.13	0.09 ± 0.01**	3.84 ± 0.07**	2.13 ± 0.08	0.23 ± 0.01**
	ECCCo+	3.88 ± 0.07**	1.20 ± 0.09	0.15 ± 0.02	3.79 ± 0.05**	1.81 ± 0.05	0.30 ± 0.01*
	ECCCo (no CP)	3.70 ± 0.08**	1.94 ± 0.13	0.10 ± 0.01**	3.85 ± 0.07**	2.13 ± 0.08	0.23 ± 0.01**
	ECCCo (no EBM)	4.03 ± 0.07	1.12 ± 0.12	0.14 ± 0.01**	4.08 ± 0.06	0.97 ± 0.08	0.31 ± 0.01*
	REVISE	3.96 ± 0.07*	0.58 ± 0.03**	0.17 ± 0.03	4.09 ± 0.07	0.63 ± 0.02**	0.33 ± 0.06
	Schut	4.00 ± 0.06	1.15 ± 0.12	0.10 ± 0.01**	4.04 ± 0.08	1.21 ± 0.08	0.30 ± 0.01*
	Wachter	4.04 ± 0.07	1.13 ± 0.12	0.16 ± 0.01	4.10 ± 0.07	0.95 ± 0.08	0.32 ± 0.01
JEM Ensemble	ECCCo	1.40 ± 0.08**	0.69 ± 0.05**	0.11 ± 0.00**	1.20 ± 0.06*	0.78 ± 0.07**	0.38 ± 0.01
	ECCCo+	1.28 ± 0.08**	0.60 ± 0.04**	0.11 ± 0.00**	1.01 ± 0.07**	0.70 ± 0.07**	0.37 ± 0.01
	ECCCo (no CP)	1.39 ± 0.08**	0.69 ± 0.05**	0.11 ± 0.00**	1.21 ± 0.07*	0.77 ± 0.07**	0.39 ± 0.01
	ECCCo (no EBM)	1.70 ± 0.09	0.99 ± 0.08	0.14 ± 0.00*	1.31 ± 0.07	0.97 ± 0.10	0.32 ± 0.01**
	REVISE	1.39 ± 0.15**	0.59 ± 0.04**	0.25 ± 0.07	1.01 ± 0.07**	0.63 ± 0.04**	0.33 ± 0.07
	Schut	1.59 ± 0.10*	1.10 ± 0.06	0.09 ± 0.00**	1.34 ± 0.07	1.21 ± 0.10	0.26 ± 0.01**
	Wachter	1.71 ± 0.09	0.99 ± 0.08	0.14 ± 0.00	1.31 ± 0.08	0.95 ± 0.10	0.33 ± 0.01

Table 2: Results for vision dataset. Formatting details are the same as in Table 1.

MNIST			
Model	Generator	Unfaithfulness ↓	Implausibility ↓
MLP	ECCCo	0.243 ± 0.000**	0.420 ± 0.001
	ECCCo+	0.246 ± 0.000*	0.306 ± 0.001**
	REVISE	0.248 ± 0.000	0.301 ± 0.004**
	Schut	0.247 ± 0.001	0.303 ± 0.008**
	Wachter	0.247 ± 0.000	0.344 ± 0.002
LeNet-5	ECCCo	0.248 ± 0.000***	0.387 ± 0.002
	ECCCo+	0.248 ± 0.000***	0.310 ± 0.002**
	REVISE	0.248 ± 0.000***	0.301 ± 0.002**
	Schut	0.250 ± 0.002	0.289 ± 0.024*
	Wachter	0.249 ± 0.000	0.335 ± 0.002

costs unnecessarily: in Figure 3 useful parts of the factual three are clearly retained.

7 Limitations

Despite having taken considerable measures to study our methodology carefully, limitations can still be identified.

Firstly, we recognise that our proposed distance-based evaluation metrics for plausibility and faithfulness may not be universally applicable to all types of data. In any case, they depend on choosing a distance metric on a case-by-case basis, as we have done in this work. Arguably, commonly used metrics for measuring other desiderata such as closeness suffer from the same pitfall. We therefore think that future work on counterfactual explanations could benefit from defining universal evaluation metrics.

Relatedly, we note that our proposed metric for measuring faithfulness depends on the availability of samples generated through SGLD, which in turn requires gradient access

for models. This means it cannot be used to evaluate non-differentiable classifiers. Consequently, we also have not applied *ECCCo* to some machine learning models commonly used for classification such as decision trees. Since *ECCCo* itself does not rely on SGLD, its defining penalty functions are indeed applicable to gradient-free counterfactual generators. This is an interesting avenue for future research.

Next, common challenges associated with energy-based modelling including sensitivity to scale, training instabilities and sensitivity to hyperparameters also apply to *ECCCo* to some extent. In grid searches for optimal hyperparameters, we have noticed that unless properly regularized, *ECCCo* is sometimes prone to overshoot for the energy constraint.

Finally, while we have used ablation to understand the roles of the different components of *ECCCo*, the scope of this work has prevented us from investigating the role of conformal prediction in this context more thoroughly. We have exclusively relied on split conformal prediction and have used fixed values for the predetermined error rate and other hyperparameters. Future work could benefit from more extensive ablation studies that tune hyperparameters and investigate different approaches to conformal prediction.

8 Conclusion

This work leverages ideas from energy-based modelling and conformal prediction in the context of counterfactual explanations. We have proposed a new way to generate counterfactuals that are maximally faithful to the black-box model they aim to explain. Our proposed generator, *ECCCo*, produces plausible counterfactuals iff the black-box model itself has learned realistic explanations for the data, which we have demonstrated through rigorous empirical analysis. This should enable researchers and practitioners to use counterfactuals in order to discern trustworthy models from unreliable ones. While the scope of this work limits its generalizability, we believe that *ECCCo* offers a solid base for future work on faithful counterfactual explanations.

References

- Altmeyer, P.; Angela, G.; Buszydlik, A.; Dobiczek, K.; van Deursen, A.; and Liem, C. C. 2023. Endogenous Macrodynamics in Algorithmic Recourse. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 418–431. IEEE.
- Angelopoulos, A. N.; and Bates, S. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.
- Artelt, A.; Vaquet, V.; Velioglu, R.; Hinder, F.; Brinkrolf, J.; Schilling, M.; and Hammer, B. 2021. Evaluating robustness of counterfactual explanations. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, 01–09. IEEE.
- Dhurandhar, A.; Chen, P.-Y.; Luss, R.; Tu, C.-C.; Ting, P.; Shanmugam, K.; and Das, P. 2018. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *Advances in neural information processing systems*, 31.
- Dombrowski, A.-K.; Gerken, J. E.; and Kessel, P. 2021. Diffeomorphic Explanations with Normalizing Flows. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*.
- Du, Y.; and Mordatch, I. 2019. Implicit generation and generalization in energy-based models. *arXiv preprint arXiv:1903.08689*.
- Grathwohl, W.; Wang, K.-C.; Jacobsen, J.-H.; Duvenaud, D.; Norouzi, M.; and Swersky, K. 2020. Your classifier is secretly an energy based model and you should treat it like one.
- Guidotti, R. 2022. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 1–55.
- Hoffman, H. 1994. German Credit Data.
- Joshi, S.; Koyejo, O.; Vijitbenjaronk, W.; Kim, B.; and Ghosh, J. 2019. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*.
- Kaggle. 2011. Give Me Some Credit, Improve on the State of the Art in Credit Scoring by Predicting the Probability That Somebody Will Experience Financial Distress in the next Two Years.
- Karimi, A.-H.; Barthe, G.; Schölkopf, B.; and Valera, I. 2020. A Survey of Algorithmic Recourse: Definitions, Formulations, Solutions, and Prospects.
- Karimi, A.-H.; Schölkopf, B.; and Valera, I. 2021. Algorithmic Recourse: From Counterfactual Explanations to Interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 353–362.
- Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2016. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles.
- LeCun, Y. 1998. The MNIST database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>.
- LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.
- Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777.
- Mahajan, D.; Tan, C.; and Sharma, A. 2019. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*.
- Manokhin, V. 2022. Awesome Conformal Prediction. "If you use Awesome Conformal Prediction, please cite it as below".
- Molnar, C. 2022. *Interpretable Machine Learning*. 2 edition.
- Mothilal, R. K.; Sharma, A.; and Tan, C. 2020. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 607–617.
- Murphy, K. P. 2023. *Probabilistic machine learning: Advanced topics*. MIT press.
- Pace, R. K.; and Barry, R. 1997. Sparse Spatial Autoregressions. *Statistics & Probability Letters*, 33(3): 291–297.
- Pawelczyk, M.; Bielawski, S.; van den Heuvel, J.; Richter, T.; and Kasneci, G. 2021. Carla: A Python Library to Benchmark Algorithmic Recourse and Counterfactual Explanation Algorithms.
- Pawelczyk, M.; Datta, T.; van-den Heuvel, J.; Kasneci, G.; and Lakkaraju, H. 2022. Probabilistically Robust Recourse: Navigating the Trade-offs between Costs and Robustness in Algorithmic Recourse. *arXiv preprint arXiv:2203.06768*.
- Poyiadzi, R.; Sokol, K.; Santos-Rodriguez, R.; De Bie, T.; and Flach, P. 2020. FACE: Feasible and Actionable Counterfactual Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 344–350.
- Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.
- Schut, L.; Key, O.; Mc Grath, R.; Costabello, L.; Sacaleanu, B.; Gal, Y.; et al. 2021. Generating Interpretable Counterfactual Explanations By Implicit Minimisation of Epistemic and Aleatoric Uncertainties. In *International Conference on Artificial Intelligence and Statistics*, 1756–1764. PMLR.
- Slack, D.; Hilgard, A.; Lakkaraju, H.; and Singh, S. 2021. Counterfactual Explanations Can Be Manipulated. *Advances in Neural Information Processing Systems*, 34.
- Spooner, T.; Dervovic, D.; Long, J.; Shepard, J.; Chen, J.; and Magazzeni, D. 2021. Counterfactual explanations for arbitrary regression models. *arXiv preprint arXiv:2106.15212*.
- Stutz, D.; Dvijotham, K. D.; Cemgil, A. T.; and Doucet, A. 2022. Learning Optimal Conformal Classifiers.
- Upadhyay, S.; Joshi, S.; and Lakkaraju, H. 2021. Towards Robust and Reliable Algorithmic Recourse.

- Ustun, B.; Spangher, A.; and Liu, Y. 2019. Actionable Re-course in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 10–19.
- Verma, S.; Dickerson, J.; and Hines, K. 2020. Counterfactual Explanations for Machine Learning: A Review.
- Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harv. JL & Tech.*, 31: 841.
- Wang, Z.; Simoncelli, E. P.; and Bovik, A. C. 2003. Multi-scale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, 1398–1402. Ieee.
- Welling, M.; and Teh, Y. W. 2011. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, 681–688. Citeseer.
- Wilson, A. G. 2020. The case for Bayesian deep learning. *arXiv preprint arXiv:2001.10995*.
- Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Appendices

The following appendices provide additional details that are relevant to the paper. Appendices A and B explain any tasks related to Energy-Based Modelling and Predictive Uncertainty Quantification through Conformal Prediction, respectively. Appendix C provides additional technical and implementation details about our proposed generator, *ECCCo*, including references to our open-sourced code base. A complete overview of our experimental setup detailing our parameter choices, training procedures and initial black-box model performance can be found in Appendix D. Finally, Appendix F reports all of our experimental results in more detail.

A Energy-Based Modelling

Since we were not able to identify any existing open-source software for Energy-Based Modelling that would be flexible enough to cater to our needs, we have developed a *Julia* package from scratch. The package has been open-sourced, but to avoid compromising the double-blind review process, we refrain from providing more information at this stage. In our development we have heavily drawn on the existing literature: Du and Mordatch (2019) describe best practices for using EBM for generative modelling; Grathwohl et al. (2020) explain how EBM can be used to train classifiers jointly for the discriminative and generative tasks. We have used the same package for training and inference, but there are some important differences between the two cases that are worth highlighting here.

Training: Joint Energy Models To train our Joint Energy Models we broadly follow the approach outlined in Grathwohl et al. (2020). Formally, JEMs are defined by the following joint distribution:

$$\log p_\theta(\mathbf{x}, \mathbf{y}) = \log p_\theta(\mathbf{y}|\mathbf{x}) + \log p_\theta(\mathbf{x}) \quad (10)$$

Training therefore involves a standard classification loss component $L_{\text{clf}}(\theta) = -\log p_\theta(\mathbf{y}|\mathbf{x})$ (e.g. cross-entropy loss) as well as a generative loss component $L_{\text{gen}}(\theta) = -\log p_\theta(\mathbf{x})$. Analogous to how we defined the conditional distribution over inputs in Definition 4.1, $p_\theta(\mathbf{x})$ denotes the unconditional distribution over inputs. The model gradient of this component of the loss function can be expressed as follows:

$$\nabla_\theta L_{\text{gen}}(\theta) = -\nabla_\theta \log p_\theta(\mathbf{x}) = -(\mathbb{E}_{p(\mathbf{x})} \{\nabla_\theta \mathcal{E}_\theta(\mathbf{x})\} - \mathbb{E}_{p_\theta(\mathbf{x})} \{\nabla_\theta \mathcal{E}_\theta(\mathbf{x})\}) \quad (11)$$

To draw samples from $p_\theta(\mathbf{x})$, we rely exclusively on the conditional sampling approach described in Grathwohl et al. (2020) for both training and inference: we first draw $\mathbf{y} \sim p(\mathbf{y})$ and then sample $\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{y})$ (Grathwohl et al. 2020) via Equation 2 with energy $\mathcal{E}_\theta(\mathbf{x}|\mathbf{y}) = \mu_\theta(\mathbf{x})[\mathbf{y}]$ where $\mu_\theta : \mathcal{X} \mapsto \mathbb{R}^K$ returns the linear predictions (logits) of our classifier M_θ . While our package also supports unconditional sampling, we found conditional sampling to work well. It is also well aligned with CE, since in this context we are interested in conditioning on the target class.

As mentioned in the body of the paper, we rely on a biased sampler involving separately specified values for the step size ϵ and the standard deviation σ of the stochastic term involving \mathbf{r} . Formally, our biased sampler performs updates as follows:

$$\hat{\mathbf{x}}_{j+1} \leftarrow \hat{\mathbf{x}}_j - \frac{\phi}{2} \mathcal{E}_\theta(\hat{\mathbf{x}}_j | \mathbf{y}^+) + \sigma \mathbf{r}_j, \quad j = 1, \dots, J \quad (12)$$

Consistent with Grathwohl et al. (2020), we have specified $\phi = 2$ and $\sigma = 0.01$ as the default values for all of our experiments. Here we have deliberately departed slightly from the notation in Equation 2 to emphasize that we use fixed values for ϕ and σ , consistent with the related literature. The number of total SGLD steps J varies by dataset (Table 3). Following best practices, we initialize \mathbf{x}_0 randomly in 5% of all cases and sample from a buffer in all other cases. The buffer itself is randomly initialised and gradually grows to a maximum of 10,000 samples during training as $\hat{\mathbf{x}}_J$ is stored in each epoch (Du and Mordatch 2019; Grathwohl et al. 2020).

It is important to realise that sampling is done during each training epoch, which makes training Joint Energy Models significantly harder than conventional neural classifiers. In each epoch the generated (batch of) sample(s) $\hat{\mathbf{x}}_J$ is used as part of the generative loss component, which compares its energy to that of observed samples \mathbf{x} :

$$L_{\text{gen}}(\theta) \approx \mu_\theta(\mathbf{x})[\mathbf{y}] - \mu_\theta(\hat{\mathbf{x}}_J)[\mathbf{y}] \quad (13)$$

Our full training objective can be summarized as follows,

$$L_{\text{JEM}}(\theta) = L_{\text{clf}}(\theta) + L_{\text{gen}}(\theta) + \lambda L_{\text{reg}}(\theta) \quad (14)$$

where $L_{\text{reg}}(\theta)$ is a Ridge penalty (L2 norm) that regularises energy magnitudes for both observed and generated samples (Du and Mordatch 2019). We have used varying degrees of regularization depending on the dataset (λ in Table 3).

Contrary to existing work, we have not typically used the entire minibatch of training data for the generative loss component but found that using a subset of the minibatch was often sufficient in attaining decent generative performance (Table 3). This has helped to reduce the computational burden for our models, which should make it easier for others to reproduce our findings. Figures 4 and 5 show generated samples for our *MNIST* and *Moons* data, to provide a sense of their generative property.

Table 3: EBM hyperparameter choices for our experiments.

Dataset	SGLD Steps	Batch Size	λ
Linearly Separable	50	50	0.10
Moons	30	10	0.10
Circles	30	50	0.01
California Housing	30	10	0.10
GMSC	30	10	0.10
German Credit	30	10	0.10
MNIST	25	10	0.01
Fashion MNIST	25	10	0.01

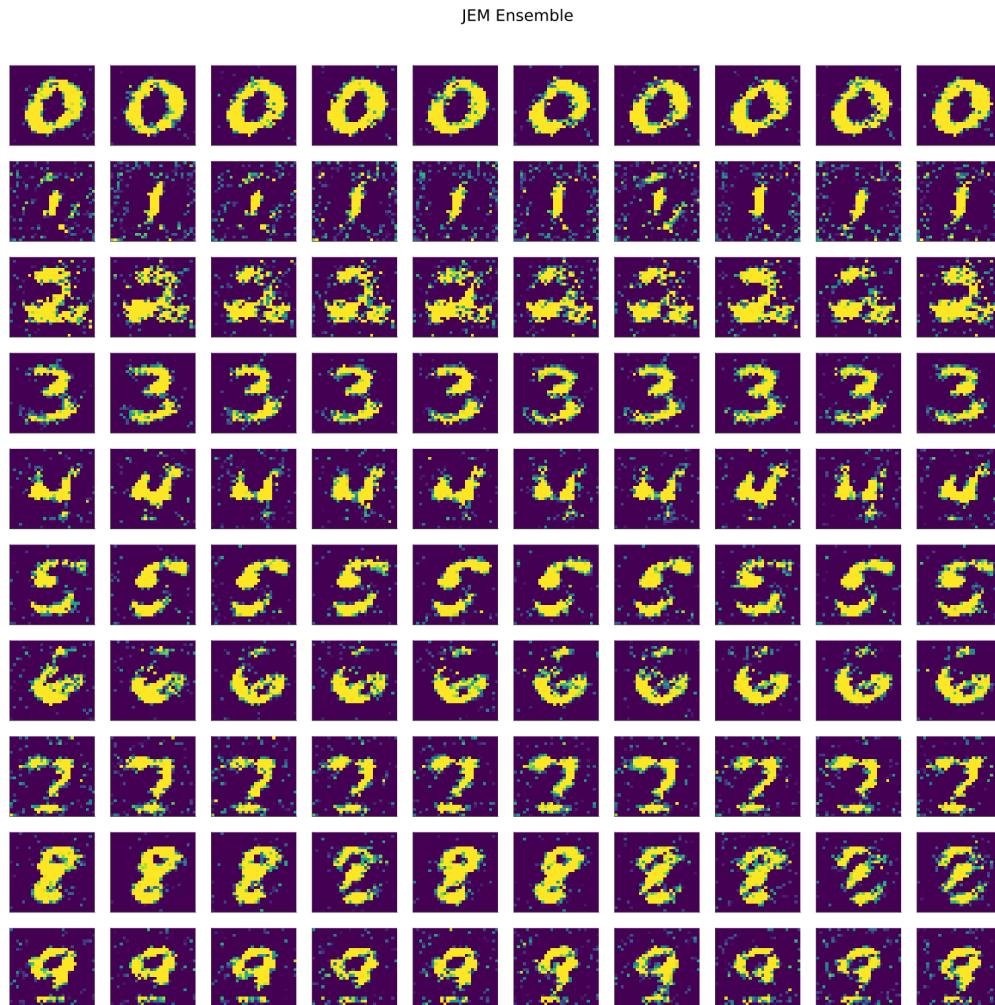


Figure 4: Conditionally generated *MNIST* images for our JEM Ensemble.

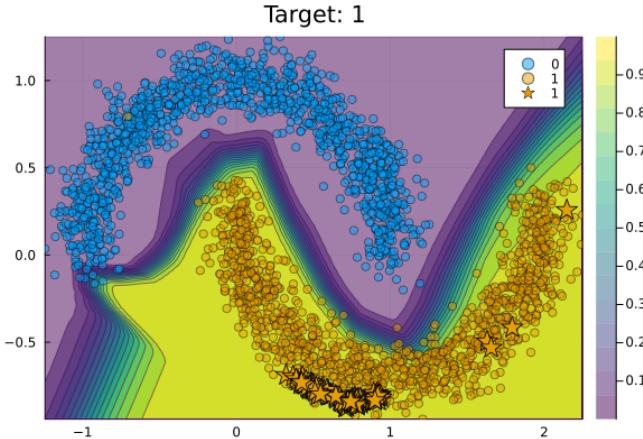


Figure 5: Conditionally generated samples (stars) for our *Moons* data using a JEM.

Inference: Quantifying Models’ Generative Property At inference time, we assume no prior knowledge about the model’s generative property. This means that we do not tap into the existing buffer of generated samples for our Joint Energy Models, but instead generate conditional samples from scratch. While we have relied on the default values $\epsilon = 2$ and $\sigma = 0.01$ also during inference, the number of total SGLD steps was set to $J = 500$ in all cases, so significantly higher than during training. For all of our synthetic datasets and models, we generated 50 conditional samples and then formed subsets containing the $n_E = 25$ lowest-energy samples. While in practice it would be sufficient to do this once for each model and dataset, we have chosen to perform sampling separately for each individual counterfactual in our experiments to account for stochasticity. To help reduce the computational burden for our real-world datasets we have generated only 10 conditional samples each time and used all of them in our counterfactual search. Using more samples, as we originally did, had no substantial impact on our results.

B Conformal Prediction

In this Appendix B we provide some more background on CP and explain in some more detail how we have used recent advances in Conformal Training for our purposes.

Background on CP Intuitively, CP works under the premise of turning heuristic notions of uncertainty into rigorous uncertainty estimates by repeatedly sifting through the data. It can be used to generate prediction intervals for regression models and prediction sets for classification models. Since the literature on CE and AR is typically concerned with classification problems, we focus on the latter. A particular variant of CP called Split Conformal Prediction (SCP) is well-suited for our purposes, because it imposes only minimal restrictions on model training.

Specifically, SCP involves splitting the data $\mathcal{D}_n = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1,\dots,n}$ into a proper training set $\mathcal{D}_{\text{train}}$ and a calibration set \mathcal{D}_{cal} . The former is used to train the classifier in any conventional fashion. The latter is then used to compute so-called nonconformity scores: $\mathcal{S} = \{s(\mathbf{x}_i, \mathbf{y}_i)\}_{i \in \mathcal{D}_{\text{cal}}}$ where $s : (\mathcal{X}, \mathcal{Y}) \mapsto \mathbb{R}$ is referred to as *score function*. In the context of classification, a common choice for the score function is just $s_i = 1 - M_\theta(\mathbf{x}_i)[\mathbf{y}_i]$, that is one minus the softmax output corresponding to the observed label \mathbf{y}_i (Angelopoulos and Bates 2021).

Finally, classification sets are formed as follows,

$$C_\theta(\mathbf{x}_i; \alpha) = \{\mathbf{y} : s(\mathbf{x}_i, \mathbf{y}) \leq \hat{q}\} \quad (15)$$

where \hat{q} denotes the $(1 - \alpha)$ -quantile of \mathcal{S} and α is a predetermined error rate. As the size of the calibration set increases, the probability that the classification set $C(\mathbf{x}_{\text{test}})$ for a newly arrived sample \mathbf{x}_{test} does not cover the true test label \mathbf{y}_{test} approaches α (Angelopoulos and Bates 2021).

Observe from Equation 15 that Conformal Prediction works on an instance-level basis, much like CE are local. The prediction set for an individual instance \mathbf{x}_i depends only on the characteristics of that sample and the specified error rate. Intuitively, the set is more likely to include multiple labels for samples that are difficult to classify, so the set size is indicative of predictive uncertainty. To see why this effect is exacerbated by small choices for α consider the case of $\alpha = 0$, which requires that the true label is covered by the prediction set with probability equal to 1.

Differentiability The fact that conformal classifiers produce set-valued predictions introduces a challenge: it is not immediately obvious how to use such classifiers in the context of gradient-based counterfactual search. Put differently, it is not clear how to use prediction sets in Equation 1. Fortunately, Stutz et al. (2022) have recently proposed a framework for Conformal

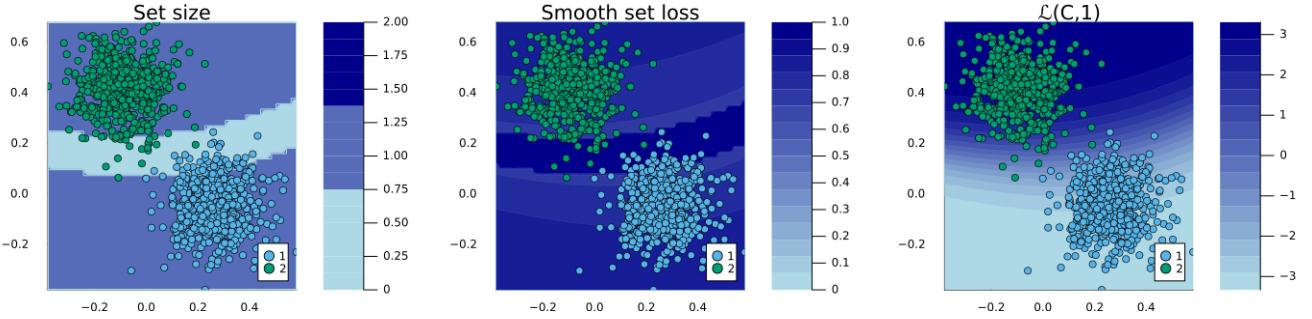


Figure 6: Prediction set size (left), smooth set size loss (centre) and configurable classification loss (right) for a JEM trained on our *Linearly Separable* data.

Training that also hinges on differentiability. Specifically, they show how Stochastic Gradient Descent can be used to train classifiers not only for the discriminative task but also for additional objectives related to Conformal Prediction. One such objective is *efficiency*: for a given target error rate α , the efficiency of a conformal classifier improves as its average prediction set size decreases. To this end, the authors introduce a smooth set size penalty defined in Equation 4 in the body of this paper. Formally, it is defined as $C_{\theta, \mathbf{y}}(\mathbf{x}_i; \alpha) := \sigma((s(\mathbf{x}_i, \mathbf{y}) - \alpha)T^{-1})$ for $\mathbf{y} \in \mathcal{Y}$, where σ is the sigmoid function and T is a hyper-parameter used for temperature scaling (Stutz et al. 2022).

In addition to the smooth set size penalty, Stutz et al. (2022) also propose a configurable classification loss function, that can be used to enforce coverage. For *MNIST* data, we found that using this function generally improved the visual quality of the generated counterfactuals, so we used it in our experiments involving real-world data. For the synthetic dataset, visual inspection of the counterfactuals showed that using the configurable loss function sometimes led to overshooting: counterfactuals would end up deep inside the target domain but far away from the observed samples. For this reason, we instead relied on standard cross-entropy loss for our synthetic datasets. As we have noted in the body of the paper, more experimental work is certainly needed in this context. Figure 6 shows the prediction set size (left), smooth set size loss (centre) and configurable classification loss (right) for a *JEM* trained on our *Linearly Separable* data.

C ECCo

In this section, we explain *ECCo* in some more detail, briefly discuss convergence conditions for counterfactual explanations and provide details concerning the actual implementation of our framework in *Julia*.

Deriving the search objective The counterfactual search objective for *ECCo* was introduced in Equation 9 in the body of the paper. We restate this equation here for reference:

$$\begin{aligned} \mathbf{Z}' = \arg \min_{\mathbf{Z}' \in \mathcal{Z}^L} & \{ \text{yloss}(M_\theta(f(\mathbf{Z}')), \mathbf{y}^+) + \lambda_1 \text{dist}(f(\mathbf{Z}'), \mathbf{x}) \\ & + \lambda_2 \mathcal{E}_\theta(\mathbf{Z}', \hat{\mathbf{X}}_{\theta, \mathbf{y}^+}) + \lambda_3 \Omega(C_\theta(f(\mathbf{Z}'); \alpha)) \} \end{aligned} \quad (16)$$

We can make the connection to energy-based modeling more explicit by restating the counterfactual search objective in terms $L_{\text{JEM}}(\theta)$, which we defined in Equation 14. In particular, consider the following counterfactual search objective,

$$\mathbf{Z}' = \arg \min_{\mathbf{Z}' \in \mathcal{Z}^L} \{ L_{\text{JEM}}(\theta; M_\theta(f(\mathbf{Z}')), \mathbf{y}^+) + \lambda_1 \text{dist}(f(\mathbf{Z}'), \mathbf{x}) + \lambda_3 \Omega(C_\theta(f(\mathbf{Z}'); \alpha)) \} \quad (17)$$

where we have simply used the JEM loss function as $\text{yloss}(M_\theta(f(\mathbf{Z}')), \mathbf{y}^+)$.

Now note that aside from the additional penalties in Equation 16, the only key difference between our counterfactual search objective and the joint-energy training objective is the parameter that is being optimized. In joint-energy training we optimize the objective with respect to the network weights θ . Recall that $\mathcal{E}_\theta(\mathbf{x}|\mathbf{y}) = \mu_\theta(\mathbf{x})[\mathbf{y}]$. Then the partial gradient with respect to the generative loss component of $L_{\text{JEM}}(\theta)$ can be expressed as follows:

$$\nabla_\theta L_{\text{gen}}(\theta) = \nabla_\theta \mu_\theta(\mathbf{x})[\mathbf{y}] - \nabla_\theta \mu_\theta(\hat{\mathbf{x}}_J)[\mathbf{y}] \quad (18)$$

During the counterfactual search, we take the network parameters as fixed and instead optimize with respect to the counterfactual itself²,

$$\nabla_\mathbf{x} L_{\text{gen}}(\theta) = \nabla_\mathbf{x} \mu_\theta(\mathbf{x})[\mathbf{y}^+] - \nabla_\mathbf{x} \mu_\theta(\hat{\mathbf{x}}_J)[\mathbf{y}^+] = \nabla_\mathbf{x} \mu_\theta(\mathbf{x})[\mathbf{y}^+] = \nabla_\mathbf{x} \mathcal{E}_\theta(\mathbf{x}|\mathbf{y}^+) \quad (19)$$

²Here we omit the notion of a latent search space to make the comparison easier.

where the second term is equal to zero because $\mu_\theta(\hat{\mathbf{x}}_J)[\mathbf{y}]$ is invariant with respect to \mathbf{x} . Since this term has zero gradients, we can remove it from the loss function altogether. For the regularization loss component of $L_{JEM}(\theta)$ we can proceed analogously such that we can rewrite Equation 17 as follows:

$$\begin{aligned}\mathbf{Z}' = \arg \min_{\mathbf{Z}' \in \mathcal{Z}^L} & \{ \text{yloss}(M_\theta(f(\mathbf{Z}')), \mathbf{y}^+) + \mathcal{E}_\theta(f(\mathbf{Z}')|\mathbf{y}^+) + \|\mathcal{E}_\theta(f(\mathbf{Z}')|\mathbf{y}^+)\|_2^2 \\ & + \lambda_1 \text{dist}(f(\mathbf{Z}'), \mathbf{x}) + \lambda_3 \Omega(C_\theta(f(\mathbf{Z}'); \alpha)) \}\end{aligned}\quad (20)$$

Now we notice that Equation 20 is equivalent to Equation 16 for $\lambda_2 = 1$. For the sake of simplicity, we omitted the regularization component from Equation 9 in the main text. Intuitively, taking iterative gradient steps according to Equation 19 has the effect of constraining the energy of the counterfactual until. The generative property of the underlying model implicitly enters this equation through θ .

The ECCCo algorithm Algorithm 1 describes how exactly *ECCCo* works. For the sake of simplicity and without loss of generality, we limit our attention to generating a single counterfactual $\mathbf{x}' = f(\mathbf{z}')$. The counterfactual state \mathbf{z}' is initialized at the factual \mathbf{x} . Other forms of initialization are also suitable but not considered here. For example, one may choose at a small random perturbation to all features (Slack et al. 2021). Next, we calibrate the model M_θ through split conformal prediction. Finally, we search counterfactuals through gradient descent where $\mathcal{L}(\mathbf{z}', \mathbf{y}^+, \hat{\mathbf{X}}_{\theta, \mathbf{y}^+}; \Lambda, \alpha)$ denotes our loss function defined in Equation 9. The search terminates once the convergence criterium is met or the maximum number of iterations T has been exhausted. Note that the choice of convergence criterium has important implications on the final counterfactual which we explain below.

Algorithm 1 The *ECCCo* generator

Input: $\mathbf{x}, \mathbf{y}^+, M_\theta, \Lambda = [\lambda_1, \lambda_2, \lambda_3], \alpha, \mathcal{D}, T$ where $M_\theta(\mathbf{x}) \neq \mathbf{y}^+$

Output: \mathbf{x}'

- 1: Initialize $\mathbf{z}' \leftarrow \mathbf{x}$
 - 2: Run SCP for M_θ using \mathcal{D} ▷ Calibrate model through split conformal prediction.
 - 3: Initialize $t \leftarrow 0$
 - 4: **while** not converged or $t < T$ **do** ▷ For convergence conditions see below.
 - 5: $\mathbf{z}' \leftarrow \mathbf{z}' - \eta \nabla_{\mathbf{z}'} \mathcal{L}(\mathbf{z}', \mathbf{y}^+; \Lambda, \alpha)$ ▷ Take gradient step of size η .
 - 6: $t \leftarrow t + 1$
 - 7: **end while**
 - 8: $\mathbf{x}' \leftarrow \mathbf{z}'$
-

The *ECCCo+* algorithm Algorithm 2 describes how exactly *ECCCo+* works. The only difference to *ECCCo* is that we encode and decode features using PCA. In particular, we let $f^{-1}(\mathbf{x})$ denote the projection of \mathbf{x} to its first n_z principal components. Conversely, $f(\mathbf{z}')$ maps back from the projection to the feature space.

Algorithm 2 The *ECCCo+* generator

Input: $\mathbf{x}, \mathbf{y}^+, M_\theta, f, \Lambda = [\lambda_1, \lambda_2, \lambda_3], \alpha, \mathcal{D}, T$ where $M_\theta(\mathbf{x}) \neq \mathbf{y}^+$

Output: \mathbf{x}'

- 1: Initialize $\mathbf{z}' \leftarrow f^{-1}(\mathbf{x})$ ▷ Map to counterfactual state space.
 - 2: Run SCP for M_θ using \mathcal{D} ▷ Calibrate model through split conformal prediction.
 - 3: Initialize $t \leftarrow 0$
 - 4: **while** not converged or $t < T$ **do** ▷ For convergence conditions see below.
 - 5: $\mathbf{z}' \leftarrow \mathbf{z}' - \eta \nabla_{\mathbf{z}'} \mathcal{L}(\mathbf{z}', \mathbf{y}^+; \Lambda, \alpha)$ ▷ Take gradient step of size η .
 - 6: $t \leftarrow t + 1$
 - 7: **end while**
 - 8: $\mathbf{x}' \leftarrow f(\mathbf{z}')$ ▷ Map back to feature space.
-

The *ECCCo-LI* algorithm Algorithm 3 describes a variation of *ECCCo* that we initially considered but ultimately discarded. For the sake of completeness we have included this approach here in the appendix. It generally yields very faithful counterfactuals but it is computationally much more expensive and struggles with plausibility.

Instead of constraining energy directly, this approach works under the premise of penalizing the distance between the counterfactual and samples generated through SGLD. The counterfactual state \mathbf{z}' is initialized as in Algorithm 1. Next, we generate n_B conditional samples $\hat{\mathbf{x}}_{\theta, \mathbf{y}^+}$ using SGLD (Equation 2) and store the n_E instances with the lowest energy. We then calibrate the model M_θ through split conformal prediction. Finally, we search counterfactuals through gradient descent where

$\mathcal{L}(\mathbf{z}', \mathbf{y}^+, \widehat{\mathbf{X}}_{\theta, \mathbf{y}^+}; \Lambda, \alpha)$ denotes our loss function defined in Equation 9, but instead of constraining energy directly we use Equation 6 (unfaithfulness metric) as a penalty.

Algorithm 3 The *ECCCo-LI* generator

Input: $\mathbf{x}, \mathbf{y}^+, M_\theta, f, \Lambda = [\lambda_1, \lambda_2, \lambda_3], \alpha, \mathcal{D}, T, \eta, n_B, n_E$ where $M_\theta(\mathbf{x}) \neq \mathbf{y}^+$
Output: \mathbf{x}'

- 1: Initialize $\mathbf{z}' \leftarrow \mathbf{x}$
- 2: Generate $\{\hat{\mathbf{x}}_{\theta, \mathbf{y}^+}\}_{n_B} \leftarrow p_\theta(\mathbf{x}_{\mathbf{y}^+})$ ▷ Generate n_B samples using SGLD (Equation 2).
- 3: Store $\widehat{\mathbf{X}}_{\theta, \mathbf{y}^+} \leftarrow \{\hat{\mathbf{x}}_{\theta, \mathbf{y}^+}\}_{n_B}$ ▷ Choose n_E lowest-energy samples.
- 4: Run SCP for M_θ using \mathcal{D} ▷ Calibrate model through split conformal prediction.
- 5: Initialize $t \leftarrow 0$
- 6: **while** not converged or $t < T$ **do** ▷ For convergence conditions see below.
- 7: $\mathbf{z}' \leftarrow \mathbf{z}' - \eta \nabla_{\mathbf{z}'} \mathcal{L}(\mathbf{z}', \mathbf{y}^+, \widehat{\mathbf{X}}_{\theta, \mathbf{y}^+}; \Lambda, \alpha)$ ▷ Take gradient step of size η .
- 8: $t \leftarrow t + 1$
- 9: **end while**
- 10: $\mathbf{x}' \leftarrow \mathbf{z}'$

A Note on Convergence Convergence is not typically discussed much in the context of CE, even though it has important implications on outcomes. One intuitive way to specify convergence is in terms of threshold probabilities: once the predicted probability $p(\mathbf{y}^+|\mathbf{x}')$ exceeds some user-defined threshold γ such that the counterfactual is valid, we could consider the search to have converged. In the binary case, for example, convergence could be defined as $p(\mathbf{y}^+|\mathbf{x}') > 0.5$ in this sense. Note, however, how this can be expected to yield counterfactuals in the proximity of the decision boundary, a region characterized by high aleatoric uncertainty. In other words, counterfactuals generated in this way would generally not be plausible. To avoid this from happening, we specify convergence in terms of gradients approaching zero for all our experiments and all of our generators. This allows us to get a cleaner read on how the different counterfactual search objectives affect counterfactual outcomes.

ECCCo.jl The core part of our code base is integrated into a larger ecosystem of *Julia* packages that we are actively developing and maintaining. To avoid compromising the double-blind review process, we only provide a link to an anonymized repository at this stage: <https://anonymous.4open.science/r/ECCCo-1252/README.md>.

D Experimental Setup

In our experiments we always generate multiple counterfactuals for each model and generator. Each time the factual and target class is drawn randomly. For each generator and model we choose $n_f = 100$ factuals for all of our synthetic and vision data. For tabular data we choose $n_f = 25$ because larger values made grid search computationally prohibitive. For vision data, grid search was computationally prohibitive in any case, so hyperparameters were tuned manually. For all other datasets grid search was performed over different combinations of penalty strengths and optimizer steps sizes (details of which can be found in the code). To select the final hyperparameter setting we used the unfaithfulness metric as our criterion.

Table 4 provides an overview of all parameters related to our experiments. The *GMSC* data were randomly undersampled for balancing purposes and all features were standardized. *MNIST* data was also randomly undersampled for reasons outlined below. Pixel values were preprocessed to fall in the range of $[-1, 1]$ and a small Gaussian noise component ($\sigma = 0.03$) was added to training samples following common practice in the EBM literature. All of our models were trained through mini-batch training using the Adam optimiser (Kingma and Ba (2014)). Table 5 shows standard evaluation metrics measuring the predictive performance of our different models grouped by dataset. These measures were computed on test data.

Table 6 summarises our hyperparameter choices for the counterfactual generators where η denotes the learning rate used for Stochastic Gradient Descent (SGD) and $\lambda_1, \lambda_2, \lambda_3$ represent the chosen penalty strengths (Equations 1 and 9). Here λ_1 also refers to the chosen penalty for the distance from factual values that applies to both *Wachter* and *REVISE*, but not *Schut* which is penalty-free. *Schut* is also the only generator that uses JSMA instead of SGD for optimization.

E Compute

To enable others to easily replicate our experiments, we have chosen to work with small neural network architectures and randomly undersampled the *MNIST* dataset (maintaining class balance). All of our final benchmarks could then be run locally on a personal machine, but grid searches and the sheer number of datasets required us to move to high-performance computing clusters to conduct our experiments efficiently.

Table 4: Parameter choices for our experiments.

Dataset	Sample Size	Network Architecture				Training	
		Hidden Units	Hidden Layers	Activation	Ensemble Size	Epochs	Batch Size
Linearly Separable	1000	16	3	swish	5	100	100
	2500	32	3	relu	5	500	128
	1000	32	1	swish	5	100	100
California Housing	16500	32	3	relu	5	100	128
	GMSC	13370	32	relu	5	100	128
	German Credit	800	32	relu	5	100	80
MNIST	10000	32	1	relu	5	100	128
	Fashion MNIST	10000	32	2	relu	5	100

Table 5: Various standard performance metrics for our different models grouped by dataset.

Dataset	Model	Performance Metrics		
		Accuracy	Precision	F1-Score
Linearly Separable	JEM	0.98	0.98	0.98
	JEM Ensemble	0.99	0.99	0.99
	MLP	0.99	0.99	0.99
	MLP Ensemble	0.99	0.99	0.99
Moons	JEM	1.00	1.00	1.00
	JEM Ensemble	1.00	1.00	1.00
	MLP	1.00	1.00	1.00
	MLP Ensemble	1.00	1.00	1.00
Circles	JEM	1.00	1.00	1.00
	JEM Ensemble	1.00	1.00	1.00
	MLP	1.00	1.00	1.00
	MLP Ensemble	1.00	1.00	1.00
California Housing	JEM	0.87	0.87	0.87
	JEM Ensemble	0.87	0.87	0.87
	MLP	0.89	0.89	0.89
	MLP Ensemble	0.89	0.89	0.89
GMSC	JEM	0.75	0.76	0.74
	JEM Ensemble	0.74	0.75	0.74
	MLP	0.74	0.75	0.74
	MLP Ensemble	0.74	0.74	0.74
German Credit	JEM	0.54	0.60	0.47
	JEM Ensemble	0.55	0.68	0.46
	MLP	0.54	0.76	0.42
	MLP Ensemble	0.51	0.75	0.36
MNIST	JEM	0.84	0.85	0.84
	JEM Ensemble	0.90	0.90	0.90
	LeNet-5	0.98	0.98	0.98
	MLP	0.95	0.95	0.95
	MLP Ensemble	0.95	0.95	0.95
Fashion MNIST	JEM	0.62	0.70	0.62
	JEM Ensemble	0.78	0.78	0.78
	LeNet-5	0.83	0.84	0.82
	MLP	0.82	0.83	0.82
	MLP Ensemble	0.84	0.84	0.84

Table 6: Generator hyperparameters: the optimiser step size (η); penalty strengths where λ_1 applies to all generators but *Schut* and the other parameter are specific to *ECCCo*; finally, the strength for the Ridge penalty on energy for *ECCCo*.

Dataset	η	λ_1	λ_2	λ_3	Ridge penalty
Linearly Separable	0.01	0.10	0.1	0.05	0.0
	Moons	0.01	0.10	0.1	0.50
	Circles	0.05	0.10	0.1	0.05
California Housing	0.05	0.10	0.1	0.10	0.0
	GMSC	0.05	0.10	0.1	0.10
	German Credit	0.05	0.20	0.2	0.20
MNIST	0.10	0.01	0.1	0.30	0.0
	Fashion MNIST	0.10	0.01	0.1	0.30

Local runs The longest runtimes we experienced for model training and counterfactual benchmarking for a single benchmark were on the order of 12-24 hours (*MNIST* data). For the synthetic data, single benchmarks could be completed in less than an hour. We have summarised our system information below:

Software:

- System Version: macOS 13.3.1
- Kernel Version: Darwin 22.4.0

Hardware:

- Model Name: MacBook Pro
- Model Identifier: MacBookPro16,1
- Processor Name: 8-Core Intel Core i9
- Processor Speed: 2.3 GHz
- Number of Processors: 1
- Total Number of Cores: 8
- L2 Cache (per Core): 256 KB
- L3 Cache: 16 MB
- Hyper-Threading Technology: Enabled
- Memory: 32 GB

HPC We used two large clusters, which we do not cite or mention by name at this point to not interfere with the double-blind review process. All of our grid searches were multi-processed on 100 CPUs at 4-8GB or memory each.

F Results

Figures 7 to 11 show examples of counterfactuals for *MNIST* generated by *ECCCo+* for our different models. Original images are shown on the diagonal and the corresponding counterfactuals are plotted across rows. Figures 12 to 16 show the same examples but for *REVISE*. Both counterfactual generators have access to the same optimizer. While the results for *REVISE* look fairly poor here, we have observed better results for optimizers with higher step sizes. Note that the seemingly poor performance by *REVISE* upon visual inspection is not driven by a weak surrogate VAE: Figure 17 shows image reconstructions generated by the VAE.

Tables 7 to 14 reports all of the evaluation metrics we have computed. Tables 15 to 22 reports the same metrics for the subset of valid counterfactuals. The ‘Unfaithfulness’ and ‘Implausibility’ metrics have been discussed extensively in the body of the paper. The ‘Cost’ metric relates to the distance between the factual and the counterfactual and is measured using the L1 Norm. The ‘Redundancy’ metric measures sparsity in is defined as the percentage of features that remain unperturbed (higher is better). The ‘Uncertainty’ metric is just the average value of the smooth set size penalty (Equation 4). Finally, ‘Validity’ is the percentage of valid counterfactuals.

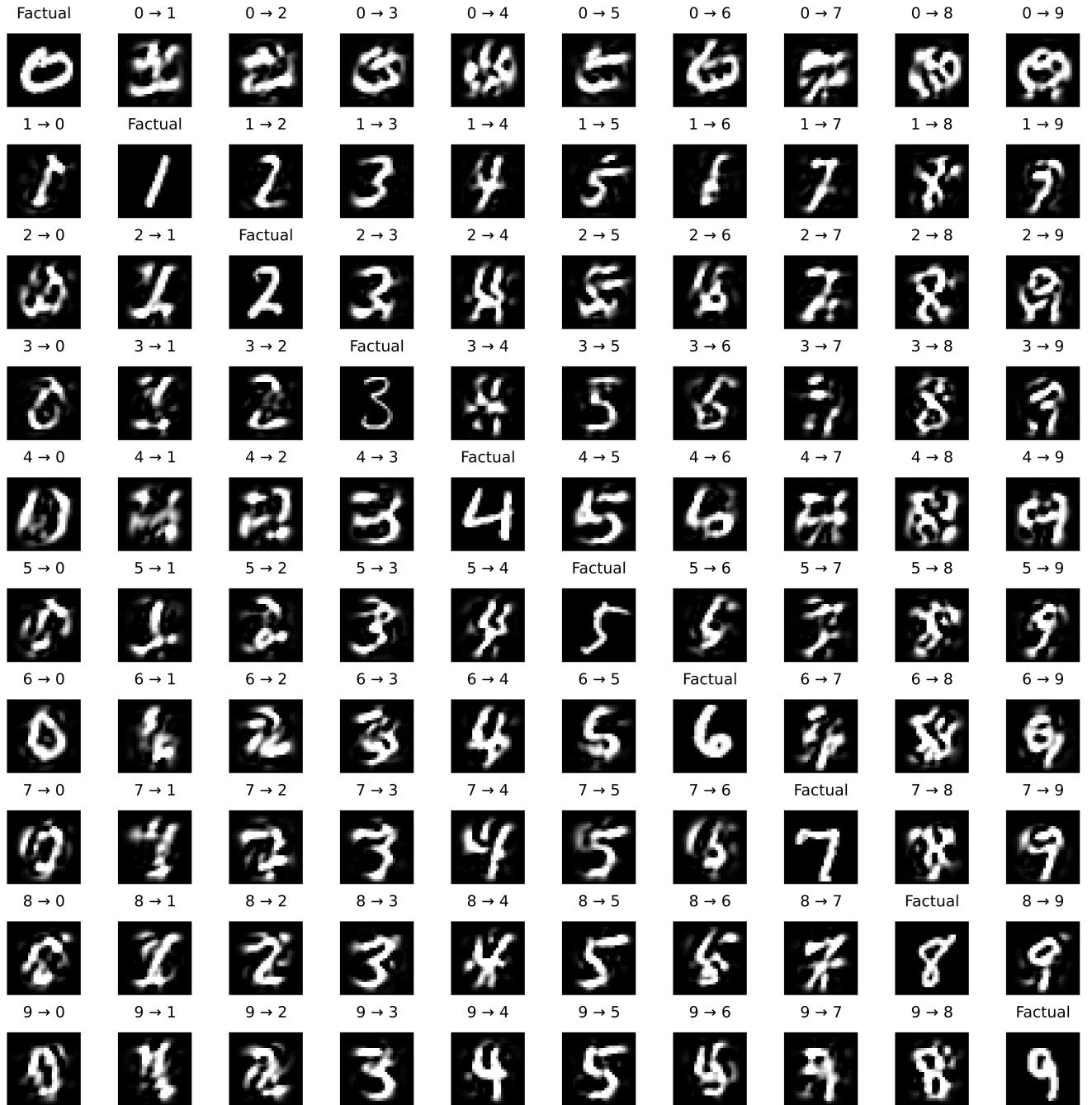


Figure 7: Counterfactuals for *MNIST* data generated by *ECCCo+*. The underlying model is a LeNet-5 CNN. Original images are shown on the diagonal with the corresponding counterfactuals plotted across rows.

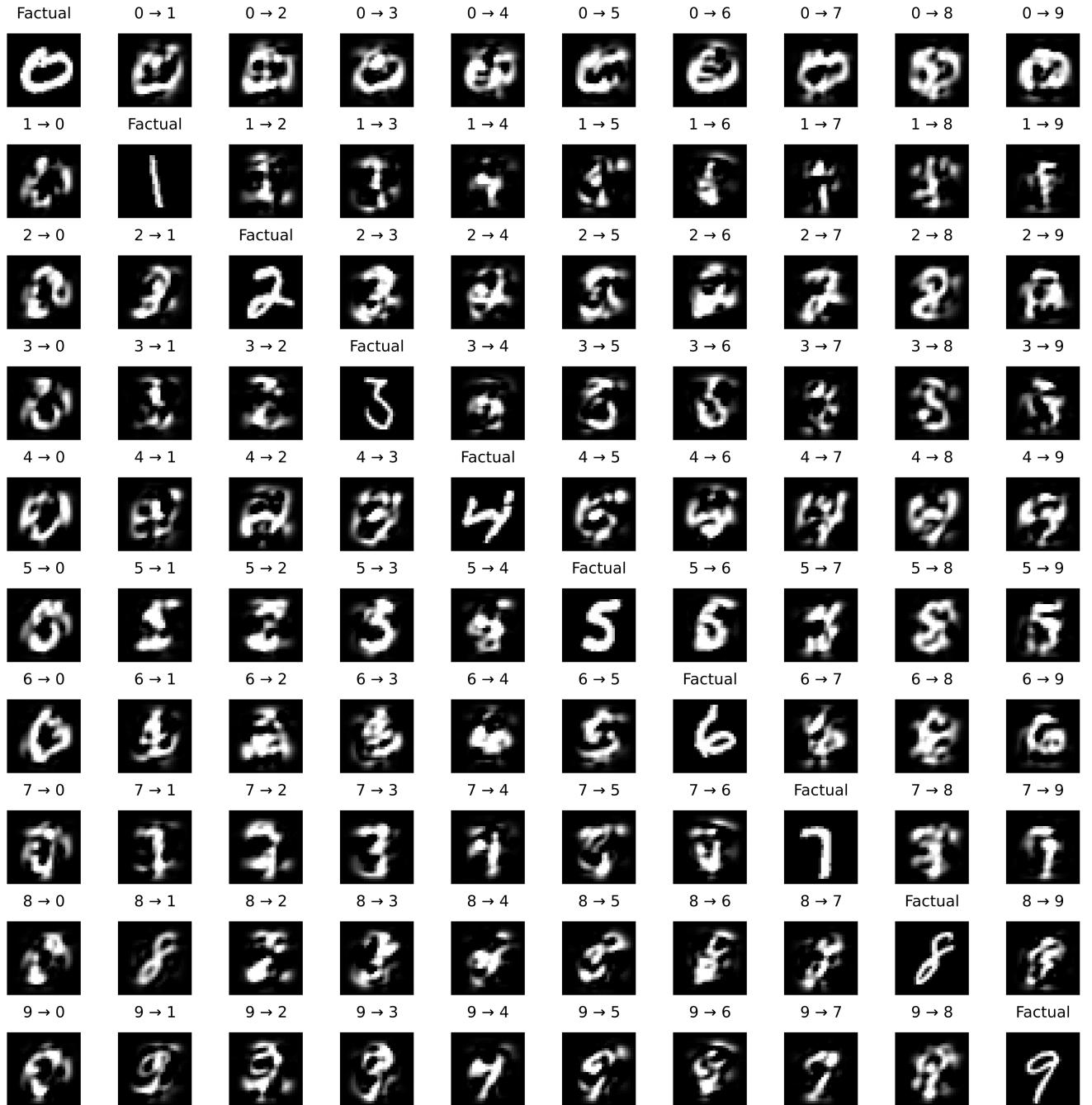


Figure 8: Counterfactuals for *MNIST* data generated by *ECCCo+*. The underlying model is an *MLP*. Original images are shown on the diagonal with the corresponding counterfactuals plotted across rows.

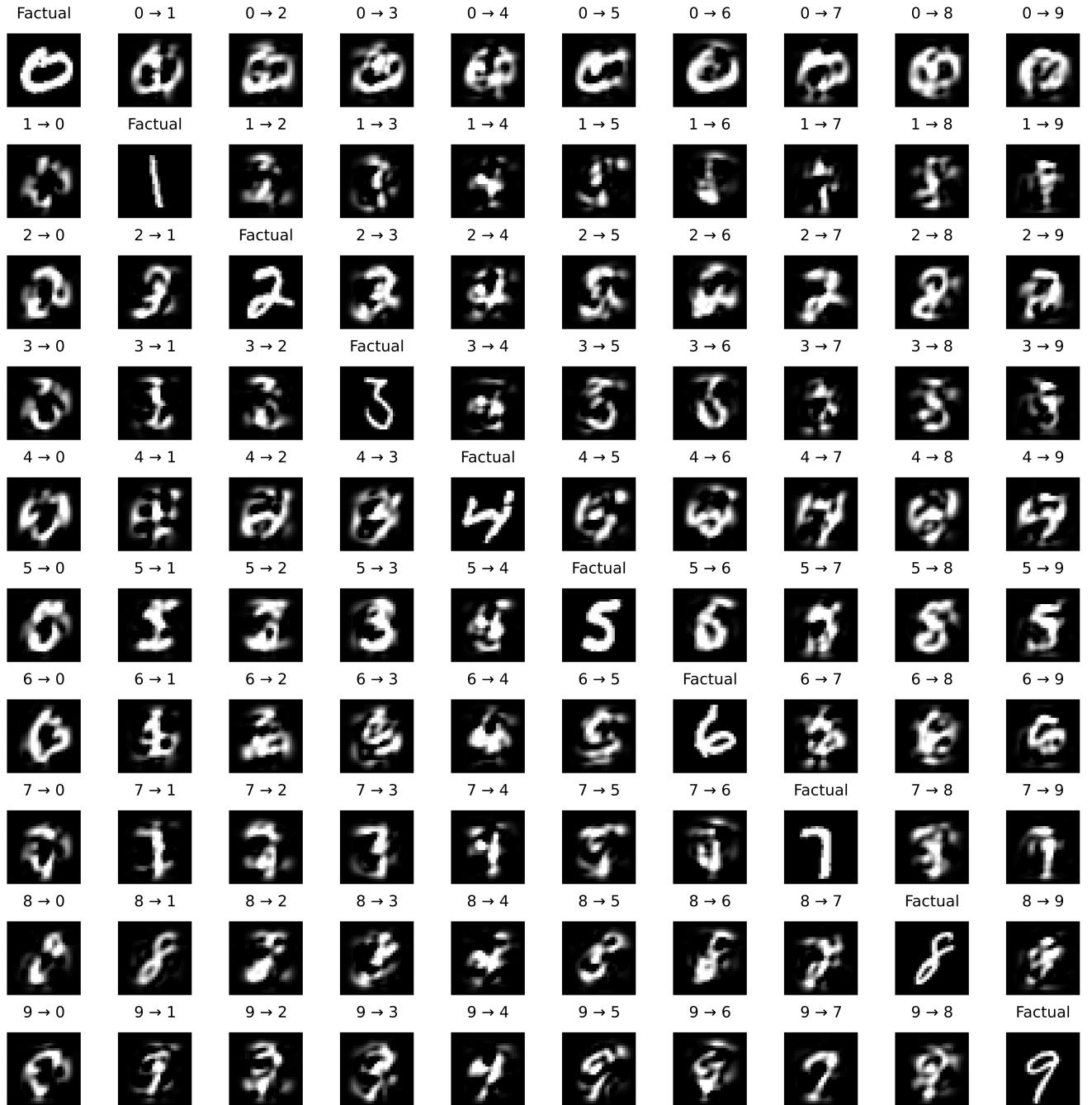


Figure 9: Counterfactuals for *MNIST* data generated by *ECCCo+*. The underlying model is an *MLP* ensemble. Original images are shown on the diagonal with the corresponding counterfactuals plotted across rows.

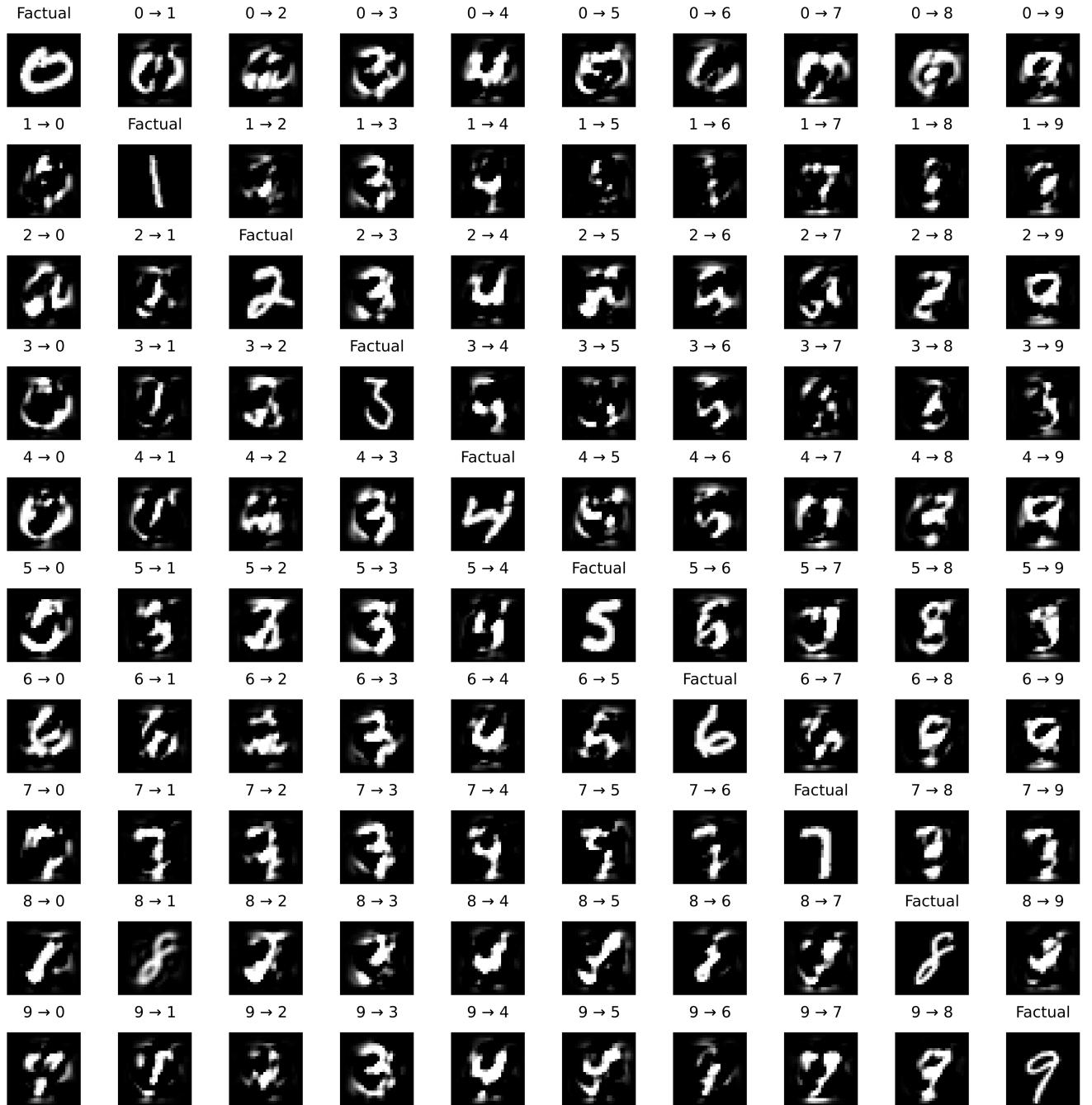


Figure 10: Counterfactuals for *MNIST* data generated by *ECCCo+*. The underlying model is a *JEM*. Original images are shown on the diagonal with the corresponding counterfactuals plotted across rows.

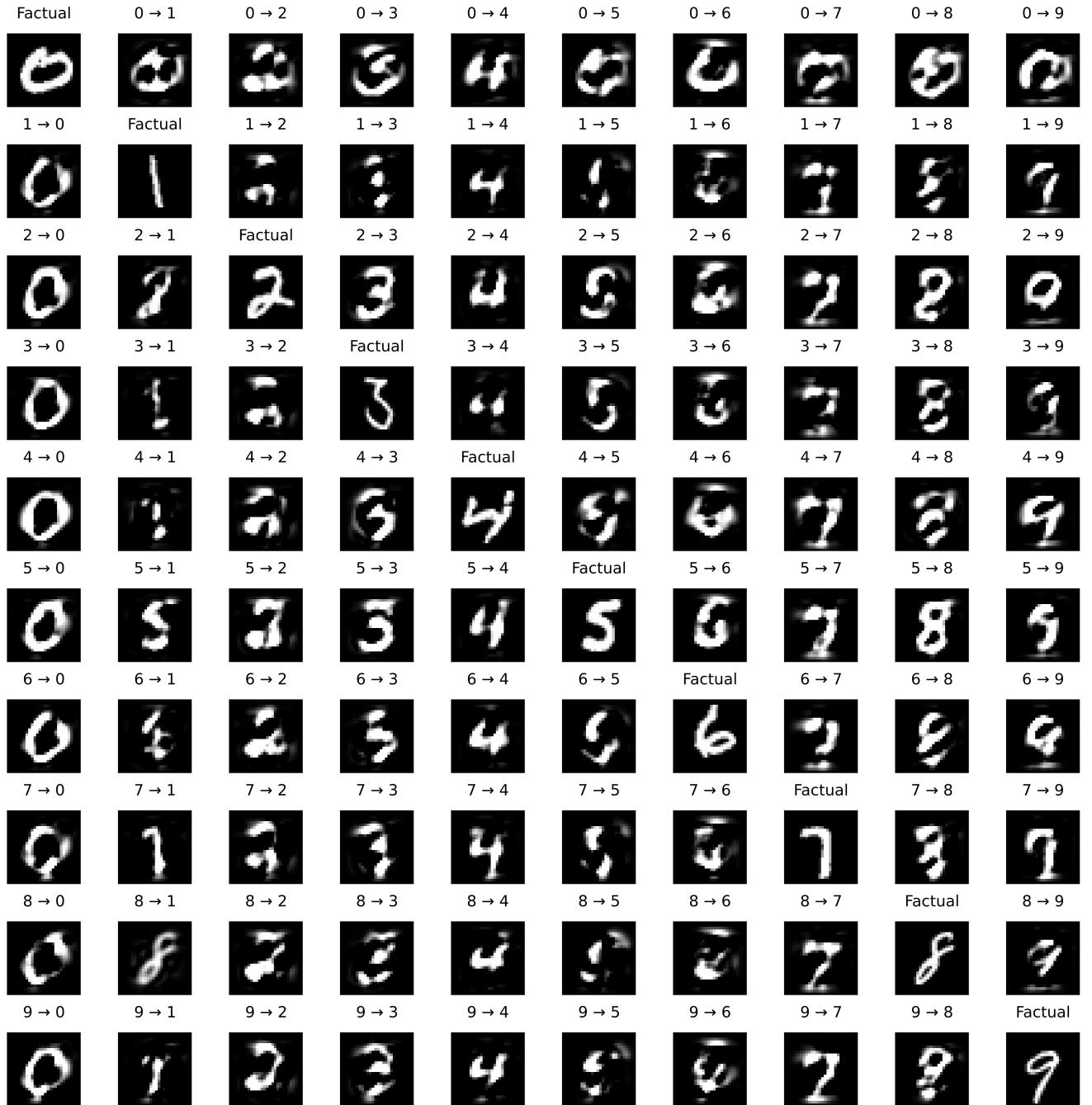


Figure 11: Counterfactuals for *MNIST* data generated by *ECCCo+*. The underlying model is a *JEM* ensemble. Original images are shown on the diagonal with the corresponding counterfactuals plotted across rows.

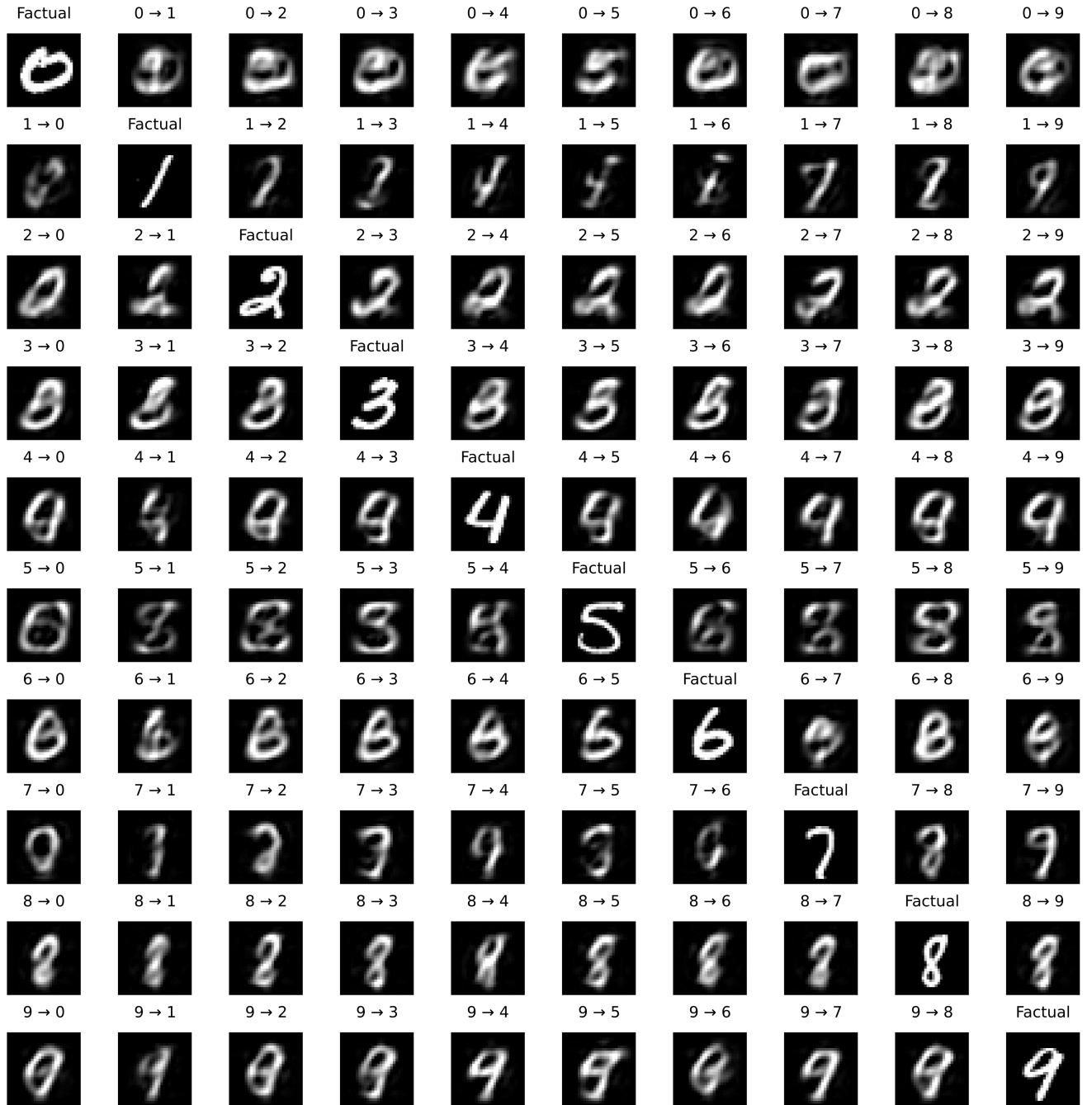


Figure 12: Counterfactuals for *MNIST* data generated by *REVISE*. The underlying model is a LeNet-5 CNN. Original images are shown on the diagonal with the corresponding counterfactuals plotted across rows.

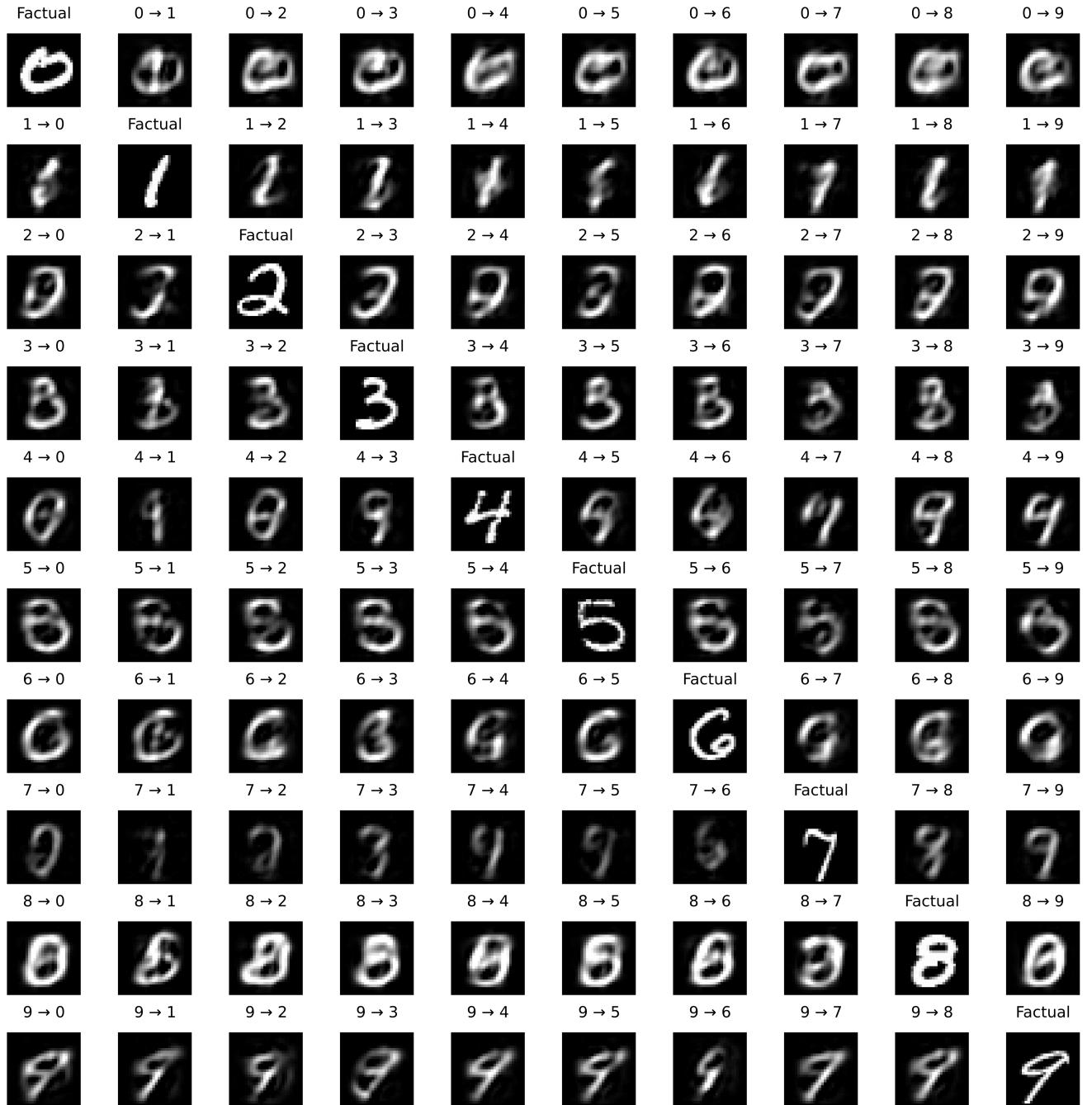


Figure 13: Counterfactuals for *MNIST* data generated by *REVISE*. The underlying model is an *MLP*. Original images are shown on the diagonal with the corresponding counterfactuals plotted across rows.

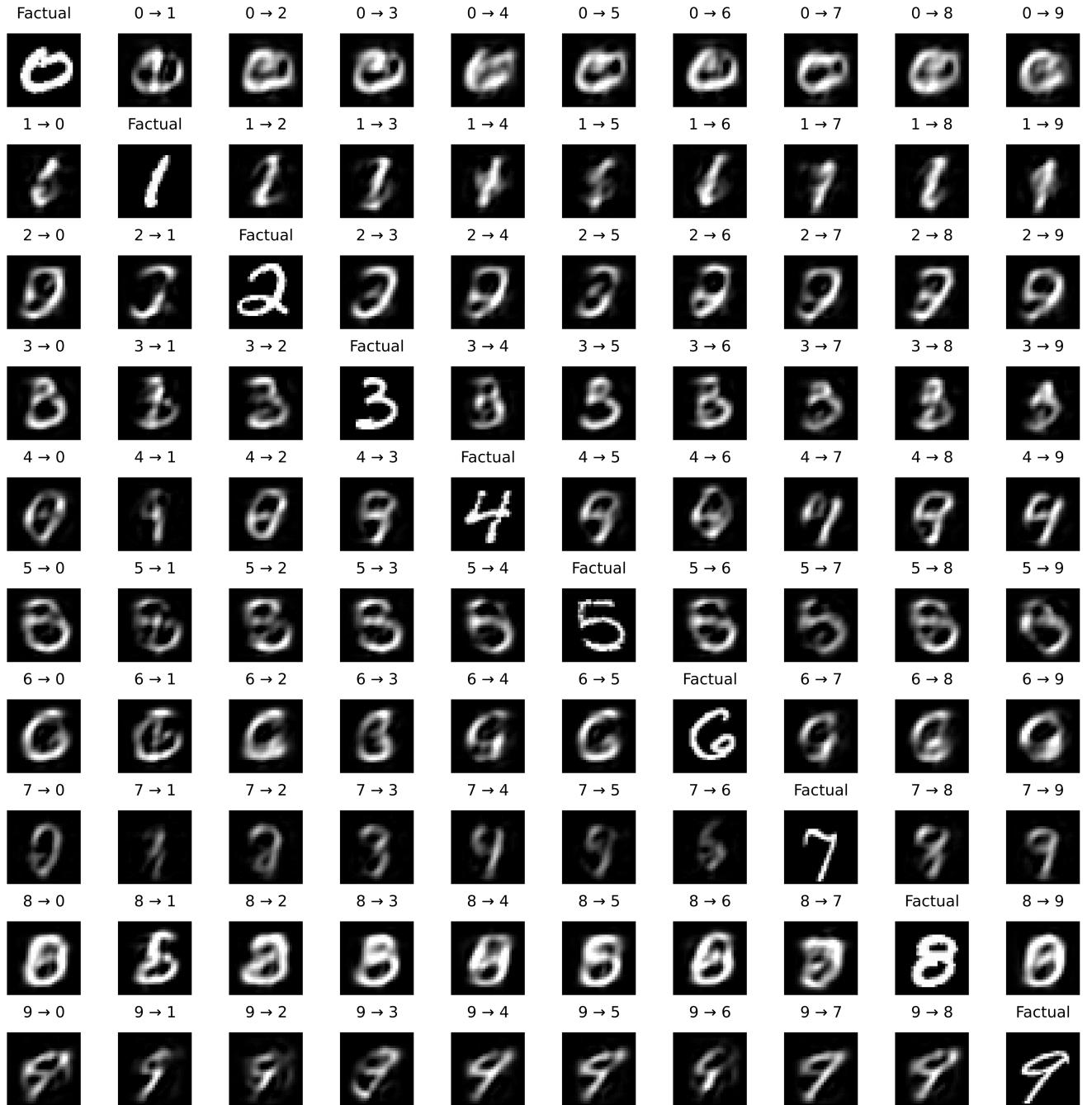


Figure 14: Counterfactuals for *MNIST* data generated by *REVISE*. The underlying model is an *MLP* ensemble. Original images are shown on the diagonal with the corresponding counterfactuals plotted across rows.

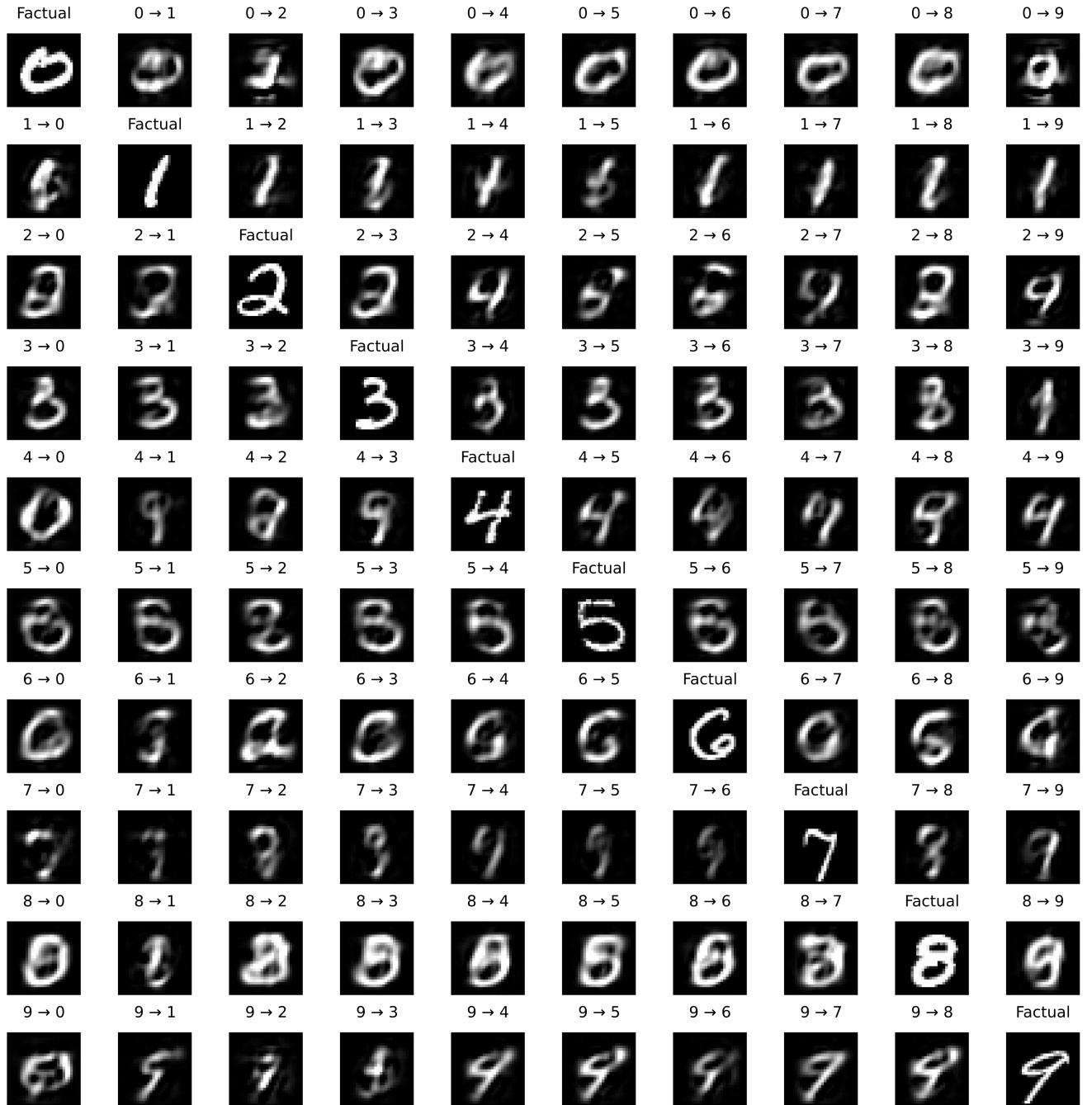


Figure 15: Counterfactuals for *MNIST* data generated by *REVISE*. The underlying model is a *JEM*. Original images are shown on the diagonal with the corresponding counterfactuals plotted across rows.

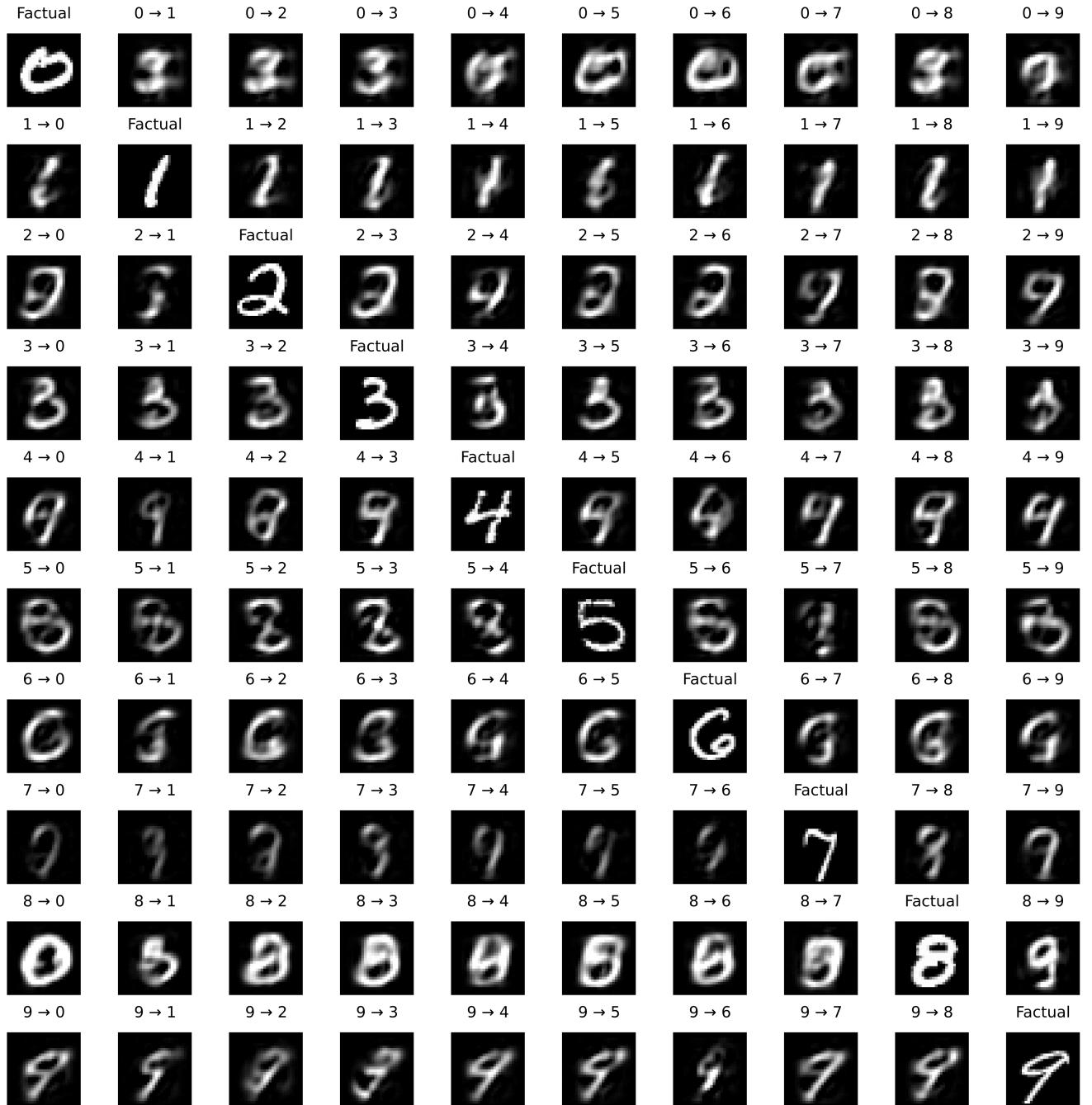


Figure 16: Counterfactuals for *MNIST* data generated by *REVISE*. The underlying model is a *JEM* ensemble. Original images are shown on the diagonal with the corresponding counterfactuals plotted across rows.

Table 7: All results for Linearly Separable dataset: sample averages +/- one standard deviation over all counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	0.06 ± 0.01**	0.16 ± 0.02**	0.00 ± 0.00	0.93 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.06 ± 0.01**	0.16 ± 0.02**	0.00 ± 0.00	0.93 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.09 ± 0.01**	0.24 ± 0.01	0.00 ± 0.00	0.88 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.19 ± 0.01**	0.19 ± 0.01**	0.00 ± 0.00	0.85 ± 0.01**	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	0.19 ± 0.01**	0.19 ± 0.01**	0.00 ± 0.00	0.85 ± 0.01**	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	0.24 ± 0.01	0.24 ± 0.01	0.00 ± 0.00	0.88 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.27 ± 0.01	0.18 ± 0.00**	0.00 ± 0.00	0.42 ± 0.01**	0.00 ± 0.00	0.47 ± 0.00
	Schut	0.29 ± 0.01	0.29 ± 0.00	0.00 ± 0.00	1.40 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	Wachter	0.24 ± 0.01	0.24 ± 0.01	0.00 ± 0.00	0.88 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
JEM Ensemble	ECCCo-L1	0.02 ± 0.00**	0.15 ± 0.01	0.00 ± 0.00	0.98 ± 0.02*	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.02 ± 0.00**	0.15 ± 0.01	0.00 ± 0.00	0.98 ± 0.02*	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.05 ± 0.00**	0.08 ± 0.00	0.00 ± 0.00	1.01 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.12 ± 0.01**	0.07 ± 0.00**	0.00 ± 0.00	0.97 ± 0.02**	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	0.12 ± 0.01**	0.07 ± 0.00**	0.00 ± 0.00	0.97 ± 0.02**	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	0.15 ± 0.01	0.08 ± 0.00	0.00 ± 0.00	1.01 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.00	0.18 ± 0.00	0.44 ± 0.05	0.42 ± 0.01**	0.00 ± 0.00	0.51 ± 0.06
	Schut	0.20 ± 0.01	0.27 ± 0.00	0.00 ± 0.00	1.38 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	Wachter	0.14 ± 0.01	0.08 ± 0.00	0.00 ± 0.00	1.01 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
MLP	ECCCo-L1	0.12 ± 0.00**	0.06 ± 0.00**	0.00 ± 0.00	0.91 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.12 ± 0.00**	0.06 ± 0.00**	0.00 ± 0.00	0.90 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.16 ± 0.00**	0.07 ± 0.00	0.00 ± 0.00	0.62 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.17 ± 0.00**	0.31 ± 0.00	0.00 ± 0.00	1.44 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	0.17 ± 0.00**	0.31 ± 0.00	0.00 ± 0.00	1.44 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	0.46 ± 0.00	0.07 ± 0.00	0.00 ± 0.00	0.62 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.53 ± 0.00	0.18 ± 0.00	0.52 ± 0.01	0.41 ± 0.02**	0.00 ± 0.00	0.53 ± 0.01
	Schut	0.48 ± 0.00	0.31 ± 0.01	0.00 ± 0.00	0.70 ± 0.02	0.46 ± 0.01**	1.00 ± 0.00
	Wachter	0.46 ± 0.00	0.07 ± 0.00	0.00 ± 0.00	0.62 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
MLP Ensemble	ECCCo-L1	0.12 ± 0.00**	0.06 ± 0.00**	0.00 ± 0.00	0.91 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.12 ± 0.00**	0.06 ± 0.00**	0.00 ± 0.00	0.91 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.16 ± 0.00**	0.07 ± 0.00	0.00 ± 0.00	0.65 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.17 ± 0.00**	0.31 ± 0.00	0.00 ± 0.00	1.44 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	0.17 ± 0.00**	0.31 ± 0.00	0.00 ± 0.00	1.44 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	0.45 ± 0.00	0.07 ± 0.00	0.00 ± 0.00	0.65 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.53 ± 0.00	0.18 ± 0.00	0.52 ± 0.01	0.42 ± 0.01**	0.00 ± 0.00	0.53 ± 0.01
	Schut	0.48 ± 0.00	0.32 ± 0.01	0.00 ± 0.00	0.73 ± 0.01	0.42 ± 0.02**	1.00 ± 0.00
	Wachter	0.45 ± 0.00	0.07 ± 0.00	0.00 ± 0.00	0.65 ± 0.01	0.00 ± 0.00	1.00 ± 0.00

Table 8: All results for Circles dataset: sample averages +/- one standard deviation over all counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	0.28 ± 0.01**	0.68 ± 0.01	0.00 ± 0.00	0.70 ± 0.01	0.00 ± 0.00	0.53 ± 0.01
	ECCCo-L1 (no CP)	0.27 ± 0.01**	0.68 ± 0.01	0.00 ± 0.00	0.70 ± 0.01	0.00 ± 0.00	0.53 ± 0.01
	ECCCo-L1 (no EBM)	0.28 ± 0.01**	0.68 ± 0.01	0.00 ± 0.00	0.70 ± 0.01	0.00 ± 0.00	0.53 ± 0.01
	ECCCo	0.69 ± 0.03	0.69 ± 0.00	0.00 ± 0.00	0.76 ± 0.02	0.00 ± 0.00	0.53 ± 0.00
	ECCCo (no CP)	0.68 ± 0.03	0.69 ± 0.00	0.00 ± 0.00	0.76 ± 0.02	0.00 ± 0.00	0.53 ± 0.00
	ECCCo (no EBM)	0.69 ± 0.02	0.68 ± 0.01	0.00 ± 0.00	0.70 ± 0.01	0.00 ± 0.00	0.53 ± 0.01
	REVISE	0.32 ± 0.01**	0.68 ± 0.00	0.00 ± 0.00	0.94 ± 0.01	0.00 ± 0.00	0.47 ± 0.00
	Schut	0.49 ± 0.02**	0.29 ± 0.00**	0.00 ± 0.00	0.48 ± 0.01**	0.41 ± 0.01**	1.00 ± 0.00**
	Wachter	0.70 ± 0.02	0.68 ± 0.01	0.00 ± 0.00	0.70 ± 0.01	0.00 ± 0.00	0.53 ± 0.01
JEM Ensemble	ECCCo-L1	0.21 ± 0.01**	0.43 ± 0.03	0.01 ± 0.01**	0.60 ± 0.02	0.00 ± 0.00	0.76 ± 0.04
	ECCCo-L1 (no CP)	0.21 ± 0.01**	0.43 ± 0.03	0.02 ± 0.01	0.60 ± 0.02	0.00 ± 0.00	0.76 ± 0.04
	ECCCo-L1 (no EBM)	0.22 ± 0.01**	0.43 ± 0.03	0.00 ± 0.00**	0.60 ± 0.02	0.00 ± 0.00	0.76 ± 0.04
	ECCCo	0.52 ± 0.03	0.51 ± 0.02	0.00 ± 0.00**	0.74 ± 0.03	0.00 ± 0.00	0.72 ± 0.04
	ECCCo (no CP)	0.52 ± 0.03	0.51 ± 0.02	0.00 ± 0.00**	0.74 ± 0.03	0.00 ± 0.00	0.72 ± 0.04
	ECCCo (no EBM)	0.52 ± 0.03	0.43 ± 0.03	0.00 ± 0.00**	0.60 ± 0.02	0.00 ± 0.00	0.76 ± 0.04
	REVISE	0.28 ± 0.01**	0.68 ± 0.00	0.00 ± 0.00**	0.93 ± 0.01	0.00 ± 0.00	0.47 ± 0.00
	Schut	0.45 ± 0.02**	0.26 ± 0.00**	0.00 ± 0.00**	0.58 ± 0.01**	0.38 ± 0.01**	1.00 ± 0.00**
	Wachter	0.53 ± 0.03	0.43 ± 0.03	0.02 ± 0.01	0.60 ± 0.02	0.00 ± 0.00	0.76 ± 0.04
MLP	ECCCo-L1	0.29 ± 0.01**	0.28 ± 0.01	0.01 ± 0.00**	0.58 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.29 ± 0.01**	0.28 ± 0.00	0.02 ± 0.01**	0.58 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.30 ± 0.01**	0.28 ± 0.01	0.01 ± 0.01**	0.57 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.71 ± 0.04*	0.56 ± 0.00	0.00 ± 0.00**	1.38 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	0.71 ± 0.04*	0.56 ± 0.00	0.00 ± 0.00**	1.38 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	0.77 ± 0.02	0.28 ± 0.01	0.01 ± 0.01**	0.57 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.64 ± 0.00**	0.68 ± 0.00	0.00 ± 0.00**	0.94 ± 0.01	0.00 ± 0.00	0.47 ± 0.00
	Schut	0.78 ± 0.02	0.25 ± 0.00**	0.00 ± 0.00**	0.55 ± 0.01**	0.39 ± 0.01**	1.00 ± 0.00
	Wachter	0.77 ± 0.02	0.29 ± 0.00	0.04 ± 0.01	0.58 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
MLP Ensemble	ECCCo-L1	0.29 ± 0.01**	0.27 ± 0.01	0.00 ± 0.00	0.58 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.29 ± 0.01**	0.28 ± 0.01	0.00 ± 0.00	0.58 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.30 ± 0.01**	0.28 ± 0.01	0.00 ± 0.00	0.58 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.71 ± 0.04*	0.57 ± 0.00	0.00 ± 0.00	1.38 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	0.71 ± 0.04*	0.57 ± 0.00	0.00 ± 0.00	1.38 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	0.77 ± 0.02	0.28 ± 0.01	0.00 ± 0.00	0.58 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.64 ± 0.00**	0.68 ± 0.00	0.00 ± 0.00	0.93 ± 0.01	0.00 ± 0.00	0.47 ± 0.00
	Schut	0.78 ± 0.02	0.25 ± 0.00**	0.00 ± 0.00	0.58 ± 0.01	0.38 ± 0.02**	1.00 ± 0.00
	Wachter	0.77 ± 0.02	0.28 ± 0.01	0.00 ± 0.00	0.58 ± 0.02	0.00 ± 0.00	1.00 ± 0.00

Table 9: All results for Moons dataset: sample averages +/- one standard deviation over all counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	0.29 ± 0.02**	0.42 ± 0.04	0.14 ± 0.03	1.02 ± 0.03	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.29 ± 0.02**	0.42 ± 0.04	0.15 ± 0.03	1.02 ± 0.03	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.30 ± 0.02**	0.42 ± 0.04	0.13 ± 0.03	1.03 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.55 ± 0.06**	0.15 ± 0.01**	0.01 ± 0.01**	1.28 ± 0.04	0.00 ± 0.00	1.00 ± 0.00**
	ECCCo (no CP)	0.57 ± 0.05**	0.15 ± 0.01**	0.01 ± 0.01**	1.28 ± 0.04	0.00 ± 0.00	1.00 ± 0.00**
	ECCCo (no EBM)	0.69 ± 0.04	0.42 ± 0.04	0.13 ± 0.03	1.03 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.54 ± 0.05**	0.31 ± 0.02**	0.15 ± 0.06	1.41 ± 0.12	0.00 ± 0.00	0.84 ± 0.07
	Schut	0.72 ± 0.05	0.46 ± 0.04	0.19 ± 0.02	1.12 ± 0.03	0.06 ± 0.02**	1.00 ± 0.01
	Wachter	0.69 ± 0.05	0.42 ± 0.04	0.14 ± 0.03	1.03 ± 0.03	0.00 ± 0.00	1.00 ± 0.00
JEM Ensemble	ECCCo-L1	0.26 ± 0.01**	0.18 ± 0.01	0.02 ± 0.01	1.04 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.26 ± 0.02**	0.18 ± 0.01	0.04 ± 0.01	1.03 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.27 ± 0.01**	0.18 ± 0.01	0.00 ± 0.00*	1.03 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.47 ± 0.03**	0.13 ± 0.00**	0.00 ± 0.00**	1.50 ± 0.04	0.00 ± 0.00	1.00 ± 0.00**
	ECCCo (no CP)	0.46 ± 0.04**	0.13 ± 0.00**	0.00 ± 0.00**	1.50 ± 0.04	0.00 ± 0.00	1.00 ± 0.00**
	ECCCo (no EBM)	0.68 ± 0.02	0.18 ± 0.01	0.00 ± 0.00*	1.03 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.64 ± 0.02*	0.34 ± 0.01	0.40 ± 0.05	1.39 ± 0.07	0.00 ± 0.00	0.70 ± 0.04
	Schut	0.52 ± 0.03**	0.24 ± 0.01	0.00 ± 0.00**	1.81 ± 0.03	0.05 ± 0.01**	1.00 ± 0.00
	Wachter	0.67 ± 0.03	0.18 ± 0.01	0.01 ± 0.01	1.02 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
MLP	ECCCo-L1	0.45 ± 0.02**	0.34 ± 0.02	0.30 ± 0.02	1.49 ± 0.07	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.45 ± 0.03**	0.35 ± 0.02	0.31 ± 0.02	1.48 ± 0.07	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.46 ± 0.02**	0.35 ± 0.02	0.30 ± 0.02	1.49 ± 0.07	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	1.33 ± 0.06**	0.59 ± 0.03	0.00 ± 0.00**	2.84 ± 0.09	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	1.32 ± 0.09*	0.59 ± 0.03	0.00 ± 0.00**	2.84 ± 0.09	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	1.50 ± 0.04	0.35 ± 0.02	0.30 ± 0.02	1.49 ± 0.07	0.00 ± 0.00	1.00 ± 0.00
	REVISE	1.49 ± 0.06	0.27 ± 0.01**	0.29 ± 0.03	1.22 ± 0.07**	0.00 ± 0.00	0.93 ± 0.03
	Schut	1.58 ± 0.09	0.52 ± 0.04	0.00 ± 0.00**	0.72 ± 0.02**	0.21 ± 0.02**	0.79 ± 0.04
	Wachter	1.50 ± 0.04	0.35 ± 0.02	0.30 ± 0.02	1.48 ± 0.07	0.00 ± 0.00	1.00 ± 0.00
MLP Ensemble	ECCCo-L1	0.44 ± 0.02**	0.28 ± 0.02	0.22 ± 0.02	1.43 ± 0.08	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.44 ± 0.02**	0.28 ± 0.02	0.24 ± 0.03	1.43 ± 0.08	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.46 ± 0.02**	0.28 ± 0.02	0.22 ± 0.02*	1.42 ± 0.08	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	1.24 ± 0.07**	0.56 ± 0.02	0.00 ± 0.00**	2.81 ± 0.08	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	1.23 ± 0.09**	0.56 ± 0.02	0.00 ± 0.00**	2.81 ± 0.08	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	1.50 ± 0.03	0.28 ± 0.02	0.22 ± 0.02*	1.42 ± 0.08	0.00 ± 0.00	1.00 ± 0.00
	REVISE	1.58 ± 0.06	0.26 ± 0.01**	0.31 ± 0.03	1.25 ± 0.07**	0.00 ± 0.00	0.94 ± 0.03
	Schut	1.64 ± 0.09	0.36 ± 0.02	0.03 ± 0.01**	0.70 ± 0.02**	0.24 ± 0.03**	0.89 ± 0.03
	Wachter	1.51 ± 0.04	0.29 ± 0.02	0.24 ± 0.02	1.42 ± 0.08	0.00 ± 0.00	1.00 ± 0.00

Table 10: All results for California Housing dataset: sample averages +/- one standard deviation over all counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	0.98 ± 0.05**	1.04 ± 0.12	0.24 ± 0.02	1.22 ± 0.07	0.00 ± 0.00	0.98 ± 0.01
	ECCCo-L1 (no CP)	0.97 ± 0.05**	1.04 ± 0.12	0.25 ± 0.02	1.18 ± 0.07	0.00 ± 0.00	0.98 ± 0.01
	ECCCo-L1 (no EBM)	1.00 ± 0.06**	1.07 ± 0.12	0.24 ± 0.02	1.12 ± 0.07	0.01 ± 0.00	0.98 ± 0.01
	ECCCo	1.94 ± 0.13*	0.78 ± 0.08**	0.18 ± 0.00**	2.05 ± 0.15	0.01 ± 0.00	1.00 ± 0.00**
	ECCCo+	1.84 ± 0.12**	0.71 ± 0.07**	0.19 ± 0.00**	3.29 ± 0.18	0.00 ± 0.00	1.00 ± 0.01**
	ECCCo (no CP)	1.95 ± 0.14*	0.78 ± 0.08**	0.18 ± 0.00**	2.03 ± 0.16	0.00 ± 0.00	1.00 ± 0.00**
	ECCCo (no EBM)	2.13 ± 0.11	1.07 ± 0.12	0.24 ± 0.02	1.12 ± 0.07	0.01 ± 0.00	0.98 ± 0.01
	REVISE	1.89 ± 0.15*	0.62 ± 0.04**	0.27 ± 0.04	5.40 ± 0.31	0.00 ± 0.00	0.67 ± 0.06
	Schut	2.05 ± 0.13	1.14 ± 0.12	0.21 ± 0.01**	2.82 ± 0.08	0.29 ± 0.01**	0.99 ± 0.01**
	Wachter	2.14 ± 0.13	1.07 ± 0.12	0.25 ± 0.02	1.08 ± 0.07	0.01 ± 0.00	0.98 ± 0.01
JEM Ensemble	ECCCo-L1	0.89 ± 0.05**	1.00 ± 0.17	0.13 ± 0.00**	1.38 ± 0.07	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.89 ± 0.06**	1.01 ± 0.17	0.14 ± 0.00*	1.34 ± 0.07	0.01 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.92 ± 0.06**	1.02 ± 0.18	0.14 ± 0.00*	1.27 ± 0.08	0.01 ± 0.00	1.00 ± 0.00
	ECCCo	1.41 ± 0.08**	0.71 ± 0.09**	0.11 ± 0.00**	2.28 ± 0.16	0.00 ± 0.00	1.00 ± 0.00
	ECCCo+	1.29 ± 0.08**	0.62 ± 0.08**	0.11 ± 0.00**	3.41 ± 0.19	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	1.40 ± 0.08**	0.71 ± 0.09**	0.11 ± 0.00**	2.25 ± 0.16	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	1.71 ± 0.11	1.02 ± 0.18	0.14 ± 0.00*	1.27 ± 0.08	0.01 ± 0.00	1.00 ± 0.00
	REVISE	1.42 ± 0.14**	0.63 ± 0.04**	0.21 ± 0.05	5.29 ± 0.38	0.00 ± 0.00	0.62 ± 0.06
	Schut	1.61 ± 0.11	1.13 ± 0.18	0.09 ± 0.00**	3.08 ± 0.10	0.31 ± 0.02**	1.00 ± 0.00
	Wachter	1.72 ± 0.11	1.02 ± 0.18	0.14 ± 0.00	1.23 ± 0.07	0.01 ± 0.00	1.00 ± 0.00
MLP	ECCCo-L1	1.22 ± 0.04**	1.18 ± 0.09	0.17 ± 0.01*	1.38 ± 0.09	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	1.23 ± 0.04**	1.18 ± 0.09	0.17 ± 0.01	1.35 ± 0.09	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	1.25 ± 0.04**	1.19 ± 0.09	0.17 ± 0.01	1.26 ± 0.09	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	3.81 ± 0.10*	2.31 ± 0.09	0.14 ± 0.01**	6.19 ± 0.14	0.00 ± 0.00	1.00 ± 0.00
	ECCCo+	3.95 ± 0.10	1.31 ± 0.08	0.15 ± 0.01**	4.62 ± 0.17	0.00 ± 0.00	0.96 ± 0.02
	ECCCo (no CP)	3.80 ± 0.11*	2.31 ± 0.09	0.14 ± 0.01**	6.17 ± 0.14	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	3.98 ± 0.08	1.19 ± 0.09	0.17 ± 0.01	1.26 ± 0.09	0.00 ± 0.00	1.00 ± 0.00
	REVISE	3.91 ± 0.07*	0.61 ± 0.04**	0.17 ± 0.04	5.37 ± 0.31	0.00 ± 0.00	0.73 ± 0.06
	Schut	3.99 ± 0.08	1.15 ± 0.09	0.15 ± 0.01**	1.59 ± 0.09	0.48 ± 0.02**	0.99 ± 0.01
	Wachter	3.98 ± 0.07	1.19 ± 0.09	0.17 ± 0.01	1.23 ± 0.09	0.00 ± 0.00	1.00 ± 0.00
MLP Ensemble	ECCCo-L1	1.23 ± 0.05**	1.17 ± 0.17	0.13 ± 0.01**	1.21 ± 0.08	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	1.22 ± 0.04**	1.17 ± 0.17	0.15 ± 0.01	1.18 ± 0.08	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	1.24 ± 0.04**	1.16 ± 0.17	0.14 ± 0.01**	1.08 ± 0.07	0.01 ± 0.00	1.00 ± 0.00
	ECCCo	3.71 ± 0.10**	1.98 ± 0.17	0.09 ± 0.01**	4.77 ± 0.12	0.01 ± 0.00	1.00 ± 0.00
	ECCCo+	3.92 ± 0.09*	1.25 ± 0.14	0.14 ± 0.02	4.22 ± 0.16	0.00 ± 0.00	0.97 ± 0.02
	ECCCo (no CP)	3.72 ± 0.10**	1.98 ± 0.17	0.10 ± 0.01**	4.77 ± 0.12	0.01 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	4.05 ± 0.08	1.16 ± 0.17	0.14 ± 0.01**	1.08 ± 0.07	0.01 ± 0.00	1.00 ± 0.00
	REVISE	3.98 ± 0.06*	0.61 ± 0.03**	0.15 ± 0.03	5.42 ± 0.34	0.00 ± 0.00	0.74 ± 0.05
	Schut	4.02 ± 0.08	1.20 ± 0.18	0.10 ± 0.01**	1.55 ± 0.08	0.58 ± 0.02**	1.00 ± 0.00
	Wachter	4.06 ± 0.09	1.16 ± 0.17	0.16 ± 0.01	1.05 ± 0.07	0.01 ± 0.00	1.00 ± 0.00

Table 11: All results for GMSC dataset: sample averages +/- one standard deviation over all counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	0.94 ± 0.05**	1.11 ± 0.18	0.29 ± 0.01**	1.03 ± 0.04	0.13 ± 0.01	0.98 ± 0.01
	ECCCo-L1 (no CP)	0.96 ± 0.05**	1.10 ± 0.18	0.31 ± 0.01*	0.96 ± 0.04	0.14 ± 0.01	0.98 ± 0.01
	ECCCo-L1 (no EBM)	0.97 ± 0.05**	1.09 ± 0.18	0.31 ± 0.01*	0.89 ± 0.03	0.14 ± 0.01	0.98 ± 0.01
	ECCCo	1.85 ± 0.12	1.01 ± 0.18	0.34 ± 0.01	1.12 ± 0.05	0.13 ± 0.01	0.99 ± 0.01*
	ECCCo+	1.72 ± 0.13*	0.90 ± 0.18*	0.35 ± 0.01	2.59 ± 0.12	0.09 ± 0.01	0.99 ± 0.01*
	ECCCo (no CP)	1.86 ± 0.12	1.00 ± 0.18	0.36 ± 0.01	1.06 ± 0.05	0.13 ± 0.01	0.99 ± 0.01*
	ECCCo (no EBM)	1.91 ± 0.11	1.09 ± 0.18	0.31 ± 0.01*	0.89 ± 0.03	0.14 ± 0.01	0.98 ± 0.01
	REVISE	1.67 ± 0.11**	0.74 ± 0.13**	0.31 ± 0.05	4.46 ± 0.23	0.04 ± 0.01	0.61 ± 0.06
	Schut	1.94 ± 0.12	1.35 ± 0.17	0.25 ± 0.01**	1.92 ± 0.05	0.46 ± 0.01**	0.98 ± 0.01
	Wachter	1.93 ± 0.12	1.08 ± 0.18	0.32 ± 0.01	0.82 ± 0.03	0.14 ± 0.01	0.98 ± 0.01
JEM Ensemble	ECCCo-L1	0.88 ± 0.05**	1.15 ± 0.20	0.30 ± 0.01**	1.13 ± 0.04	0.14 ± 0.01	0.98 ± 0.01
	ECCCo-L1 (no CP)	0.89 ± 0.06**	1.15 ± 0.20	0.31 ± 0.01	1.05 ± 0.05	0.14 ± 0.01	0.98 ± 0.01
	ECCCo-L1 (no EBM)	0.90 ± 0.05**	1.15 ± 0.20	0.31 ± 0.01*	1.00 ± 0.04	0.17 ± 0.01	0.98 ± 0.01
	ECCCo	1.29 ± 0.11*	0.94 ± 0.19	0.37 ± 0.01	1.11 ± 0.06	0.13 ± 0.01	0.99 ± 0.01
	ECCCo+	1.08 ± 0.11**	0.85 ± 0.17*	0.36 ± 0.01	2.64 ± 0.17	0.10 ± 0.01	0.99 ± 0.01*
	ECCCo (no CP)	1.29 ± 0.12	0.94 ± 0.19*	0.39 ± 0.01	1.04 ± 0.06	0.13 ± 0.01	0.99 ± 0.01
	ECCCo (no EBM)	1.40 ± 0.11	1.15 ± 0.20	0.31 ± 0.01*	1.00 ± 0.04	0.17 ± 0.01	0.98 ± 0.01
	REVISE	1.05 ± 0.09**	0.74 ± 0.13**	0.30 ± 0.04	4.40 ± 0.28	0.05 ± 0.02	0.60 ± 0.06
	Schut	1.42 ± 0.12	1.38 ± 0.20	0.25 ± 0.01**	2.08 ± 0.05	0.44 ± 0.01**	0.98 ± 0.01
	Wachter	1.40 ± 0.11	1.13 ± 0.21	0.33 ± 0.01	0.92 ± 0.04	0.17 ± 0.01	0.98 ± 0.01
MLP	ECCCo-L1	1.21 ± 0.03**	1.06 ± 0.11	0.28 ± 0.01*	1.39 ± 0.15	0.14 ± 0.01	0.97 ± 0.01
	ECCCo-L1 (no CP)	1.21 ± 0.03**	1.05 ± 0.11	0.29 ± 0.01	1.34 ± 0.14	0.14 ± 0.01	0.97 ± 0.01
	ECCCo-L1 (no EBM)	1.23 ± 0.04**	1.03 ± 0.11	0.29 ± 0.01	1.25 ± 0.15	0.14 ± 0.01	0.97 ± 0.01
	ECCCo	3.84 ± 0.08	1.69 ± 0.10	0.24 ± 0.01**	3.25 ± 0.21	0.14 ± 0.01	0.98 ± 0.01
	ECCCo+	3.76 ± 0.08*	1.14 ± 0.08	0.26 ± 0.01**	3.91 ± 0.20	0.10 ± 0.01	0.97 ± 0.01
	ECCCo (no CP)	3.84 ± 0.09	1.69 ± 0.10	0.24 ± 0.01**	3.23 ± 0.21	0.14 ± 0.01	0.98 ± 0.01
	ECCCo (no EBM)	3.83 ± 0.07	1.03 ± 0.11	0.29 ± 0.01	1.25 ± 0.15	0.14 ± 0.01	0.97 ± 0.01
	REVISE	3.78 ± 0.06*	0.63 ± 0.02**	0.28 ± 0.04	4.52 ± 0.28	0.04 ± 0.01	0.69 ± 0.05
	Schut	3.89 ± 0.09	1.20 ± 0.15	0.28 ± 0.02	1.51 ± 0.07	0.50 ± 0.02**	0.93 ± 0.02
	Wachter	3.85 ± 0.08	1.02 ± 0.11	0.30 ± 0.01	1.19 ± 0.15	0.14 ± 0.01	0.97 ± 0.01
MLP Ensemble	ECCCo-L1	1.21 ± 0.03**	1.04 ± 0.10	0.29 ± 0.01**	1.46 ± 0.22	0.14 ± 0.01	0.99 ± 0.01
	ECCCo-L1 (no CP)	1.21 ± 0.02**	1.03 ± 0.10	0.31 ± 0.01*	1.39 ± 0.22	0.14 ± 0.01	0.99 ± 0.01
	ECCCo-L1 (no EBM)	1.24 ± 0.03**	1.00 ± 0.10	0.31 ± 0.01*	1.27 ± 0.22	0.14 ± 0.01	0.99 ± 0.01
	ECCCo	3.86 ± 0.08**	2.16 ± 0.09	0.23 ± 0.01**	4.41 ± 0.25	0.13 ± 0.01	0.99 ± 0.01
	ECCCo+	3.81 ± 0.06**	1.84 ± 0.07	0.30 ± 0.01*	5.45 ± 0.24	0.10 ± 0.01	0.99 ± 0.01
	ECCCo (no CP)	3.87 ± 0.07**	2.15 ± 0.10	0.23 ± 0.01**	4.37 ± 0.25	0.13 ± 0.01	0.99 ± 0.01
	ECCCo (no EBM)	4.10 ± 0.07	1.00 ± 0.10	0.31 ± 0.01*	1.27 ± 0.22	0.14 ± 0.01	0.99 ± 0.01
	REVISE	4.08 ± 0.06	0.63 ± 0.02**	0.29 ± 0.04	4.54 ± 0.36	0.04 ± 0.01	0.66 ± 0.04
	Schut	4.10 ± 0.10	1.32 ± 0.17	0.30 ± 0.01*	1.69 ± 0.06	0.56 ± 0.02**	0.98 ± 0.01
	Wachter	4.11 ± 0.07	0.99 ± 0.10	0.32 ± 0.01	1.20 ± 0.22	0.14 ± 0.01	0.99 ± 0.01

Table 12: All results for German Credit dataset: sample averages +/- one standard deviation over all counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	1.69 ± 0.03**	5.16 ± 0.08	0.59 ± 0.01**	3.06 ± 0.19	0.29 ± 0.01	1.00 ± 0.01
	ECCCo-L1 (no CP)	1.69 ± 0.03**	5.16 ± 0.07	0.60 ± 0.01	2.98 ± 0.19	0.29 ± 0.01	1.00 ± 0.01
	ECCCo-L1 (no EBM)	1.70 ± 0.04**	5.18 ± 0.08	0.59 ± 0.01**	3.04 ± 0.19	0.30 ± 0.01	1.00 ± 0.01
	ECCCo	4.47 ± 0.09*	5.07 ± 0.07*	0.60 ± 0.01	3.65 ± 0.28	0.27 ± 0.01	1.00 ± 0.01
	ECCCo+	4.08 ± 0.07**	3.82 ± 0.06**	0.22 ± 0.01**	13.04 ± 0.29	0.09 ± 0.01	0.39 ± 0.01
	ECCCo (no CP)	4.48 ± 0.10	5.06 ± 0.07*	0.61 ± 0.01	3.59 ± 0.28	0.27 ± 0.01	1.00 ± 0.01
	ECCCo (no EBM)	4.55 ± 0.09	5.18 ± 0.08	0.59 ± 0.01**	3.04 ± 0.19	0.30 ± 0.01	1.00 ± 0.01
	REVISE	4.11 ± 0.09**	3.88 ± 0.05**	0.21 ± 0.01**	18.00 ± 0.41	0.04 ± 0.01	0.39 ± 0.00
	Schut	4.58 ± 0.08	5.22 ± 0.08	0.44 ± 0.03**	2.44 ± 0.09**	0.80 ± 0.01**	0.75 ± 0.03
	Wachter	4.56 ± 0.08	5.17 ± 0.08	0.60 ± 0.01	2.96 ± 0.19	0.30 ± 0.01	1.00 ± 0.01
JEM Ensemble	ECCCo-L1	1.61 ± 0.04**	5.07 ± 0.09	0.88 ± 0.01**	3.51 ± 0.18	0.28 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no CP)	1.60 ± 0.04**	5.08 ± 0.10	0.90 ± 0.01	3.45 ± 0.19	0.28 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no EBM)	1.62 ± 0.04**	5.09 ± 0.10	0.88 ± 0.01**	3.49 ± 0.19	0.29 ± 0.01	1.00 ± 0.00
	ECCCo	3.72 ± 0.11**	4.13 ± 0.06**	0.79 ± 0.02**	7.44 ± 0.31	0.17 ± 0.01	0.95 ± 0.02
	ECCCo+	3.62 ± 0.11**	3.61 ± 0.04**	0.24 ± 0.00**	13.88 ± 0.38	0.08 ± 0.00	0.39 ± 0.00
	ECCCo (no CP)	3.70 ± 0.13**	4.13 ± 0.06**	0.80 ± 0.02**	7.43 ± 0.31	0.17 ± 0.01	0.95 ± 0.02
	ECCCo (no EBM)	4.24 ± 0.12	5.09 ± 0.10	0.88 ± 0.01**	3.49 ± 0.19	0.29 ± 0.01	1.00 ± 0.00
	REVISE	3.41 ± 0.13**	3.80 ± 0.07**	0.20 ± 0.00**	17.46 ± 0.51	0.03 ± 0.01	0.39 ± 0.00
	Schut	4.17 ± 0.11	5.05 ± 0.10	0.58 ± 0.04**	3.38 ± 0.08	0.76 ± 0.01**	0.74 ± 0.04
	Wachter	4.23 ± 0.12	5.10 ± 0.10	0.90 ± 0.01	3.42 ± 0.18	0.29 ± 0.01	1.00 ± 0.00
MLP	ECCCo-L1	1.77 ± 0.03**	4.61 ± 0.09	0.76 ± 0.00*	4.98 ± 0.21	0.23 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no CP)	1.77 ± 0.03**	4.61 ± 0.09	0.77 ± 0.00	4.93 ± 0.21	0.23 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no EBM)	1.77 ± 0.03**	4.63 ± 0.09	0.76 ± 0.00*	4.95 ± 0.21	0.23 ± 0.01	1.00 ± 0.00**
	ECCCo	5.36 ± 0.07	4.63 ± 0.09	0.81 ± 0.00	4.83 ± 0.21	0.24 ± 0.01**	1.00 ± 0.00**
	ECCCo+	5.10 ± 0.06**	3.55 ± 0.02**	0.22 ± 0.00**	13.50 ± 0.29	0.02 ± 0.00	0.39 ± 0.00
	ECCCo (no CP)	5.35 ± 0.07	4.62 ± 0.09	0.82 ± 0.00	4.78 ± 0.21	0.23 ± 0.01*	1.00 ± 0.00**
	ECCCo (no EBM)	5.35 ± 0.06	4.63 ± 0.09	0.76 ± 0.00*	4.95 ± 0.21	0.23 ± 0.01	1.00 ± 0.00**
	REVISE	5.18 ± 0.06**	3.89 ± 0.07**	0.20 ± 0.00**	17.21 ± 0.53	0.02 ± 0.01	0.39 ± 0.00
	Schut	5.40 ± 0.07	4.74 ± 0.09	0.37 ± 0.03**	3.29 ± 0.06**	0.72 ± 0.01**	0.56 ± 0.03
	Wachter	5.34 ± 0.06	4.63 ± 0.09	0.77 ± 0.01	4.90 ± 0.21	0.23 ± 0.01	1.00 ± 0.00
MLP Ensemble	ECCCo-L1	1.64 ± 0.03**	4.16 ± 0.05	0.78 ± 0.02*	7.03 ± 0.28	0.23 ± 0.01	0.95 ± 0.02
	ECCCo-L1 (no CP)	1.64 ± 0.03**	4.16 ± 0.05	0.80 ± 0.02	6.95 ± 0.29	0.24 ± 0.01	0.95 ± 0.02
	ECCCo-L1 (no EBM)	1.65 ± 0.03**	4.17 ± 0.05	0.78 ± 0.02*	6.99 ± 0.28	0.24 ± 0.01	0.95 ± 0.02
	ECCCo	4.48 ± 0.06	4.16 ± 0.05	0.82 ± 0.02	6.95 ± 0.29	0.23 ± 0.01	0.96 ± 0.02
	ECCCo+	4.62 ± 0.08	3.43 ± 0.01**	0.22 ± 0.00**	13.11 ± 0.33	0.02 ± 0.00	0.39 ± 0.00
	ECCCo (no CP)	4.49 ± 0.07	4.16 ± 0.05	0.83 ± 0.02	6.89 ± 0.29	0.24 ± 0.01	0.96 ± 0.02
	ECCCo (no EBM)	4.47 ± 0.07	4.17 ± 0.05	0.78 ± 0.02*	6.99 ± 0.28	0.24 ± 0.01	0.95 ± 0.02
	REVISE	4.78 ± 0.09	3.85 ± 0.06**	0.20 ± 0.00**	16.71 ± 0.46	0.03 ± 0.01	0.39 ± 0.00
	Schut	4.84 ± 0.07	4.52 ± 0.06	0.26 ± 0.02**	3.76 ± 0.06**	0.71 ± 0.01**	0.44 ± 0.02
	Wachter	4.49 ± 0.06	4.17 ± 0.05	0.80 ± 0.02	6.91 ± 0.29	0.24 ± 0.01	0.95 ± 0.02

Table 13: All results for MNIST dataset: sample averages +/- one standard deviation over all counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	0.10 ± 0.00**	0.35 ± 0.00	4.59 ± 0.01**	23.75 ± 0.97	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.25 ± 0.00**	0.46 ± 0.00	4.50 ± 0.00**	160.43 ± 1.31	0.00 ± 0.00	1.00 ± 0.00
	ECCCo+	0.25 ± 0.00**	0.31 ± 0.00**	4.51 ± 0.00**	196.62 ± 3.62	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.00	0.29 ± 0.00**	3.82 ± 0.14**	239.71 ± 6.72	0.00 ± 0.00	0.96 ± 0.02
	Schut	0.25 ± 0.00	0.34 ± 0.00**	0.73 ± 0.15**	7.01 ± 0.18**	0.99 ± 0.00**	0.16 ± 0.03
	Wachter	0.25 ± 0.00	0.35 ± 0.00	4.62 ± 0.01	22.90 ± 0.99	0.00 ± 0.00	1.00 ± 0.00
JEM Ensemble	ECCCo-L1	0.10 ± 0.00**	0.34 ± 0.00	2.44 ± 0.01**	35.89 ± 1.47	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.25 ± 0.00**	0.45 ± 0.00	1.62 ± 0.01**	157.21 ± 1.17	0.00 ± 0.00	1.00 ± 0.00
	ECCCo+	0.24 ± 0.00**	0.30 ± 0.00**	1.58 ± 0.01**	193.39 ± 2.21	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.00**	0.29 ± 0.00**	1.75 ± 0.09**	233.72 ± 3.16	0.00 ± 0.00	0.96 ± 0.02
	Schut	0.25 ± 0.00	0.34 ± 0.00**	0.14 ± 0.04**	7.82 ± 0.07**	0.99 ± 0.00**	0.06 ± 0.01
	Wachter	0.25 ± 0.00	0.34 ± 0.00	2.53 ± 0.01	34.89 ± 1.45	0.00 ± 0.00	1.00 ± 0.00
LeNet-5	ECCCo-L1	0.10 ± 0.00**	0.34 ± 0.00	8.62 ± 0.09	46.71 ± 1.28	0.00 ± 0.00	0.99 ± 0.01
	ECCCo	0.25 ± 0.00**	0.39 ± 0.00	0.01 ± 0.03**	121.18 ± 1.95	0.00 ± 0.00**	1.00 ± 0.00**
	ECCCo+	0.25 ± 0.00**	0.31 ± 0.00**	0.16 ± 0.10**	173.48 ± 1.80	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.00**	0.30 ± 0.00**	2.08 ± 0.21**	228.22 ± 3.77	0.00 ± 0.00	0.85 ± 0.04
	Schut	0.25 ± 0.00	0.34 ± 0.00	0.06 ± 0.06**	8.91 ± 0.08**	0.98 ± 0.00**	0.01 ± 0.01
	Wachter	0.25 ± 0.00	0.34 ± 0.00	8.69 ± 0.06	45.50 ± 1.26	0.00 ± 0.00	0.99 ± 0.01
MLP	ECCCo-L1	0.10 ± 0.00**	0.35 ± 0.00	1.08 ± 0.00**	41.48 ± 1.47	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.24 ± 0.00**	0.42 ± 0.00	0.86 ± 0.00**	131.37 ± 1.79	0.00 ± 0.00**	1.00 ± 0.00
	ECCCo+	0.25 ± 0.00*	0.31 ± 0.00**	0.86 ± 0.00**	175.89 ± 2.75	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.00	0.30 ± 0.00**	0.68 ± 0.04**	224.01 ± 3.66	0.00 ± 0.00	0.86 ± 0.02
	Schut	0.25 ± 0.00	0.34 ± 0.00**	0.04 ± 0.02**	8.18 ± 0.16**	0.99 ± 0.00**	0.04 ± 0.02
	Wachter	0.25 ± 0.00	0.34 ± 0.00	1.16 ± 0.00	39.80 ± 1.44	0.00 ± 0.00	1.00 ± 0.00
MLP Ensemble	ECCCo-L1	0.10 ± 0.00**	0.34 ± 0.00	0.61 ± 0.01**	45.56 ± 1.46	0.00 ± 0.00	0.99 ± 0.01
	ECCCo	0.24 ± 0.00**	0.40 ± 0.00	0.33 ± 0.00**	116.67 ± 1.95	0.00 ± 0.00**	1.00 ± 0.00**
	ECCCo+	0.25 ± 0.00**	0.30 ± 0.00**	0.34 ± 0.00**	174.80 ± 3.42	0.00 ± 0.00	0.99 ± 0.01
	REVISE	0.25 ± 0.00	0.30 ± 0.00**	0.33 ± 0.03**	224.93 ± 3.99	0.00 ± 0.00	0.87 ± 0.04
	Schut	0.25 ± 0.00	0.34 ± 0.00	0.01 ± 0.01**	8.38 ± 0.11**	0.99 ± 0.00**	0.02 ± 0.01
	Wachter	0.25 ± 0.00	0.34 ± 0.00	0.65 ± 0.01	43.98 ± 1.41	0.00 ± 0.00	0.99 ± 0.01

Table 14: All results for Fashion MNIST dataset: sample averages +/- one standard deviation over all counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	0.10 ± 0.00**	0.40 ± 0.00	4.15 ± 0.10	31.58 ± 1.34	0.00 ± 0.00	0.84 ± 0.02*
	ECCCo	0.25 ± 0.00**	0.43 ± 0.00	4.56 ± 0.02	79.06 ± 1.35	0.00 ± 0.00**	1.00 ± 0.00**
	ECCCo+	0.25 ± 0.00**	0.38 ± 0.00**	2.98 ± 0.15**	184.42 ± 4.53	0.00 ± 0.00	0.62 ± 0.03
	REVISE	0.25 ± 0.00	0.37 ± 0.00**	0.74 ± 0.17**	196.14 ± 4.92	0.00 ± 0.00	0.16 ± 0.04
	Schut	0.25 ± 0.00	0.40 ± 0.00*	0.88 ± 0.13**	8.21 ± 0.10**	0.99 ± 0.00**	0.18 ± 0.03
	Wachter	0.25 ± 0.00	0.40 ± 0.00	4.07 ± 0.11	29.91 ± 1.30	0.00 ± 0.00	0.81 ± 0.02
JEM Ensemble	ECCCo-L1	0.10 ± 0.00**	0.40 ± 0.00	0.79 ± 0.10	40.74 ± 1.07	0.00 ± 0.00	0.26 ± 0.04*
	ECCCo	0.25 ± 0.00**	0.42 ± 0.00	2.68 ± 0.04	112.32 ± 2.27	0.00 ± 0.00*	0.91 ± 0.02**
	ECCCo+	0.25 ± 0.00**	0.37 ± 0.00**	1.61 ± 0.13	193.58 ± 3.17	0.00 ± 0.00	0.55 ± 0.04**
	REVISE	0.25 ± 0.00	0.37 ± 0.00**	0.46 ± 0.12*	219.25 ± 4.89	0.00 ± 0.00	0.15 ± 0.03
	Schut	0.25 ± 0.00	0.39 ± 0.00	0.23 ± 0.06**	8.93 ± 0.12**	0.99 ± 0.00**	0.08 ± 0.02
	Wachter	0.25 ± 0.00	0.39 ± 0.00	0.65 ± 0.09	36.91 ± 1.07	0.00 ± 0.00	0.22 ± 0.03
LeNet-5	ECCCo-L1	0.10 ± 0.00**	0.40 ± 0.00	0.83 ± 0.08	30.24 ± 0.90	0.00 ± 0.00	0.24 ± 0.02*
	ECCCo	0.25 ± 0.00**	0.42 ± 0.00	2.60 ± 0.07	94.33 ± 1.90	0.00 ± 0.00**	0.93 ± 0.02**
	ECCCo+	0.25 ± 0.00**	0.38 ± 0.00**	1.79 ± 0.16	173.75 ± 2.75	0.00 ± 0.00	0.59 ± 0.05**
	REVISE	0.25 ± 0.00	0.38 ± 0.00**	0.32 ± 0.07**	189.62 ± 3.95	0.00 ± 0.00	0.10 ± 0.02
	Schut	0.25 ± 0.00	0.39 ± 0.00*	0.19 ± 0.07**	8.59 ± 0.14**	0.97 ± 0.00**	0.06 ± 0.02
	Wachter	0.25 ± 0.00	0.40 ± 0.00	0.70 ± 0.05	27.66 ± 0.88	0.00 ± 0.00	0.20 ± 0.01
MLP	ECCCo-L1	0.10 ± 0.00**	0.40 ± 0.00	1.68 ± 0.10	42.02 ± 1.09	0.00 ± 0.00	0.47 ± 0.03**
	ECCCo	0.24 ± 0.00**	0.42 ± 0.00	2.89 ± 0.02	112.70 ± 1.67	0.00 ± 0.00**	0.99 ± 0.01**
	ECCCo+	0.25 ± 0.00**	0.37 ± 0.00**	2.21 ± 0.08	189.26 ± 3.80	0.00 ± 0.00	0.71 ± 0.03**
	REVISE	0.25 ± 0.00	0.37 ± 0.00**	0.45 ± 0.09**	202.31 ± 5.63	0.00 ± 0.00	0.15 ± 0.03
	Schut	0.25 ± 0.00	0.40 ± 0.00*	0.29 ± 0.09**	8.61 ± 0.12**	0.99 ± 0.00**	0.09 ± 0.03
	Wachter	0.25 ± 0.00	0.40 ± 0.00	1.52 ± 0.11	38.73 ± 1.03	0.00 ± 0.00	0.42 ± 0.03
MLP Ensemble	ECCCo-L1	0.10 ± 0.00**	0.39 ± 0.00	0.66 ± 0.12	36.10 ± 1.54	0.00 ± 0.00	0.24 ± 0.04*
	ECCCo	0.24 ± 0.00**	0.41 ± 0.00	1.91 ± 0.04	116.23 ± 2.62	0.00 ± 0.00**	0.92 ± 0.02**
	ECCCo+	0.25 ± 0.00**	0.37 ± 0.00**	1.61 ± 0.08	188.33 ± 3.71	0.00 ± 0.00	0.71 ± 0.04**
	REVISE	0.25 ± 0.00	0.37 ± 0.00**	0.37 ± 0.05**	198.52 ± 6.17	0.00 ± 0.00	0.14 ± 0.03
	Schut	0.25 ± 0.00	0.39 ± 0.00	0.11 ± 0.04**	8.81 ± 0.11**	0.99 ± 0.00**	0.05 ± 0.02
	Wachter	0.25 ± 0.00	0.39 ± 0.00	0.52 ± 0.04	31.82 ± 1.47	0.00 ± 0.00	0.19 ± 0.01

Table 15: All results for Linearly Separable dataset: sample averages +/- one standard deviation over all valid counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	0.06 ± 0.01**	0.16 ± 0.02**	0.00 ± 0.00	0.93 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.06 ± 0.01**	0.16 ± 0.02**	0.00 ± 0.00	0.93 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.09 ± 0.01**	0.24 ± 0.01	0.00 ± 0.00	0.88 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.19 ± 0.01**	0.19 ± 0.01**	0.00 ± 0.00	0.85 ± 0.01**	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	0.19 ± 0.01**	0.19 ± 0.01**	0.00 ± 0.00	0.85 ± 0.01**	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	0.24 ± 0.01	0.24 ± 0.01	0.00 ± 0.00	0.88 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.24 ± 0.02	0.17 ± 0.00**	0.00 ± 0.00	0.42 ± 0.02**	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.29 ± 0.01	0.29 ± 0.00	0.00 ± 0.00	1.40 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	Wachter	0.24 ± 0.01	0.24 ± 0.01	0.00 ± 0.00	0.88 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
JEM Ensemble	ECCCo-L1	0.02 ± 0.00**	0.15 ± 0.01	0.00 ± 0.00	0.98 ± 0.02*	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.02 ± 0.00**	0.15 ± 0.01	0.00 ± 0.00	0.98 ± 0.02*	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.05 ± 0.00**	0.08 ± 0.00	0.00 ± 0.00	1.01 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.12 ± 0.01**	0.07 ± 0.00**	0.00 ± 0.00	0.97 ± 0.02**	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	0.12 ± 0.01**	0.07 ± 0.00**	0.00 ± 0.00	0.97 ± 0.02**	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	0.15 ± 0.01	0.08 ± 0.00	0.00 ± 0.00	1.01 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.01	0.17 ± 0.00	0.45 ± 0.08	0.42 ± 0.02**	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.20 ± 0.01	0.27 ± 0.00	0.00 ± 0.00	1.38 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	Wachter	0.14 ± 0.01	0.08 ± 0.00	0.00 ± 0.00	1.01 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
MLP	ECCCo-L1	0.12 ± 0.00**	0.06 ± 0.00**	0.00 ± 0.00	0.91 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.12 ± 0.00**	0.06 ± 0.00**	0.00 ± 0.00	0.90 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.16 ± 0.00**	0.07 ± 0.00	0.00 ± 0.00	0.62 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.17 ± 0.00**	0.31 ± 0.00	0.00 ± 0.00	1.44 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	0.17 ± 0.00**	0.31 ± 0.00	0.00 ± 0.00	1.44 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	0.46 ± 0.00	0.07 ± 0.00	0.00 ± 0.00	0.62 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.50 ± 0.00	0.18 ± 0.00	0.97 ± 0.02	0.41 ± 0.02**	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.48 ± 0.00	0.31 ± 0.01	0.00 ± 0.00	0.70 ± 0.02	0.46 ± 0.01**	1.00 ± 0.00
	Wachter	0.46 ± 0.00	0.07 ± 0.00	0.00 ± 0.00	0.62 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
MLP Ensemble	ECCCo-L1	0.12 ± 0.00**	0.06 ± 0.00**	0.00 ± 0.00	0.91 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.12 ± 0.00**	0.06 ± 0.00**	0.00 ± 0.00	0.91 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.16 ± 0.00**	0.07 ± 0.00	0.00 ± 0.00	0.65 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.17 ± 0.00**	0.31 ± 0.00	0.00 ± 0.00	1.44 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	0.17 ± 0.00**	0.31 ± 0.00	0.00 ± 0.00	1.44 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	0.45 ± 0.00	0.07 ± 0.00	0.00 ± 0.00	0.65 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.50 ± 0.00	0.18 ± 0.00	0.98 ± 0.01	0.42 ± 0.02**	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.48 ± 0.00	0.32 ± 0.01	0.00 ± 0.00	0.73 ± 0.01	0.42 ± 0.02**	1.00 ± 0.00
	Wachter	0.45 ± 0.00	0.07 ± 0.00	0.00 ± 0.00	0.65 ± 0.01	0.00 ± 0.00	1.00 ± 0.00

Table 16: All results for Circles dataset: sample averages +/- one standard deviation over all valid counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness \downarrow	Implausibility \downarrow	Uncertainty \downarrow	Cost \downarrow	Redundancy \uparrow	Validity \uparrow
JEM	ECCCo-L1	$0.24 \pm 0.01^{**}$	0.33 ± 0.01	0.00 ± 0.00	0.50 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	$0.24 \pm 0.02^{**}$	0.33 ± 0.01	0.00 ± 0.00	0.50 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	$0.25 \pm 0.02^{**}$	0.33 ± 0.01	0.00 ± 0.00	0.50 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.61 ± 0.05	0.33 ± 0.01	0.00 ± 0.00	0.62 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	0.60 ± 0.05	0.33 ± 0.01	0.00 ± 0.00	0.62 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	0.62 ± 0.04	0.33 ± 0.01	0.00 ± 0.00	0.50 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	REVISE	$0.12 \pm 0.02^{**}$	0.43 ± 0.00	0.00 ± 0.00	1.27 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	Schut	$0.49 \pm 0.02^{**}$	$0.29 \pm 0.00^{**}$	0.00 ± 0.00	$0.48 \pm 0.01^{**}$	$0.41 \pm 0.01^{**}$	1.00 ± 0.00
	Wachter	0.63 ± 0.05	0.33 ± 0.01	0.00 ± 0.00	0.50 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
JEM Ensemble	ECCCo-L1	$0.19 \pm 0.01^{**}$	0.30 ± 0.01	$0.01 \pm 0.01^{**}$	0.56 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	$0.19 \pm 0.01^{**}$	0.30 ± 0.01	0.02 ± 0.01	0.56 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	$0.20 \pm 0.01^{**}$	0.30 ± 0.01	$0.00 \pm 0.00^{**}$	0.56 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	$0.45 \pm 0.04^*$	0.35 ± 0.00	$0.00 \pm 0.00^{**}$	0.69 ± 0.03	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	$0.46 \pm 0.03^*$	0.35 ± 0.00	$0.00 \pm 0.00^{**}$	0.69 ± 0.03	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	0.48 ± 0.03	0.30 ± 0.01	$0.00 \pm 0.00^{**}$	0.56 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	REVISE	$0.05 \pm 0.00^{**}$	0.43 ± 0.00	$0.00 \pm 0.00^{**}$	1.26 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	Schut	$0.45 \pm 0.02^*$	$0.26 \pm 0.00^{**}$	$0.00 \pm 0.00^{**}$	0.58 ± 0.01	$0.38 \pm 0.01^{**}$	1.00 ± 0.00
	Wachter	0.49 ± 0.04	0.30 ± 0.01	0.02 ± 0.01	0.56 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
MLP	ECCCo-L1	$0.29 \pm 0.01^{**}$	0.28 ± 0.01	$0.01 \pm 0.00^{**}$	0.58 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	$0.29 \pm 0.01^{**}$	0.28 ± 0.00	$0.02 \pm 0.01^{**}$	0.58 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	$0.30 \pm 0.01^{**}$	0.28 ± 0.01	$0.01 \pm 0.01^{**}$	0.57 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	$0.71 \pm 0.04^*$	0.56 ± 0.00	$0.00 \pm 0.00^{**}$	1.38 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	$0.71 \pm 0.04^*$	0.56 ± 0.00	$0.00 \pm 0.00^{**}$	1.38 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	0.77 ± 0.02	0.28 ± 0.01	$0.01 \pm 0.01^{**}$	0.57 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	REVISE	$0.01 \pm 0.00^{**}$	0.43 ± 0.00	$0.00 \pm 0.00^{**}$	1.27 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.78 ± 0.02	$0.25 \pm 0.00^{**}$	$0.00 \pm 0.00^{**}$	$0.55 \pm 0.01^{**}$	$0.39 \pm 0.01^{**}$	1.00 ± 0.00
	Wachter	0.77 ± 0.02	0.29 ± 0.00	0.04 ± 0.01	0.58 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
MLP Ensemble	ECCCo-L1	$0.29 \pm 0.01^{**}$	0.27 ± 0.01	0.00 ± 0.00	0.58 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	$0.29 \pm 0.01^{**}$	0.28 ± 0.01	0.00 ± 0.00	0.58 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	$0.30 \pm 0.01^{**}$	0.28 ± 0.01	0.00 ± 0.00	0.58 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	$0.71 \pm 0.04^*$	0.57 ± 0.00	0.00 ± 0.00	1.38 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	$0.71 \pm 0.04^*$	0.57 ± 0.00	0.00 ± 0.00	1.38 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	0.77 ± 0.02	0.28 ± 0.01	0.00 ± 0.00	0.58 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	REVISE	$0.01 \pm 0.00^{**}$	0.43 ± 0.00	0.00 ± 0.00	1.26 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.78 ± 0.02	$0.25 \pm 0.00^{**}$	0.00 ± 0.00	0.58 ± 0.01	$0.38 \pm 0.02^{**}$	1.00 ± 0.00
	Wachter	0.77 ± 0.02	0.28 ± 0.01	0.00 ± 0.00	0.58 ± 0.02	0.00 ± 0.00	1.00 ± 0.00

Table 17: All results for Moons dataset: sample averages +/- one standard deviation over all valid counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	$0.29 \pm 0.02^{**}$	0.42 ± 0.04	0.14 ± 0.03	1.02 ± 0.03	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	$0.29 \pm 0.02^{**}$	0.42 ± 0.04	0.15 ± 0.03	1.02 ± 0.03	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	$0.30 \pm 0.02^{**}$	0.42 ± 0.04	0.13 ± 0.03	1.03 ± 0.03	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	$0.55 \pm 0.06^{**}$	$0.15 \pm 0.01^{**}$	$0.01 \pm 0.01^{**}$	1.28 ± 0.04	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	$0.57 \pm 0.05^{**}$	$0.15 \pm 0.01^{**}$	$0.01 \pm 0.01^{**}$	1.28 ± 0.04	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	0.69 ± 0.04	0.42 ± 0.04	0.13 ± 0.03	1.03 ± 0.03	0.00 ± 0.00	1.00 ± 0.00
	REVISE	$0.54 \pm 0.05^{**}$	$0.27 \pm 0.01^{**}$	0.14 ± 0.05	1.42 ± 0.13	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.72 ± 0.05	0.46 ± 0.04	0.19 ± 0.02	1.12 ± 0.03	$0.06 \pm 0.02^{**}$	1.00 ± 0.00
	Wachter	0.69 ± 0.05	0.42 ± 0.04	0.14 ± 0.03	1.02 ± 0.03	0.00 ± 0.00	1.00 ± 0.00
JEM Ensemble	ECCCo-L1	$0.26 \pm 0.01^{**}$	0.18 ± 0.01	0.02 ± 0.01	1.04 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	$0.26 \pm 0.02^{**}$	0.18 ± 0.01	0.04 ± 0.01	1.03 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	$0.26 \pm 0.01^{**}$	0.18 ± 0.01	$0.00 \pm 0.00^*$	1.03 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	$0.47 \pm 0.03^{**}$	$0.13 \pm 0.00^{**}$	$0.00 \pm 0.00^{**}$	1.50 ± 0.04	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	$0.46 \pm 0.04^{**}$	$0.13 \pm 0.00^{**}$	$0.00 \pm 0.00^{**}$	1.50 ± 0.04	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	0.67 ± 0.02	0.18 ± 0.01	$0.00 \pm 0.00^*$	1.03 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.66 ± 0.03	0.27 ± 0.01	0.44 ± 0.05	1.37 ± 0.10	0.00 ± 0.00	1.00 ± 0.00
	Schut	$0.51 \pm 0.03^{**}$	0.24 ± 0.01	$0.00 \pm 0.00^{**}$	1.81 ± 0.03	$0.05 \pm 0.01^{**}$	1.00 ± 0.00
	Wachter	0.67 ± 0.03	0.18 ± 0.01	0.01 ± 0.01	1.03 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
MLP	ECCCo-L1	$0.45 \pm 0.02^{**}$	0.34 ± 0.02	0.30 ± 0.02	1.49 ± 0.07	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	$0.45 \pm 0.03^{**}$	0.35 ± 0.02	0.31 ± 0.02	1.48 ± 0.07	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	$0.46 \pm 0.02^{**}$	0.35 ± 0.02	0.30 ± 0.02	1.49 ± 0.07	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	$1.33 \pm 0.06^{**}$	0.59 ± 0.03	$0.00 \pm 0.00^{**}$	2.84 ± 0.09	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	$1.32 \pm 0.09^*$	0.59 ± 0.03	$0.00 \pm 0.00^{**}$	2.84 ± 0.09	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	1.50 ± 0.04	0.35 ± 0.02	0.30 ± 0.02	1.49 ± 0.07	0.00 ± 0.00	1.00 ± 0.00
	REVISE	1.50 ± 0.06	$0.26 \pm 0.01^{**}$	0.30 ± 0.03	$1.20 \pm 0.07^{**}$	0.00 ± 0.00	1.00 ± 0.00
	Schut	1.59 ± 0.06	0.38 ± 0.03	$0.00 \pm 0.00^{**}$	$0.78 \pm 0.02^{**}$	$0.21 \pm 0.03^{**}$	1.00 ± 0.00
	Wachter	1.50 ± 0.04	0.35 ± 0.02	0.30 ± 0.02	1.48 ± 0.07	0.00 ± 0.00	1.00 ± 0.00
MLP Ensemble	ECCCo-L1	$0.44 \pm 0.02^{**}$	0.28 ± 0.02	0.22 ± 0.02	1.43 ± 0.08	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	$0.44 \pm 0.02^{**}$	0.28 ± 0.02	0.24 ± 0.03	1.43 ± 0.08	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	$0.46 \pm 0.02^{**}$	0.28 ± 0.02	$0.22 \pm 0.02^*$	1.42 ± 0.08	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	$1.24 \pm 0.07^{**}$	0.56 ± 0.02	$0.00 \pm 0.00^{**}$	2.81 ± 0.08	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	$1.23 \pm 0.09^{**}$	0.56 ± 0.02	$0.00 \pm 0.00^{**}$	2.81 ± 0.08	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	1.50 ± 0.03	0.28 ± 0.02	$0.22 \pm 0.02^*$	1.42 ± 0.08	0.00 ± 0.00	1.00 ± 0.00
	REVISE	1.59 ± 0.06	$0.25 \pm 0.01^{**}$	0.32 ± 0.03	$1.23 \pm 0.07^{**}$	0.00 ± 0.00	1.00 ± 0.00
	Schut	1.61 ± 0.06	0.28 ± 0.01	$0.03 \pm 0.01^{**}$	$0.75 \pm 0.02^{**}$	$0.26 \pm 0.03^{**}$	1.00 ± 0.00
	Wachter	1.51 ± 0.04	0.29 ± 0.02	0.24 ± 0.02	1.42 ± 0.08	0.00 ± 0.00	1.00 ± 0.00

Table 18: All results for California Housing dataset: sample averages +/- one standard deviation over all valid counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	0.98 ± 0.05**	0.99 ± 0.09	0.25 ± 0.02	1.20 ± 0.07	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.97 ± 0.05**	0.99 ± 0.09	0.26 ± 0.02	1.15 ± 0.06	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	1.00 ± 0.05**	1.02 ± 0.09	0.25 ± 0.01	1.10 ± 0.07	0.01 ± 0.00	1.00 ± 0.00
	ECCCo	1.93 ± 0.13*	0.77 ± 0.07**	0.18 ± 0.00**	2.05 ± 0.16	0.01 ± 0.00	1.00 ± 0.00
	ECCCo+	1.83 ± 0.12**	0.69 ± 0.04**	0.19 ± 0.00**	3.27 ± 0.18	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	1.95 ± 0.15*	0.77 ± 0.07**	0.18 ± 0.00**	2.03 ± 0.16	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	2.11 ± 0.11	1.02 ± 0.09	0.25 ± 0.01	1.10 ± 0.07	0.01 ± 0.00	1.00 ± 0.00
	REVISE	1.87 ± 0.16*	0.58 ± 0.04**	0.31 ± 0.06	5.50 ± 0.39	0.00 ± 0.00	1.00 ± 0.00
	Schut	2.03 ± 0.12	1.10 ± 0.08	0.21 ± 0.01**	2.81 ± 0.08	0.29 ± 0.01**	1.00 ± 0.00
	Wachter	2.12 ± 0.11	1.02 ± 0.09	0.26 ± 0.02	1.05 ± 0.07	0.01 ± 0.00	1.00 ± 0.00
JEM Ensemble	ECCCo-L1	0.89 ± 0.04**	0.98 ± 0.08	0.13 ± 0.00**	1.37 ± 0.07	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.89 ± 0.05**	0.98 ± 0.08	0.14 ± 0.00*	1.33 ± 0.07	0.01 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.92 ± 0.05**	0.99 ± 0.08	0.14 ± 0.00*	1.27 ± 0.07	0.01 ± 0.00	1.00 ± 0.00
	ECCCo	1.40 ± 0.08**	0.69 ± 0.05**	0.11 ± 0.00**	2.27 ± 0.10	0.00 ± 0.00	1.00 ± 0.00
	ECCCo+	1.28 ± 0.08**	0.60 ± 0.04**	0.11 ± 0.00**	3.39 ± 0.14	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	1.39 ± 0.08**	0.69 ± 0.05**	0.11 ± 0.00**	2.23 ± 0.10	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	1.70 ± 0.09	0.99 ± 0.08	0.14 ± 0.00*	1.27 ± 0.07	0.01 ± 0.00	1.00 ± 0.00
	REVISE	1.39 ± 0.15**	0.59 ± 0.04**	0.25 ± 0.07	5.37 ± 0.52	0.00 ± 0.00	1.00 ± 0.00
	Schut	1.59 ± 0.10*	1.10 ± 0.06	0.09 ± 0.00**	3.07 ± 0.10	0.31 ± 0.02**	1.00 ± 0.00
	Wachter	1.71 ± 0.09	0.99 ± 0.08	0.14 ± 0.00	1.23 ± 0.07	0.01 ± 0.00	1.00 ± 0.00
MLP	ECCCo-L1	1.22 ± 0.04**	1.18 ± 0.09	0.17 ± 0.01*	1.37 ± 0.09	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	1.23 ± 0.04**	1.18 ± 0.09	0.17 ± 0.01	1.34 ± 0.10	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	1.25 ± 0.04**	1.19 ± 0.09	0.17 ± 0.01	1.26 ± 0.09	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	3.81 ± 0.10*	2.31 ± 0.09	0.14 ± 0.01**	6.20 ± 0.14	0.00 ± 0.00	1.00 ± 0.00
	ECCCo+	3.93 ± 0.10	1.30 ± 0.08	0.16 ± 0.01*	4.65 ± 0.16	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	3.80 ± 0.11*	2.31 ± 0.09	0.14 ± 0.01**	6.17 ± 0.14	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	3.98 ± 0.08	1.19 ± 0.09	0.17 ± 0.01	1.26 ± 0.09	0.00 ± 0.00	1.00 ± 0.00
	REVISE	3.88 ± 0.08*	0.58 ± 0.03**	0.20 ± 0.04	5.38 ± 0.36	0.00 ± 0.00	1.00 ± 0.00
	Schut	3.99 ± 0.08	1.15 ± 0.09	0.15 ± 0.01**	1.58 ± 0.09	0.49 ± 0.02**	1.00 ± 0.00
	Wachter	3.98 ± 0.07	1.19 ± 0.09	0.17 ± 0.01	1.23 ± 0.09	0.00 ± 0.00	1.00 ± 0.00
MLP Ensemble	ECCCo-L1	1.22 ± 0.04**	1.13 ± 0.12	0.13 ± 0.01**	1.20 ± 0.07	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	1.22 ± 0.03**	1.13 ± 0.12	0.15 ± 0.01	1.17 ± 0.07	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	1.24 ± 0.04**	1.12 ± 0.12	0.14 ± 0.01**	1.07 ± 0.06	0.01 ± 0.00	1.00 ± 0.00
	ECCCo	3.69 ± 0.08**	1.94 ± 0.13	0.09 ± 0.01**	4.76 ± 0.13	0.01 ± 0.00	1.00 ± 0.00
	ECCCo+	3.88 ± 0.07**	1.20 ± 0.09	0.15 ± 0.02	4.23 ± 0.14	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	3.70 ± 0.08**	1.94 ± 0.13	0.10 ± 0.01**	4.76 ± 0.13	0.01 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	4.03 ± 0.07	1.12 ± 0.12	0.14 ± 0.01**	1.07 ± 0.06	0.01 ± 0.00	1.00 ± 0.00
	REVISE	3.96 ± 0.07*	0.58 ± 0.03**	0.17 ± 0.03	5.45 ± 0.37	0.00 ± 0.00	1.00 ± 0.00
	Schut	4.00 ± 0.06	1.15 ± 0.12	0.10 ± 0.01**	1.55 ± 0.08	0.58 ± 0.02**	1.00 ± 0.00
	Wachter	4.04 ± 0.07	1.13 ± 0.12	0.16 ± 0.01	1.04 ± 0.06	0.01 ± 0.00	1.00 ± 0.00

Table 19: All results for GMSC dataset: sample averages +/- one standard deviation over all valid counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	0.92 ± 0.04**	0.94 ± 0.09	0.30 ± 0.01**	1.02 ± 0.04	0.13 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.93 ± 0.04**	0.93 ± 0.09	0.31 ± 0.01*	0.94 ± 0.04	0.14 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.95 ± 0.04**	0.93 ± 0.10	0.31 ± 0.01*	0.88 ± 0.03	0.14 ± 0.01	1.00 ± 0.00
	ECCCo	1.78 ± 0.10	0.86 ± 0.09	0.35 ± 0.01	1.13 ± 0.05	0.13 ± 0.01	1.00 ± 0.00
	ECCCo+	1.65 ± 0.10**	0.73 ± 0.05**	0.35 ± 0.01	2.58 ± 0.11	0.09 ± 0.01	1.00 ± 0.00
	ECCCo (no CP)	1.79 ± 0.10	0.85 ± 0.09	0.36 ± 0.01	1.06 ± 0.05	0.13 ± 0.01	1.00 ± 0.00
	ECCCo (no EBM)	1.84 ± 0.10	0.93 ± 0.10	0.31 ± 0.01*	0.88 ± 0.03	0.14 ± 0.01	1.00 ± 0.00
	REVISE	1.66 ± 0.13*	0.63 ± 0.04**	0.34 ± 0.08	4.42 ± 0.30	0.03 ± 0.01	1.00 ± 0.00
	Schut	1.87 ± 0.09	1.19 ± 0.07	0.26 ± 0.01**	1.90 ± 0.05	0.46 ± 0.01**	1.00 ± 0.00
	Wachter	1.85 ± 0.11	0.91 ± 0.10	0.33 ± 0.01	0.80 ± 0.03	0.14 ± 0.01	1.00 ± 0.00
JEM Ensemble	ECCCo-L1	0.85 ± 0.04**	0.98 ± 0.10	0.31 ± 0.01**	1.10 ± 0.04	0.14 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.86 ± 0.05**	0.97 ± 0.10	0.32 ± 0.01*	1.03 ± 0.04	0.14 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.87 ± 0.04**	0.97 ± 0.10	0.32 ± 0.01**	0.98 ± 0.03	0.17 ± 0.01	1.00 ± 0.00
	ECCCo	1.20 ± 0.06*	0.78 ± 0.07**	0.38 ± 0.01	1.10 ± 0.06	0.13 ± 0.01	1.00 ± 0.00
	ECCCo+	1.01 ± 0.07**	0.70 ± 0.07**	0.37 ± 0.01	2.63 ± 0.16	0.10 ± 0.01	1.00 ± 0.00
	ECCCo (no CP)	1.21 ± 0.07*	0.77 ± 0.07**	0.39 ± 0.01	1.03 ± 0.06	0.13 ± 0.01	1.00 ± 0.00
	ECCCo (no EBM)	1.31 ± 0.07	0.97 ± 0.10	0.32 ± 0.01**	0.98 ± 0.03	0.17 ± 0.01	1.00 ± 0.00
	REVISE	1.01 ± 0.07**	0.63 ± 0.04**	0.33 ± 0.07	4.35 ± 0.36	0.04 ± 0.01	1.00 ± 0.00
	Schut	1.34 ± 0.07	1.21 ± 0.10	0.26 ± 0.01**	2.07 ± 0.05	0.43 ± 0.01**	1.00 ± 0.00
	Wachter	1.31 ± 0.08	0.95 ± 0.10	0.33 ± 0.01	0.90 ± 0.03	0.16 ± 0.01	1.00 ± 0.00
MLP	ECCCo-L1	1.20 ± 0.03**	0.98 ± 0.08	0.29 ± 0.01*	1.24 ± 0.06	0.15 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no CP)	1.20 ± 0.03**	0.97 ± 0.08	0.30 ± 0.01	1.18 ± 0.06	0.15 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no EBM)	1.21 ± 0.04**	0.95 ± 0.08	0.29 ± 0.01	1.09 ± 0.05	0.14 ± 0.01	1.00 ± 0.00
	ECCCo	3.80 ± 0.07	1.63 ± 0.07	0.24 ± 0.01**	3.13 ± 0.13	0.14 ± 0.01	1.00 ± 0.00
	ECCCo+	3.74 ± 0.08	1.10 ± 0.05	0.27 ± 0.01**	3.77 ± 0.11	0.10 ± 0.01	1.00 ± 0.00
	ECCCo (no CP)	3.81 ± 0.08	1.62 ± 0.08	0.24 ± 0.01**	3.10 ± 0.13	0.14 ± 0.01	1.00 ± 0.00
	ECCCo (no EBM)	3.79 ± 0.07	0.95 ± 0.08	0.29 ± 0.01	1.09 ± 0.05	0.14 ± 0.01	1.00 ± 0.00
	REVISE	3.81 ± 0.08	0.63 ± 0.03**	0.32 ± 0.05	4.34 ± 0.34	0.04 ± 0.01	1.00 ± 0.00
	Schut	3.83 ± 0.07	1.07 ± 0.07	0.30 ± 0.02	1.51 ± 0.06	0.50 ± 0.02**	1.00 ± 0.00
	Wachter	3.82 ± 0.07	0.94 ± 0.08	0.30 ± 0.01	1.03 ± 0.05	0.15 ± 0.01	1.00 ± 0.00
MLP Ensemble	ECCCo-L1	1.21 ± 0.03**	1.01 ± 0.08	0.30 ± 0.01**	1.30 ± 0.04	0.14 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no CP)	1.21 ± 0.02**	0.99 ± 0.08	0.31 ± 0.01*	1.23 ± 0.05	0.14 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no EBM)	1.23 ± 0.03**	0.97 ± 0.08	0.31 ± 0.01*	1.11 ± 0.04	0.14 ± 0.01	1.00 ± 0.00
	ECCCo	3.84 ± 0.07**	2.13 ± 0.08	0.23 ± 0.01**	4.24 ± 0.08	0.13 ± 0.01	1.00 ± 0.00
	ECCCo+	3.79 ± 0.05**	1.81 ± 0.05	0.30 ± 0.01*	5.29 ± 0.11	0.10 ± 0.01	1.00 ± 0.00
	ECCCo (no CP)	3.85 ± 0.07**	2.13 ± 0.08	0.23 ± 0.01**	4.20 ± 0.08	0.13 ± 0.01	1.00 ± 0.00
	ECCCo (no EBM)	4.08 ± 0.06	0.97 ± 0.08	0.31 ± 0.01*	1.11 ± 0.04	0.14 ± 0.01	1.00 ± 0.00
	REVISE	4.09 ± 0.07	0.63 ± 0.02**	0.33 ± 0.06	4.45 ± 0.44	0.04 ± 0.01	1.00 ± 0.00
	Schut	4.04 ± 0.08	1.21 ± 0.08	0.30 ± 0.01*	1.68 ± 0.06	0.56 ± 0.02**	1.00 ± 0.00
	Wachter	4.10 ± 0.07	0.95 ± 0.08	0.32 ± 0.01	1.04 ± 0.04	0.14 ± 0.01	1.00 ± 0.00

Table 20: All results for German Credit dataset: sample averages +/- one standard deviation over all valid counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	1.69 ± 0.03**	5.16 ± 0.07	0.59 ± 0.00**	3.04 ± 0.20	0.29 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no CP)	1.69 ± 0.03**	5.16 ± 0.07	0.60 ± 0.00	2.96 ± 0.19	0.29 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no EBM)	1.70 ± 0.04**	5.18 ± 0.08	0.59 ± 0.00**	3.02 ± 0.19	0.30 ± 0.01	1.00 ± 0.00
	ECCCo	4.48 ± 0.09*	5.07 ± 0.07*	0.60 ± 0.01	3.63 ± 0.28	0.27 ± 0.01	1.00 ± 0.00
	ECCCo+	4.85 ± 0.10	3.76 ± 0.10**	0.55 ± 0.01**	12.77 ± 0.40	0.08 ± 0.01	1.00 ± 0.00
	ECCCo (no CP)	4.48 ± 0.10	5.07 ± 0.07*	0.62 ± 0.01	3.58 ± 0.28	0.27 ± 0.01	1.00 ± 0.00
	ECCCo (no EBM)	4.55 ± 0.09	5.18 ± 0.08	0.59 ± 0.00**	3.02 ± 0.19	0.30 ± 0.01	1.00 ± 0.00
	REVISE	4.73 ± 0.16	3.46 ± 0.07**	0.53 ± 0.01**	19.21 ± 0.73	0.03 ± 0.01	1.00 ± 0.00
	Schut	4.67 ± 0.10	5.14 ± 0.08	0.59 ± 0.01*	2.53 ± 0.09**	0.79 ± 0.01**	1.00 ± 0.00
	Wachter	4.57 ± 0.08	5.17 ± 0.08	0.60 ± 0.00	2.95 ± 0.19	0.30 ± 0.01	1.00 ± 0.00
JEM Ensemble	ECCCo-L1	1.61 ± 0.04**	5.07 ± 0.10	0.88 ± 0.01**	3.51 ± 0.19	0.28 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no CP)	1.60 ± 0.04**	5.08 ± 0.10	0.90 ± 0.01	3.44 ± 0.19	0.28 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no EBM)	1.62 ± 0.04**	5.09 ± 0.10	0.88 ± 0.01**	3.48 ± 0.19	0.29 ± 0.01	1.00 ± 0.00
	ECCCo	3.80 ± 0.12**	4.13 ± 0.06**	0.83 ± 0.01**	7.28 ± 0.30	0.17 ± 0.01	1.00 ± 0.00
	ECCCo+	5.80 ± 0.12	3.55 ± 0.09**	0.60 ± 0.01**	12.12 ± 0.44	0.07 ± 0.01	1.00 ± 0.00
	ECCCo (no CP)	3.77 ± 0.14**	4.13 ± 0.06**	0.84 ± 0.01**	7.26 ± 0.30	0.17 ± 0.01	1.00 ± 0.00
	ECCCo (no EBM)	4.24 ± 0.12	5.09 ± 0.10	0.88 ± 0.01**	3.48 ± 0.19	0.29 ± 0.01	1.00 ± 0.00
	REVISE	5.33 ± 0.16	3.45 ± 0.07**	0.52 ± 0.01**	17.92 ± 0.89	0.03 ± 0.01	1.00 ± 0.00
	Schut	4.59 ± 0.15	4.90 ± 0.13*	0.78 ± 0.02**	3.46 ± 0.10	0.76 ± 0.01**	1.00 ± 0.00
	Wachter	4.23 ± 0.12	5.10 ± 0.10	0.90 ± 0.01	3.42 ± 0.19	0.29 ± 0.01	1.00 ± 0.00
MLP	ECCCo-L1	1.77 ± 0.03**	4.61 ± 0.09	0.76 ± 0.00*	4.98 ± 0.21	0.23 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no CP)	1.77 ± 0.03**	4.61 ± 0.09	0.77 ± 0.00	4.93 ± 0.21	0.23 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no EBM)	1.77 ± 0.03**	4.63 ± 0.09	0.76 ± 0.00*	4.95 ± 0.21	0.23 ± 0.01	1.00 ± 0.00
	ECCCo	5.36 ± 0.07	4.63 ± 0.09	0.81 ± 0.00	4.83 ± 0.21	0.24 ± 0.01**	1.00 ± 0.00
	ECCCo+	6.51 ± 0.05	3.20 ± 0.04**	0.56 ± 0.00**	10.93 ± 0.48	0.04 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	5.35 ± 0.07	4.62 ± 0.09	0.82 ± 0.00	4.78 ± 0.21	0.23 ± 0.01*	1.00 ± 0.00
	ECCCo (no EBM)	5.35 ± 0.06	4.63 ± 0.09	0.76 ± 0.00*	4.95 ± 0.21	0.23 ± 0.01	1.00 ± 0.00
	REVISE	6.53 ± 0.05	3.48 ± 0.07**	0.52 ± 0.01**	16.89 ± 0.68	0.03 ± 0.01	1.00 ± 0.00
	Schut	6.10 ± 0.13	4.40 ± 0.10**	0.66 ± 0.02**	2.55 ± 0.10**	0.78 ± 0.01**	1.00 ± 0.00
	Wachter	5.35 ± 0.06	4.63 ± 0.09	0.77 ± 0.00	4.90 ± 0.21	0.23 ± 0.01	1.00 ± 0.00
MLP Ensemble	ECCCo-L1	1.65 ± 0.03**	4.12 ± 0.05	0.82 ± 0.01**	6.77 ± 0.30	0.24 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no CP)	1.65 ± 0.03**	4.12 ± 0.05	0.84 ± 0.01	6.67 ± 0.31	0.24 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no EBM)	1.66 ± 0.03**	4.12 ± 0.05	0.82 ± 0.01**	6.72 ± 0.30	0.25 ± 0.01	1.00 ± 0.00
	ECCCo	4.52 ± 0.06	4.13 ± 0.05	0.85 ± 0.01	6.73 ± 0.30	0.23 ± 0.01	1.00 ± 0.00
	ECCCo+	6.29 ± 0.07	3.17 ± 0.02**	0.57 ± 0.00**	9.64 ± 0.53	0.04 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	4.52 ± 0.08	4.13 ± 0.05	0.86 ± 0.01	6.67 ± 0.30	0.24 ± 0.01	1.00 ± 0.00
	ECCCo (no EBM)	4.52 ± 0.07	4.12 ± 0.05	0.82 ± 0.01**	6.72 ± 0.30	0.25 ± 0.01	1.00 ± 0.00
	REVISE	6.37 ± 0.06	3.47 ± 0.06**	0.52 ± 0.01**	15.78 ± 0.82	0.03 ± 0.01	1.00 ± 0.00
	Schut	6.24 ± 0.21	4.10 ± 0.08	0.59 ± 0.02**	2.90 ± 0.10**	0.81 ± 0.01**	1.00 ± 0.00
	Wachter	4.54 ± 0.07	4.12 ± 0.05	0.84 ± 0.01	6.63 ± 0.30	0.25 ± 0.01	1.00 ± 0.00

Table 21: All results for MNIST dataset: sample averages +/- one standard deviation over all valid counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	0.10 ± 0.00**	0.35 ± 0.00	4.59 ± 0.01**	23.75 ± 0.97	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.25 ± 0.00**	0.46 ± 0.00	4.50 ± 0.00**	160.43 ± 1.31	0.00 ± 0.00	1.00 ± 0.00
	ECCCo+	0.25 ± 0.00**	0.31 ± 0.00**	4.51 ± 0.00**	196.62 ± 3.62	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.00	0.29 ± 0.00**	3.90 ± 0.14**	240.16 ± 6.54	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.25 ± 0.00	0.32 ± 0.01**	4.61 ± 0.05	6.02 ± 0.60**	0.99 ± 0.00**	1.00 ± 0.00
	Wachter	0.25 ± 0.00	0.35 ± 0.00	4.62 ± 0.01	22.90 ± 0.99	0.00 ± 0.00	1.00 ± 0.00
JEM Ensemble	ECCCo-L1	0.10 ± 0.00**	0.34 ± 0.00	2.44 ± 0.01**	35.89 ± 1.47	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.25 ± 0.00**	0.45 ± 0.00	1.62 ± 0.01**	157.21 ± 1.17	0.00 ± 0.00	1.00 ± 0.00
	ECCCo+	0.24 ± 0.00**	0.30 ± 0.00**	1.58 ± 0.01**	193.39 ± 2.21	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.00**	0.29 ± 0.00**	1.78 ± 0.09**	234.28 ± 3.19	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.25 ± 0.00	0.31 ± 0.01**	2.46 ± 0.11	7.20 ± 1.04**	0.99 ± 0.00**	1.00 ± 0.00
	Wachter	0.25 ± 0.00	0.34 ± 0.00	2.53 ± 0.01	34.89 ± 1.45	0.00 ± 0.00	1.00 ± 0.00
LeNet-5	ECCCo-L1	0.10 ± 0.00**	0.34 ± 0.00	8.67 ± 0.04*	46.52 ± 1.16	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.25 ± 0.00**	0.39 ± 0.00	0.01 ± 0.03**	121.18 ± 1.95	0.00 ± 0.00**	1.00 ± 0.00
	ECCCo+	0.25 ± 0.00**	0.31 ± 0.00**	0.17 ± 0.10**	173.43 ± 1.78	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.00**	0.30 ± 0.00**	2.25 ± 0.20**	230.65 ± 3.73	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.25 ± 0.00	0.29 ± 0.02*	4.26 ± 3.87*	7.28 ± 1.04**	0.98 ± 0.00**	1.00 ± 0.00
	Wachter	0.25 ± 0.00	0.33 ± 0.00	8.74 ± 0.00	45.32 ± 1.14	0.00 ± 0.00	1.00 ± 0.00
MLP	ECCCo-L1	0.10 ± 0.00**	0.35 ± 0.00	1.08 ± 0.00**	41.48 ± 1.47	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.24 ± 0.00**	0.42 ± 0.00	0.86 ± 0.00**	131.37 ± 1.79	0.00 ± 0.00**	1.00 ± 0.00
	ECCCo+	0.25 ± 0.00*	0.31 ± 0.00**	0.86 ± 0.00**	175.89 ± 2.75	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.00	0.30 ± 0.00**	0.74 ± 0.04**	224.80 ± 4.65	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.25 ± 0.00	0.30 ± 0.01**	1.12 ± 0.09	7.88 ± 0.67**	0.99 ± 0.00**	1.00 ± 0.00
	Wachter	0.25 ± 0.00	0.34 ± 0.00	1.16 ± 0.00	39.80 ± 1.44	0.00 ± 0.00	1.00 ± 0.00
MLP Ensemble	ECCCo-L1	0.10 ± 0.00**	0.34 ± 0.00	0.61 ± 0.01**	45.21 ± 1.45	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.24 ± 0.00**	0.40 ± 0.00	0.33 ± 0.00**	116.67 ± 1.95	0.00 ± 0.00**	1.00 ± 0.00
	ECCCo+	0.25 ± 0.00**	0.30 ± 0.00**	0.34 ± 0.00**	174.29 ± 3.50	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.00	0.30 ± 0.00**	0.36 ± 0.03**	226.43 ± 4.59	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.25 ± 0.00	0.29 ± 0.02**	0.68 ± 0.21	7.83 ± 1.24**	0.99 ± 0.00**	1.00 ± 0.00
	Wachter	0.25 ± 0.00	0.34 ± 0.00	0.66 ± 0.01	43.42 ± 1.44	0.00 ± 0.00	1.00 ± 0.00

Table 22: All results for Fashion MNIST dataset: sample averages +/- one standard deviation over all valid counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	0.10 ± 0.00**	0.39 ± 0.01	4.95 ± 0.01**	29.93 ± 1.60	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.25 ± 0.00**	0.43 ± 0.00	4.57 ± 0.01**	78.99 ± 1.28	0.00 ± 0.00**	1.00 ± 0.00
	ECCCo+	0.25 ± 0.00**	0.37 ± 0.00**	4.79 ± 0.04**	175.11 ± 4.20	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.00*	0.30 ± 0.02**	3.82 ± 0.49**	208.22 ± 16.18	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.25 ± 0.00	0.34 ± 0.01**	4.78 ± 0.04**	8.29 ± 0.23**	0.99 ± 0.00**	1.00 ± 0.00
	Wachter	0.25 ± 0.00	0.39 ± 0.01	5.01 ± 0.02	28.33 ± 1.56	0.00 ± 0.00	1.00 ± 0.00
JEM Ensemble	ECCCo-L1	0.10 ± 0.00**	0.37 ± 0.01	2.98 ± 0.04	37.85 ± 3.04	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.25 ± 0.00**	0.42 ± 0.00	2.95 ± 0.03	109.67 ± 2.72	0.00 ± 0.00*	1.00 ± 0.00
	ECCCo+	0.25 ± 0.00**	0.35 ± 0.00**	2.94 ± 0.02*	183.89 ± 5.64	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.00*	0.31 ± 0.01**	2.04 ± 0.34**	241.12 ± 12.81	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.25 ± 0.00*	0.30 ± 0.01**	2.90 ± 0.04*	8.44 ± 0.36**	0.99 ± 0.00**	1.00 ± 0.00
	Wachter	0.25 ± 0.00	0.37 ± 0.01	2.97 ± 0.05	35.32 ± 4.17	0.00 ± 0.00	1.00 ± 0.00
LeNet-5	ECCCo-L1	0.10 ± 0.00**	0.37 ± 0.01	3.39 ± 0.02*	23.45 ± 2.28	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.25 ± 0.00*	0.41 ± 0.00	2.80 ± 0.02**	92.19 ± 2.20	0.00 ± 0.00**	1.00 ± 0.00
	ECCCo+	0.25 ± 0.00**	0.36 ± 0.01	3.04 ± 0.04**	167.29 ± 5.17	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.00**	0.29 ± 0.02**	2.43 ± 0.28**	201.66 ± 17.81	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.25 ± 0.00	0.30 ± 0.02**	3.26 ± 0.14*	7.82 ± 0.36**	0.98 ± 0.00**	1.00 ± 0.00
	Wachter	0.25 ± 0.00	0.36 ± 0.01	3.42 ± 0.02	20.30 ± 1.82	0.00 ± 0.00	1.00 ± 0.00
MLP	ECCCo-L1	0.10 ± 0.00**	0.37 ± 0.01	3.56 ± 0.02**	34.94 ± 2.24	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.24 ± 0.00**	0.42 ± 0.00	2.91 ± 0.01**	112.25 ± 1.61	0.00 ± 0.00**	1.00 ± 0.00
	ECCCo+	0.25 ± 0.00**	0.36 ± 0.00	3.13 ± 0.02**	188.22 ± 6.05	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.00	0.30 ± 0.01**	2.40 ± 0.38**	224.88 ± 16.50	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.25 ± 0.00	0.28 ± 0.02**	3.30 ± 0.09**	8.27 ± 0.34**	0.99 ± 0.00**	1.00 ± 0.00
	Wachter	0.25 ± 0.00	0.36 ± 0.01	3.63 ± 0.03	31.09 ± 2.89	0.00 ± 0.00	1.00 ± 0.00
MLP Ensemble	ECCCo-L1	0.10 ± 0.00**	0.33 ± 0.02	2.71 ± 0.04	27.86 ± 1.82	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.24 ± 0.00**	0.40 ± 0.00	2.07 ± 0.02**	112.98 ± 3.11	0.00 ± 0.00**	1.00 ± 0.00
	ECCCo+	0.25 ± 0.00	0.35 ± 0.00	2.26 ± 0.02**	181.76 ± 4.91	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.00	0.28 ± 0.02**	1.81 ± 0.14**	207.62 ± 16.42	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.25 ± 0.00	0.27 ± 0.02*	2.45 ± 0.15*	8.08 ± 0.43**	0.99 ± 0.00**	1.00 ± 0.00
	Wachter	0.25 ± 0.00	0.32 ± 0.01	2.71 ± 0.04	23.84 ± 1.88	0.00 ± 0.00	1.00 ± 0.00

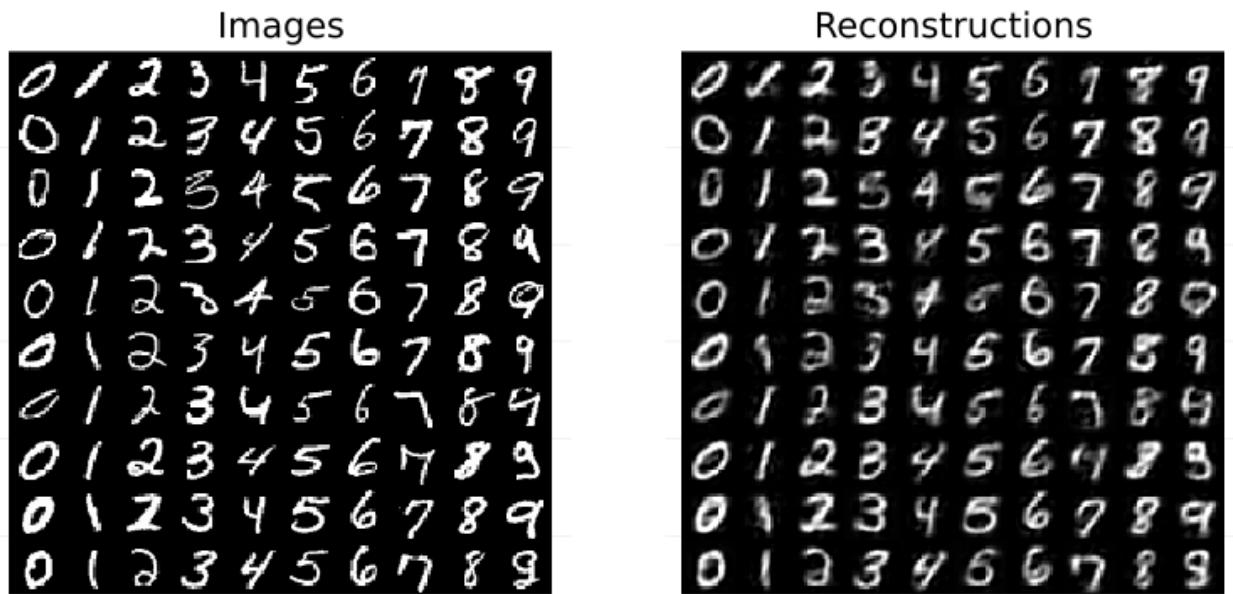


Figure 17: Randomly drawn MNIST images and their reconstructions generated by the VAE used by REVISE.