

ECCCos from the Black Box: Faithful 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 (ECCCos)** that are only as plausible as the model permits. Through extensive empirical studies, we demonstrate that ECCCos reconcile 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 ECCCos 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 help affected individuals through the means of algorithmic recourse. 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 $M_\theta : \mathcal{X} \mapsto \mathcal{Y}$ ini-

tially 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 (Mothilal, Sharma, and Tan 2020; Molnar 2020).

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 very distinct explanations can all lead to the same model prediction, especially when dealing with heavily parameterized models like deep neural networks, which are typically 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 and algorithmic recourses 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:

- 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).
- We introduce a novel algorithmic approach aimed at generating Energy-Constrained Conformal Counterfactuals (ECCCos) that faithfully explain model behaviour in Section 5.
- We provide extensive empirical evidence demonstrating

that ECCCos faithfully explain model behaviour and attain 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 ECCCos 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, proximity (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.; Guidotti).

The notion of plausibility is central to all of the desiderata. For example, Artelt et al. 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).

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 called *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*

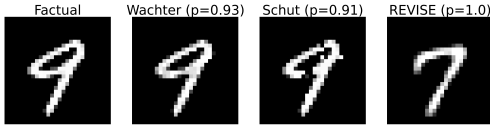


Figure 1: Counterfactuals for turning a 9 (nine) into a 7 (seven): original image (left); then from left to right the counterfactuals generated using *Wachter*, *Schut* and *REVISE*.

dataset (LeCun 1998): a Multi-Layer Perceptron (*MLP*) with above 90 percent test accuracy. 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 in Figure 1. Captions on top of the individual images indicate the generator along with the predicted probability that the image belongs to the target class. In all three cases that probability is above 90 percent and yet the counterfactuals look very different from each other.

Since *Wachter* is only concerned with proximity, 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 A New Notion of Faithfulness

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

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 $p_\theta(\mathbf{x}|\mathbf{y}^+)$. To this end, we draw on recent advances in energy-based modelling (EBM), a subdomain of machine learning that is concerned with generative or hybrid modelling (Grathwohl et al. 2020; Du and Mordatch). In particular, note that if we fix \mathbf{y} to our target value \mathbf{y}^+ , we can conditionally draw from $p_\theta(\mathbf{x}|\mathbf{y}^+)$ by randomly initializing \mathbf{x}_0 and then using Stochastic Gradient Langevin Dynamics (SGLD) as follows,

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

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). The term $\mathcal{E}(\mathbf{x}_j|\mathbf{y}^+)$ 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 $p_\theta(\mathbf{x}|\mathbf{y}^+)$ if $\epsilon \rightarrow 0$ and $J \rightarrow \infty$, the bias introduced for a small finite ϵ is negligible in practice (Murphy; Grathwohl et al. 2020). Appendix A provides additional implementation details for any tasks related to energy-based modelling.

Generating multiple samples using SGLD thus yields an empirical distribution $\hat{\mathbf{X}}_{\theta, \mathbf{y}^+}$ 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 and generate faithful model explanations.

4.2 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 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,

$$\text{impl}(\mathbf{x}', \mathbf{X}_{\mathbf{y}^+}) = \frac{1}{|\mathbf{X}_{\mathbf{y}^+}|} \sum_{\mathbf{x} \in \mathbf{X}_{\mathbf{y}^+}} \text{dist}(\mathbf{x}', \mathbf{x}) \quad (3)$$

where \mathbf{x}' denotes the counterfactual and $\mathbf{X}_{\mathbf{y}^+}$ is a subsample of the training data in the target class \mathbf{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 3 gives rise to a similar evaluation metric for unfaithfulness. We merely swap out the subsample of individuals in the target class for a subset $\hat{\mathbf{X}}_{\mathbf{y}^+}$ of the generated conditional samples:

$$\text{unfaith}(\mathbf{x}', \hat{\mathbf{X}}_{y^+}) = \frac{1}{n_E} \sum_{\mathbf{x} \in \hat{\mathbf{X}}_{y^+}} \text{dist}(\mathbf{x}', \mathbf{x}) \quad (4)$$

Specifically, we form this subset based on the n_E generated samples with the lowest energy.

5 Energy-Constrained Conformal Counterfactuals

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

We begin by stating our proposed objective function, which involves tailored loss and penalty functions that we will explain in the following. In particular, we extend Equation 1 as follows:

$$\begin{aligned} \mathbf{Z}' = \arg \min_{\mathbf{Z}' \in \mathcal{Z}^L} \{ & \text{yloss}(M_\theta(f(\mathbf{Z}')), \mathbf{y}^+) + \lambda_1 \text{dist}(f(\mathbf{Z}'), \mathbf{x}) \\ & + \lambda_2 \text{unfaith}(f(\mathbf{Z}'), \hat{\mathbf{X}}_{y^+}) + \lambda_3 \Omega(C_\theta(f(\mathbf{Z}'); \alpha)) \} \end{aligned} \quad (5)$$

The first penalty term involving λ_1 induces proximity like in Wachter, Mittelstadt, and Russell (2017). Our default choice for $\text{dist}(\cdot)$ is the L1 Norm due to its sparsity-inducing properties. The second penalty term involving λ_2 induces faithfulness by constraining the energy of the generated counterfactual where $\text{unfaith}(\cdot)$ corresponds to the metric defined in Equation 4. The third and final penalty term involving λ_3 introduces a new concept: it ensures that the generated counterfactual is associated with low predictive uncertainty. As mentioned in Section 2, Schut et al. (2021) have shown that plausible counterfactuals can be generated implicitly through predictive uncertainty minimization. Unfortunately, this relies on the assumption that the model itself can provide predictive uncertainty estimates, which may be too restrictive in practice.

To relax this assumption, we leverage recent advances in conformal prediction (CP), an approach to predictive uncertainty quantification that has recently gained popularity (Angelopoulos and Bates 2021; Manokhin). Crucially for our intended application, CP is model-agnostic and can be applied during inference without placing any restrictions on model training. Intuitively, CP works under the premise of turning heuristic notions of uncertainty into rigorous uncertainty 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 tend to be larger for inputs that do not conform with the training data and are characterized by high predictive uncertainty.

In order to generate counterfactuals that are associated with low predictive uncertainty, we use a smooth set size

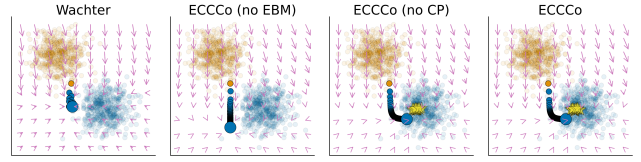


Figure 2: Gradient fields and counterfactual paths for different generators. The objective is to generate a counterfactual in the ‘blue’ class for a sample from the ‘orange’ class. Bright yellow stars indicate conditional samples generated through SGLD. The underlying classifier is a Joint Energy Model.

penalty introduced by Stutz et al. (2022) in the context of conformal training:

$$\Omega(C_\theta(\mathbf{x}; \alpha)) = \max \left(0, \sum_{\mathbf{y} \in \mathcal{Y}} C_{\theta, \mathbf{y}}(\mathbf{x}; \alpha) - \kappa \right) \quad (6)$$

Here, $\kappa \in \{0, 1\}$ is a hyper-parameter and $C_{\theta, \mathbf{y}}(\mathbf{x}; \alpha)$ can be interpreted as the probability of label \mathbf{y} being included in the prediction set. 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. Further details are provided in Appendix B.

Figure 2 illustrates how the different components in Equation 5 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 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 5. 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 attracted by the generated samples shown in bright yellow. 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

Algorithm 1 The *ECCCo* 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  $\mathbf{z}' \leftarrow f^{-1}(\mathbf{x})$  ▷ Map to counterfactual state space.
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  $\hat{\mathbf{X}}_{\mathbf{y}^+} \leftarrow \{\hat{\mathbf{x}}_{\theta, \mathbf{y}^+}\}_{n_B}$  ▷ Choose  $n_E$  lowest-energy samples.
4: Run SCP for  $M_\theta$  using  $\hat{\mathcal{D}}$  ▷ Calibrate model through split conformal prediction.
5: Initialize  $t \leftarrow 0$ 
6: while not converged or  $t < T$  do ▷ For convergence conditions see Appendix C.
7:    $\mathbf{z}' \leftarrow \mathbf{z}' - \eta \nabla_{\mathbf{z}'} \mathcal{L}(\mathbf{z}', \mathbf{y}^+, \hat{\mathbf{X}}_{\mathbf{y}^+}; \Lambda, \alpha)$  ▷ Take gradient step of size  $\eta$ .
8:    $t \leftarrow t + 1$ 
9: end while
10:  $\mathbf{x}' \leftarrow f(\mathbf{z}')$  ▷ Map back to feature space.
```

smooth set size penalty leads to somewhat faster convergence.

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 by passing the factual \mathbf{x} through a simple feature transformer f^{-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 $\mathcal{L}(\mathbf{z}', \mathbf{y}^+, \hat{\mathbf{X}}_{\mathbf{y}^+}; \Lambda, \alpha)$ denotes our loss function defined in Equation 5. 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 in Appendix C.

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 ECCCos more faithful than counterfactuals produced by state-of-the-art generators?*

Research Question 6.2 (Balancing Objectives). *Compared to state-of-the-art generators, how do ECCCos 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 with respect to plausibility; and, finally, *Wachter*, which serves as our baseline.

We use both synthetic and real-world datasets from different domains, all of which are publicly available and commonly used to train and benchmark classification algorithms. We synthetically generate a dataset containing two *Linearly Separable* Gaussian clusters ($n = 1000$), as well as the well-known *Circles* ($n = 1000$) and *Moons* ($n = 2500$) data. Since these data are generated by distributions of varying degrees of complexity, they allow us to assess how the generators and our proposed evaluation metrics handle this.

As for real-world data, we follow Schut et al. (2021) and use the *MNIST* (LeCun 1998) dataset containing images of handwritten digits such as the example shown above in Figure 1. From the social sciences domain, we include Give Me Some Credit (*GMSC*) (Kaggle 2011): a tabular dataset that has been studied extensively in the literature on algorithmic recourse (Pawelczyk et al. 2021). It consists of 11 numeric features that can be used to predict the binary outcome variable indicating whether retail borrowers experience financial distress.

For the predictive modelling tasks, we use simple neural networks (*MLP*) and Joint Energy Models (*JEM*). For the more complex real-world datasets we also use ensembling in each case. 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 ECCCos. To account for stochasticity, we generate multiple counterfactuals for each target class, generator, model and dataset. Specifically, we randomly sample n^- times from the subset of individuals for which the given model predicts the non-target class \mathbf{y}^- given the current target. We set $n^- = 25$ for all of our synthetic datasets, $n^- = 10$ for *GMSC* and $n^- = 5$ for *MNIST*. Full details concerning our parameter choices, training procedures and model performance can be found in Appendix D.

6.2 Results for Synthetic Data

Table 1 shows the key results for the synthetic datasets separated by model (first column) and generator (second column). The numerical columns show sample averages and standard deviations of our key evaluation metrics computed

across all counterfactuals. We have highlighted the best outcome for each model and metric in bold. To provide some sense of effect sizes, we have added asterisks to indicate that a given value is at least one (*) or two (**) standard deviations lower than the baseline (*Wachter*).

Starting with the high-level results for our *Linearly Separable* data, we find that *ECCECo* produces the most faithful counterfactuals for both black-box models. This is consistent with our design since *ECCECo* directly enforces faithfulness through regularization. Crucially though, *ECCECo* also produces the most plausible counterfactuals for both models. This dataset is so simple that even the *MLP* has learned plausible explanations of the input data. Zooming in on the granular details for the *Linearly Separable* data, the results for *ECCECo* (no CP) and *ECCECo* (no EBM) indicate that the positive results are dominated by the effect of quantifying and leveraging the model’s generative property (EBM). Conformal prediction alone only leads to marginally improved faithfulness and plausibility.

The findings for the *Moons* dataset are broadly in line with the findings so far: for the *JEM*, *ECCECo* yields substantially more faithful and plausible counterfactuals than all other generators. For the *MLP*, faithfulness is maintained but counterfactuals are not plausible. This high-level pattern is broadly consistent with other more complex datasets and supportive of our narrative, so it is worth highlighting: *ECCECos* consistently achieve high faithfulness, which—subject to the quality of the model itself—coincides with high plausibility. By comparison, *REVISE* yields the most plausible counterfactuals for the *MLP*, but it does so at the cost of faithfulness. We also observe that the best results for *ECCECo* are achieved when using both penalties. Once again though, the generative component (EBM) has a stronger impact on the positive results for the *JEM*.

For the *Circles* data, it appears that *REVISE* performs well, but we note that it generates valid counterfactuals only half of the time (see Appendix E for a complete overview including additional common evaluation metrics). The underlying VAE with default parameters has not adequately learned the data-generating process. Of course, it is possible to improve generative performance through hyperparameter tuning but this example serves to illustrate that *REVISE* depends on the quality of its surrogate. Independent of the outcome for *REVISE*, however, the results do not seem to indicate that *ECCECo* substantially improves faithfulness and plausibility for the *Circles* data. We think this points to a limitation of our evaluation metrics rather than *ECCECo* itself: computing average distances fails to account for the ‘wraparound’ effect associated with circular data (Gill and Hangartner).

6.3 Results for Real-World Data

The results for our real-world datasets are shown in Table 2. Once again the findings indicate that the plausibility of *ECCECos* is positively correlated with the capacity of the black-box model to distinguish plausible from implausible inputs. The case is very clear for *MNIST*: *ECCECos* are consistently more faithful than the counterfactuals produced by our benchmark generators and their plausibility gradually

improves through ensembling and joint-energy modelling. Interestingly, faithfulness also gradually improves for *REVISE*. This indicates that as our models improve, their generative capacity approaches that of the surrogate VAE used by *REVISE*. The VAE still outperforms our classifiers in this regard, as evident from the fact that *ECCECo* never quite reaches the same level of plausibility as *REVISE*. With reference to Appendix E we note that the results for *Schut* need to be discounted as it rarely produces valid counterfactuals for *MNIST*. Relatedly, we find that *ECCECo* is the only generator that consistently achieves full validity. Finally, it is worth noting that *ECCECo* produces counterfactual images with the lowest average predictive uncertainty for all models.

For the tabular credit dataset (*GMSC*) it is inherently challenging to use deep neural networks in order to achieve good discriminative performance (Borisov et al. 2021; Grinsztajn, Oyallon, and Varoquaux 2022) and generative performance (Liu et al.), respectively. In order to achieve high plausibility, *ECCECo* effectively requires classifiers to achieve good performance for both tasks. Since this is a challenging task even for Joint Energy Models, it is not surprising to find that even though *ECCECo* once again achieves state-of-the-art faithfulness, it is outperformed by *REVISE* and *Schut* with respect to plausibility.

6.4 Key Takeways

To conclude this section, we summarize our findings with reference to the opening questions. The results clearly demonstrate that *ECCECo* consistently achieves state-of-the-art faithfulness, as it was designed to do (Research Question 6.1). A related important finding is that *ECCECo* yields highly plausible explanations provided that they faithfully describe model behaviour (Research Question 6.2). *ECCECo* achieves this result primarily by leveraging the model’s generative property.

7 Limitations

Even though we have taken considerable measures to study our proposed methodology carefully, limitations can still be identified. In particular, we have found that the performance of *ECCECo* is sensitive to hyperparameter choices. In order to achieve faithfulness, we generally had to penalise the distance from generated samples slightly more than the distance from factual values.

Conversely, we have not found that strongly penalising prediction set sizes had any discernable effect. Our results indicate that CP alone is often not sufficient to achieve faithfulness and plausibility, although we acknowledge that this needs to be investigated more thoroughly through future work.

While our approach is readily applicable to models with gradient access like deep neural networks, more work is needed to generalise it to other machine learning models such as decision trees. Relatedly, common challenges associated with energy-based modelling including sensitivity to scale, training instabilities and sensitivity to hyperparameters also apply to *ECCECo*.

Table 1: Results for synthetic datasets: sample averages \pm one standard deviation across 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	Linearly Separable		Moons		Circles	
		Unfaithfulness \downarrow	Implausibility \downarrow	Unfaithfulness \downarrow	Implausibility \downarrow	Unfaithfulness \downarrow	Implausibility \downarrow
JEM	ECCCo	0.03 \pm 0.06**	0.20 \pm 0.08**	0.31 \pm 0.30*	1.20 \pm 0.15**	0.52 \pm 0.36	1.22 \pm 0.46
	ECCCo (no CP)	0.03 \pm 0.06**	0.20 \pm 0.08**	0.37 \pm 0.30*	1.21 \pm 0.17**	0.54 \pm 0.39	1.21 \pm 0.46
	ECCCo (no EBM)	0.16 \pm 0.11	0.34 \pm 0.19	0.91 \pm 0.32	1.71 \pm 0.25	0.70 \pm 0.33	1.30 \pm 0.37
	REVISE	0.19 \pm 0.03	0.41 \pm 0.01**	0.78 \pm 0.23	1.57 \pm 0.26	0.48 \pm 0.16*	0.95 \pm 0.32*
	Schut	0.39 \pm 0.07	0.73 \pm 0.17	0.67 \pm 0.27	1.50 \pm 0.22*	0.54 \pm 0.43	1.28 \pm 0.53
	Wachter	0.18 \pm 0.10	0.44 \pm 0.17	0.80 \pm 0.27	1.78 \pm 0.24	0.68 \pm 0.34	1.33 \pm 0.32
MLP	ECCCo	0.29 \pm 0.05**	0.23 \pm 0.06**	0.80 \pm 0.62	1.69 \pm 0.40	0.65 \pm 0.53	1.17 \pm 0.41
	ECCCo (no CP)	0.29 \pm 0.05**	0.23 \pm 0.07**	0.79 \pm 0.62	1.68 \pm 0.42	0.49 \pm 0.35	1.19 \pm 0.44
	ECCCo (no EBM)	0.46 \pm 0.05	0.28 \pm 0.04**	1.34 \pm 0.47	1.68 \pm 0.47	0.84 \pm 0.51	1.23 \pm 0.31
	REVISE	0.56 \pm 0.05	0.41 \pm 0.01	1.45 \pm 0.44	1.64 \pm 0.31	0.58 \pm 0.52	0.95 \pm 0.32
	Schut	0.43 \pm 0.06*	0.47 \pm 0.36	1.45 \pm 0.55	1.73 \pm 0.48	0.58 \pm 0.37	1.23 \pm 0.43
	Wachter	0.51 \pm 0.04	0.40 \pm 0.08	1.32 \pm 0.41	1.69 \pm 0.32	0.83 \pm 0.50	1.24 \pm 0.29

Table 2: Results for real-world datasets: sample averages \pm one standard deviation across 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	MNIST		GMSC	
		Unfaithfulness \downarrow	Implausibility \downarrow	Unfaithfulness \downarrow	Implausibility \downarrow
JEM	ECCCo	19.28 \pm 5.01**	314.76 \pm 32.36*	79.16 \pm 11.67**	18.26 \pm 4.92**
	REVISE	188.70 \pm 26.18*	255.26 \pm 41.50**	186.40 \pm 28.06	5.34 \pm 2.38**
	Schut	211.62 \pm 27.13	290.56 \pm 40.66*	200.98 \pm 28.49	6.50 \pm 2.01**
	Wachter	222.90 \pm 26.56	361.88 \pm 39.74	214.08 \pm 45.35	61.04 \pm 2.58
JEM Ensemble	ECCCo	15.99 \pm 3.06**	294.72 \pm 30.75**	83.28 \pm 13.26**	17.21 \pm 4.46**
	REVISE	173.59 \pm 20.65**	246.32 \pm 37.46**	194.24 \pm 35.41	4.95 \pm 1.26**
	Schut	204.36 \pm 23.14	290.64 \pm 39.49*	208.45 \pm 34.60	6.12 \pm 1.91**
	Wachter	217.67 \pm 23.78	363.23 \pm 39.24	186.19 \pm 33.88	60.70 \pm 44.32
MLP	ECCCo	41.95 \pm 6.50**	591.58 \pm 36.24	75.93 \pm 14.27**	17.20 \pm 3.15**
	REVISE	365.82 \pm 15.35*	249.49 \pm 41.55**	196.75 \pm 41.25	4.84 \pm 0.60**
	Schut	379.66 \pm 17.16	290.07 \pm 42.65*	212.00 \pm 41.15	6.44 \pm 1.34**
	Wachter	386.05 \pm 16.60	361.83 \pm 42.18	218.34 \pm 53.26	45.84 \pm 39.39
MLP Ensemble	ECCCo	31.43 \pm 3.91**	490.88 \pm 27.19	73.86 \pm 14.63**	17.92 \pm 4.17**
	REVISE	337.74 \pm 11.89*	247.67 \pm 38.36**	207.21 \pm 43.20	5.78 \pm 2.10**
	Schut	354.80 \pm 13.05	285.79 \pm 41.33*	205.36 \pm 32.11	7.00 \pm 2.15**
	Wachter	360.79 \pm 14.39	357.73 \pm 42.55	213.71 \pm 54.17	73.09 \pm 64.50

8 Conclusion

This work leverages recent advances in energy-based modelling and conformal prediction in the context of Explainable Artificial Intelligence. 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 if and only if 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 baseline for future work on faithful counterfactual explanations.

9 Acknowledgments

Some of the members of TU Delft were partially funded by ICAI AI for Fintech Research, an ING — TU Delft collaboration.

References

- Altmeyer, P.; Angela, G.; Buszydluk, A.; Dobiczek, K.; van Deursen, A.; and Liem, C. 2023. Endogenous Macrodynamics in Algorithmic Recourse. In *First IEEE Conference on Secure and Trustworthy Machine Learning*.
- Angelopoulos, A. N.; and Bates, S. 2021. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification.
- Artelt, A.; Vaquet, V.; Velioglu, R.; Hinder, F.; Brinkrolf, J.; Schilling, M.; and Hammer, B. ??? Evaluating Robustness of Counterfactual Explanations. Technical report, arXiv. ArXiv:2103.02354 [cs] type: article.
- Borisov, V.; Leemann, T.; Seßler, K.; Haug, J.; Pawelczyk, M.; and Kasneci, G. 2021. Deep Neural Networks and Tabular Data: A Survey.
- 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. ??? Implicit Generation and Generalization in Energy-Based Models. Technical report, arXiv. ArXiv:1903.08689 [cs, stat] type: article.
- Gill, J.; and Hangartner, D. ??? Circular Data in Political Science and How to Handle It. 18(3): 316–336.
- 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.
- Grinsztajn, L.; Oyallon, E.; and Varoquaux, G. 2022. Why Do Tree-Based Models Still Outperform Deep Learning on Tabular Data?

- Guidotti, R. 2019. Counterfactual explanations and how to find them: literature review and benchmarking.
- 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.
- 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. 2015. Adam: A Method for Stochastic Optimization. Technical report, arXiv. ArXiv:1412.6980 [cs] type: article.
- 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.
- Liu, T.; Qian, Z.; Berrevoets, J.; and Schaar, M. v. d. 2020. GOGGLE: Generative Modelling for Tabular Data by Learning Relational Structure.
- 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. Technical report, arXiv. ArXiv:1912.03277 [cs, stat] type: article.
- Manokhin, V. 2020. Awesome Conformal Prediction.
- Molnar, C. 2020. *Interpretable Machine Learning*. Lulu.com.
- 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. 2012. *Probabilistic machine learning: Advanced topics*. MIT Press.
- 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.
- Spooner, T.; Dervovic, D.; Long, J.; Shepard, J.; Chen, J.; and Magazzeni, D. 2021. Counterfactual Explanations for Arbitrary Regression Models.
- 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 Recourse 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.
- Welling, M.; and Teh, Y. 2011. Bayesian Learning via Stochastic Gradient Langevin Dynamics.
- Wilson, A. G. 2020. The Case for Bayesian Deep Learning.

Table 3: EBM hyperparameter choices for our experiments.

	Dataset	SGLD Steps	Batch Size	λ
	Linearly Separable	30	50	0.10
	Moons	30	10	0.10
	Circles	20	100	0.01
	MNIST	25	10	0.01
	GMSC	30	10	0.10

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 E 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 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). These models are trained to optimize a hybrid objective that 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 $L_{\text{gen}}(\theta) = -\log p_{\theta}(\mathbf{x})$.

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 $\mathcal{E}(\mathbf{x}|\mathbf{y}) = \mu_{\theta}(\mathbf{x})[\mathbf{y}]$ 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:

$$\hat{\mathbf{x}}_{j+1} \leftarrow \hat{\mathbf{x}}_j - \frac{\epsilon}{2} \mathcal{E}(\hat{\mathbf{x}}_j|\mathbf{y}^+) + \sigma \mathbf{r}_j, \quad j = 1, \dots, J \quad (7)$$

Consistent with Grathwohl et al. (2020), we have specified $\epsilon = 2$ and $\sigma = 0.01$ as the default values for all of our experiments. 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; 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} : $L_{\text{gen}}(\theta) = \mu_{\theta}(\mathbf{x})[\mathbf{y}] - \mu_{\theta}(\hat{\mathbf{x}}_J)[\mathbf{y}]$. Our full training objective can be summarized as follows,

$$L(\theta) = L_{\text{clf}}(\theta) + L_{\text{gen}}(\theta) + \lambda L_{\text{reg}}(\theta) \quad (8)$$

where $L_{\text{reg}}(\theta)$ is a Ridge penalty (L2 norm) that regularises energy magnitudes for both observed and generated samples (Du and Mordatch). 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 3 and 4 show generated samples for our *MNIST* and *Moons* data, to provide a sense of their generative property.

JEM Ensemble

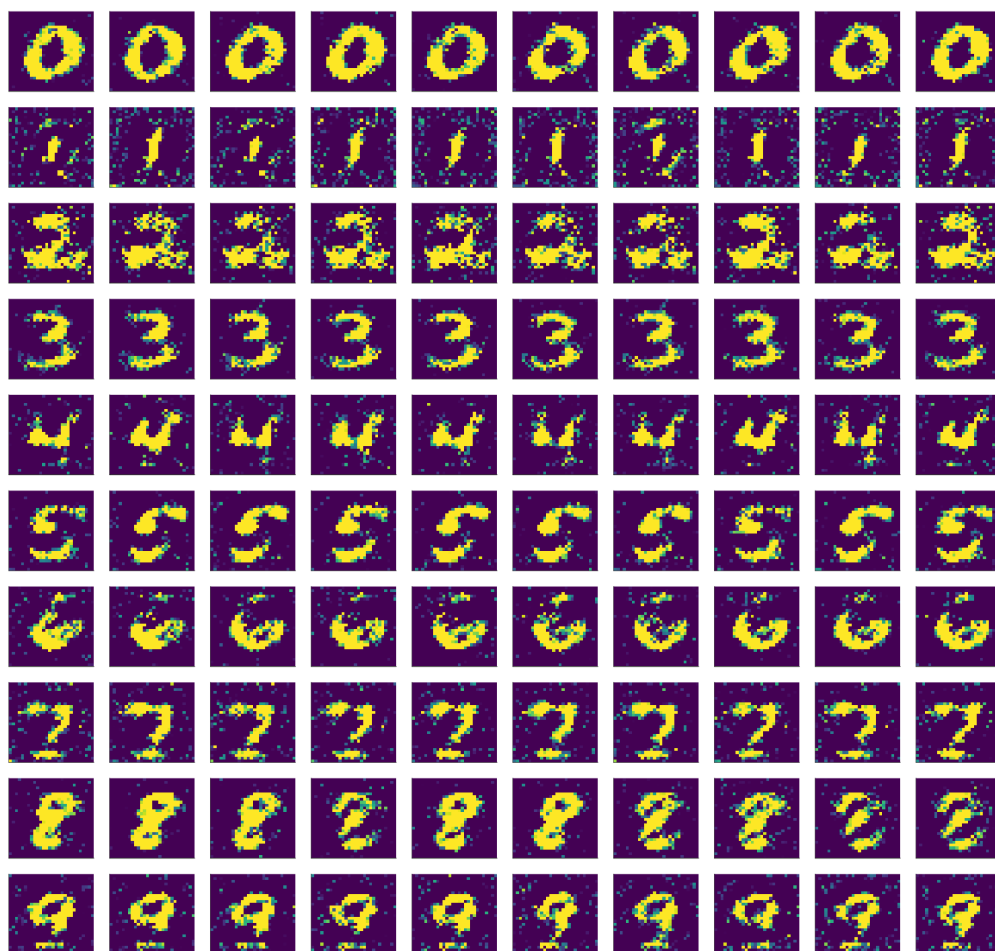


Figure 3: Conditionally generated *MNIST* images for our JEM Ensemble.

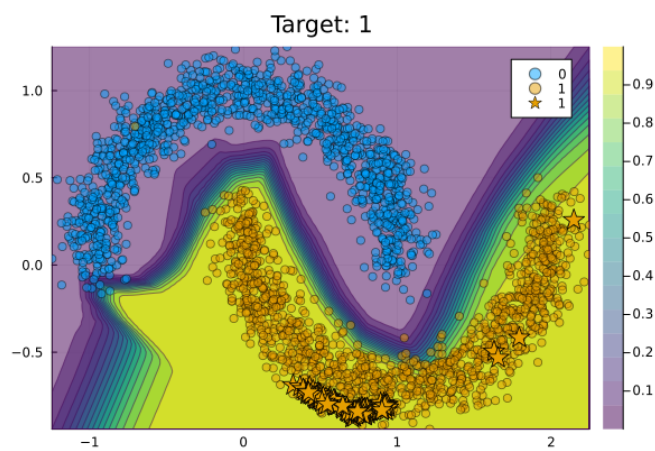


Figure 4: Conditionally generated samples (stars) for our *Moons* data using a JEM.

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,

$$C_\theta(\mathbf{x}_i; \alpha) = \{\mathbf{y} : s(\mathbf{x}_i, \mathbf{y}) \leq \hat{q}\} \quad (9)$$

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 9 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 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 6 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 5 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 ECCC0

In this section, we briefly discuss convergence conditions for CE and provide details concerning the actual implementation of our framework in *Julia*.

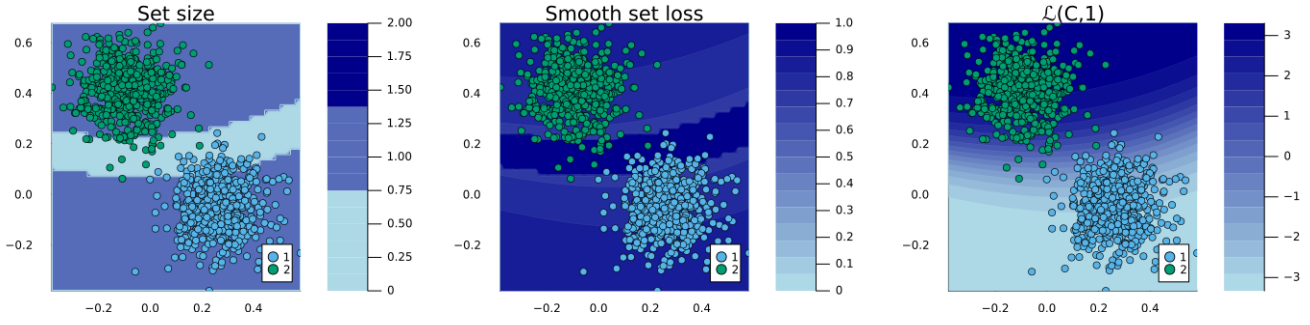


Figure 5: Prediction set size (left), smooth set size loss (centre) and configurable classification loss (right) for a JEM trained on our *Linearly Separable* data.

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
Moons	2500	32	3	relu	5	500	128
Circles	1000	32	3	swish	5	100	100
MNIST	10000	128	1	swish	5	100	128
GMSC	13370	128	2	swish	5	100	250

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 $p(y^+|x')$ 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 $p(y^+|x') > 0.5$ 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

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). 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 5). 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.

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 experiments could then be run locally on a personal machine. The longest runtimes we experienced for model training and counterfactual benchmarking were on the order of 8-12 hours (*MNIST* data). For the synthetic data, all experiments could be completed in less than an hour.

We have summarised our system information below:

Software:

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

Dataset	Model	Performance Metrics		
		Accuracy	Precision	F1-Score
Linearly Separable	JEM	0.99	0.99	0.99
	MLP	0.99	0.99	0.99
Moons	JEM	1.00	1.00	1.00
	MLP	1.00	1.00	1.00
Circles	JEM	0.98	0.98	0.98
	MLP	1.00	1.00	1.00
MNIST	JEM	0.83	0.84	0.83
	JEM Ensemble	0.90	0.90	0.89
	MLP	0.95	0.95	0.95
	MLP Ensemble	0.95	0.95	0.95
GMSC	JEM	0.73	0.75	0.73
	JEM Ensemble	0.73	0.75	0.73
	MLP	0.75	0.75	0.75
	MLP Ensemble	0.75	0.75	0.75

Table 6: Generator hyperparameters.

Dataset	η	λ_1	λ_2	λ_3
Linearly Separable	0.01	0.25	0.75	0.75
Moons	0.05	0.25	0.75	0.75
Circles	0.01	0.25	0.75	0.75
MNIST	0.10	0.10	0.25	0.25
GMSC	0.05	0.10	0.50	0.50

- 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

E Results

Figure 6 shows examples of counterfactuals for *MNIST* data where the underlying model is our *JEM Ensemble*. Original images are shown on the diagonal and the corresponding counterfactuals are plotted across rows.

Table 7 reports all of the evaluation metrics we have computed. Table 8 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. 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 6). Finally, ‘Validity’ is the percentage of valid counterfactuals.

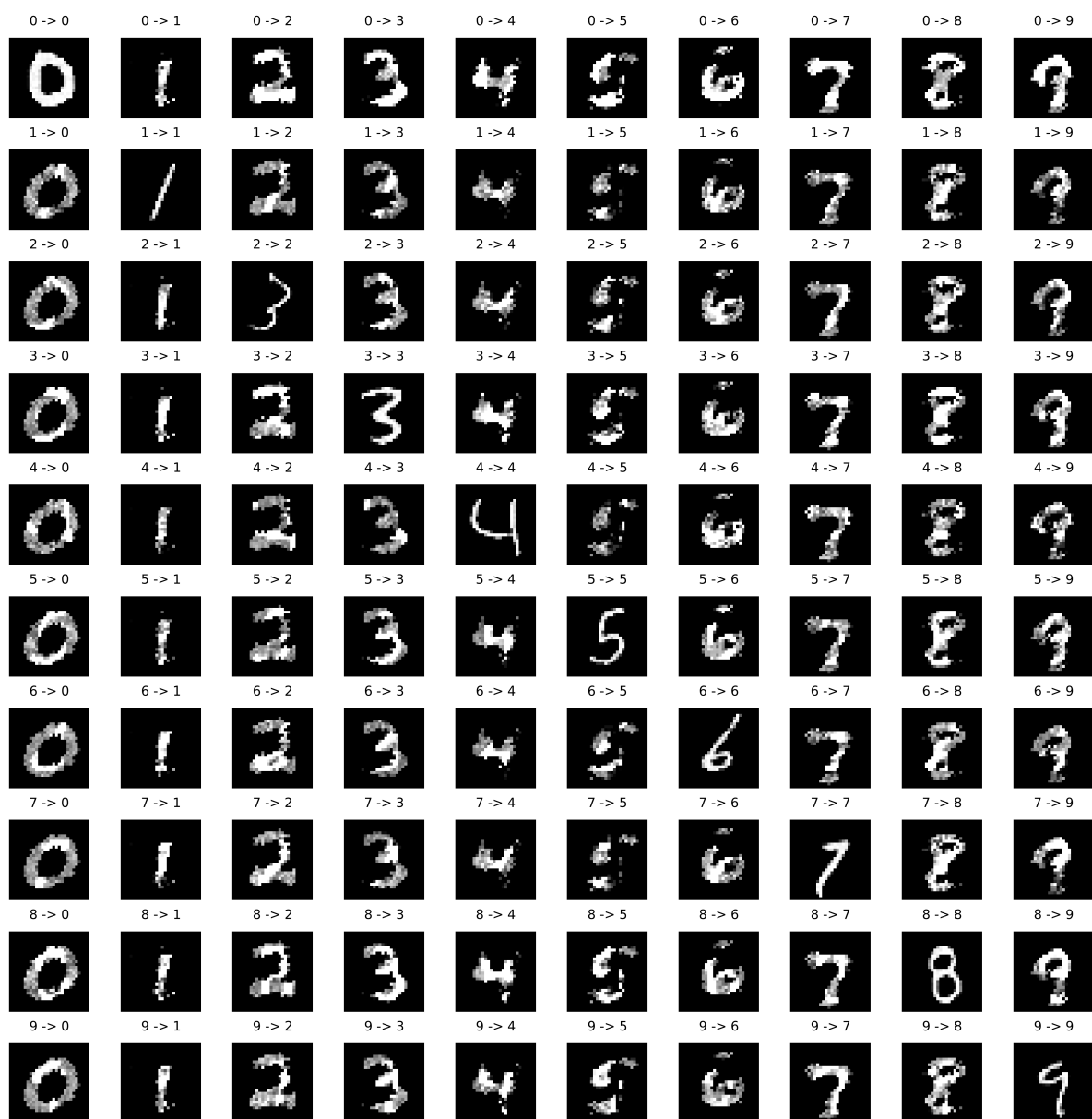


Figure 6: Counterfactuals for *MNIST* data and our *JEM Ensemble*. Original images are shown on the diagonal with the corresponding counterfactuals plotted across rows.

Table 7: All results for all datasets: 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	Data	Generator	Cost ↓	Unfaithfulness ↓	Implausibility ↓	Redundancy ↑	Uncertainty ↓	Validity ↑
Circles	JEM	ECCTCo	0.74 ± 0.21	0.52 ± 0.36	1.22 ± 0.46	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00**
		ECCTCo (no CP)	0.72 ± 0.21	0.54 ± 0.39	1.21 ± 0.46	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00**
		ECCTCo (no EBM)	0.52 ± 0.15	0.70 ± 0.33	1.30 ± 0.37	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00**
		REVISE	0.97 ± 0.34	0.48 ± 0.16*	0.95 ± 0.32*	0.00 ± 0.00	0.00 ± 0.00	0.50 ± 0.51
		Schut	1.06 ± 0.43	0.54 ± 0.43	1.28 ± 0.53	0.26 ± 0.25*	0.00 ± 0.00	1.00 ± 0.00**
		Wachter	0.44 ± 0.16	0.68 ± 0.34	1.33 ± 0.32	0.00 ± 0.00	0.00 ± 0.00	0.98 ± 0.14
	MLP	ECCTCo	0.67 ± 0.19	0.65 ± 0.53	1.17 ± 0.41	0.00 ± 0.00	0.09 ± 0.19**	1.00 ± 0.00
		ECCTCo (no CP)	0.71 ± 0.16	0.49 ± 0.35	1.19 ± 0.44	0.00 ± 0.00	0.05 ± 0.16**	1.00 ± 0.00
		ECCTCo (no EBM)	0.45 ± 0.11	0.84 ± 0.51	1.23 ± 0.31	0.00 ± 0.00	0.15 ± 0.23*	1.00 ± 0.00
		REVISE	0.96 ± 0.31	0.58 ± 0.52	0.95 ± 0.32	0.00 ± 0.00	0.00 ± 0.00**	0.50 ± 0.51
		Schut	0.57 ± 0.11	0.58 ± 0.37	1.23 ± 0.43	0.43 ± 0.18**	0.00 ± 0.00**	1.00 ± 0.00
		Wachter	0.40 ± 0.09	0.83 ± 0.50	1.24 ± 0.29	0.00 ± 0.00	0.53 ± 0.01	1.00 ± 0.00
	GMSC	ECCTCo	17.45 ± 2.92**	79.16 ± 11.67**	18.26 ± 4.92**	0.00 ± 0.00	0.10 ± 0.01	1.00 ± 0.00
		REVISE	3.43 ± 1.67**	186.40 ± 28.06	5.34 ± 2.38**	0.00 ± 0.00	0.51 ± 0.22	1.00 ± 0.00
		Schut	1.27 ± 0.33**	200.98 ± 28.49	6.50 ± 2.01**	0.77 ± 0.07**	0.07 ± 0.00	1.00 ± 0.00
		Wachter	57.71 ± 0.47	214.08 ± 45.35	61.04 ± 2.58	0.00 ± 0.00	0.07 ± 0.00	1.00 ± 0.00
	JEM Ensemble	ECCTCo	17.43 ± 3.04**	83.28 ± 13.26**	17.21 ± 4.46**	0.00 ± 0.00	0.16 ± 0.11	1.00 ± 0.00
		REVISE	2.94 ± 1.13**	194.24 ± 35.41	4.95 ± 1.26**	0.00 ± 0.00	0.51 ± 0.29	1.00 ± 0.00
		Schut	1.03 ± 0.20**	208.45 ± 34.60	6.12 ± 1.91**	0.85 ± 0.05**	0.09 ± 0.04	1.00 ± 0.00
		Wachter	56.79 ± 44.68	186.19 ± 33.88	60.70 ± 44.32	0.00 ± 0.00	0.07 ± 0.00	1.00 ± 0.00
	MLP	ECCTCo	17.05 ± 2.87**	75.93 ± 14.27**	17.20 ± 3.15**	0.00 ± 0.00	0.19 ± 0.08	1.00 ± 0.00**
		REVISE	2.93 ± 1.24**	196.75 ± 41.25	4.84 ± 0.60**	0.00 ± 0.00	0.38 ± 0.18	1.00 ± 0.00**
		Schut	1.49 ± 0.87**	212.00 ± 41.15	6.44 ± 1.34**	0.77 ± 0.13**	0.12 ± 0.01	1.00 ± 0.00**
		Wachter	42.97 ± 39.50	218.34 ± 53.26	45.84 ± 39.39	0.00 ± 0.00	0.06 ± 0.06	0.50 ± 0.51
	MLP Ensemble	ECCTCo	16.63 ± 2.62**	73.86 ± 14.63**	17.92 ± 4.17**	0.00 ± 0.00	0.23 ± 0.07	1.00 ± 0.00**
		REVISE	3.73 ± 2.36**	207.21 ± 43.20	5.78 ± 2.10**	0.00 ± 0.00	0.33 ± 0.19	1.00 ± 0.00**
		Schut	1.20 ± 0.47**	205.36 ± 32.11	7.00 ± 2.15**	0.79 ± 0.09**	0.12 ± 0.01	1.00 ± 0.00**
		Wachter	69.30 ± 66.00	213.71 ± 54.17	73.09 ± 64.50	0.00 ± 0.00	0.06 ± 0.06	0.50 ± 0.51
Linearly Separable	JEM	ECCTCo	0.75 ± 0.17	0.03 ± 0.06**	0.20 ± 0.08**	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
		ECCTCo (no CP)	0.75 ± 0.17	0.03 ± 0.06**	0.20 ± 0.08**	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
		ECCTCo (no EBM)	0.70 ± 0.16	0.16 ± 0.11	0.34 ± 0.19	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
		REVISE	0.41 ± 0.15	0.19 ± 0.03	0.41 ± 0.01**	0.00 ± 0.00	0.36 ± 0.36	0.50 ± 0.51
		Schut	1.15 ± 0.35	0.39 ± 0.07	0.73 ± 0.17	0.25 ± 0.25	0.00 ± 0.00	1.00 ± 0.00
		Wachter	0.50 ± 0.13	0.18 ± 0.10	0.44 ± 0.17	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
	MLP	ECCTCo	0.95 ± 0.16	0.29 ± 0.05**	0.23 ± 0.06**	0.00 ± 0.00	0.00 ± 0.00**	1.00 ± 0.00
		ECCTCo (no CP)	0.94 ± 0.16	0.29 ± 0.05**	0.23 ± 0.07**	0.00 ± 0.00	0.00 ± 0.00**	1.00 ± 0.00
		ECCTCo (no EBM)	0.60 ± 0.15	0.46 ± 0.05	0.28 ± 0.04**	0.00 ± 0.00	0.02 ± 0.10**	1.00 ± 0.00
		REVISE	0.42 ± 0.14	0.56 ± 0.05	0.41 ± 0.01	0.00 ± 0.00	0.47 ± 0.50	0.48 ± 0.50
		Schut	0.77 ± 0.17	0.43 ± 0.06*	0.47 ± 0.36	0.20 ± 0.25	0.00 ± 0.00**	1.00 ± 0.00
		Wachter	0.51 ± 0.15	0.51 ± 0.04	0.40 ± 0.08	0.00 ± 0.00	0.59 ± 0.02	1.00 ± 0.00
	MNIST	ECCTCo	334.61 ± 46.37	19.28 ± 5.01**	314.76 ± 32.36*	0.00 ± 0.00	4.43 ± 0.56	0.98 ± 0.12
		REVISE	170.68 ± 63.26	188.70 ± 26.18*	255.26 ± 41.50**	0.00 ± 0.00	4.39 ± 0.91	0.96 ± 0.20
		Schut	9.44 ± 1.60**	211.00 ± 27.21	286.61 ± 39.85*	0.99 ± 0.00**	1.08 ± 1.95*	0.24 ± 0.43
		Wachter	128.36 ± 14.95	222.90 ± 26.56	361.88 ± 39.74	0.00 ± 0.00	4.37 ± 0.98	0.95 ± 0.21
	JEM Ensemble	ECCTCo	342.64 ± 41.14	15.99 ± 3.06**	294.72 ± 30.75**	0.00 ± 0.00	2.07 ± 0.06**	1.00 ± 0.00**
		REVISE	170.21 ± 58.02	173.59 ± 20.65**	246.32 ± 37.46**	0.00 ± 0.00	2.56 ± 0.83	0.93 ± 0.26
		Schut	9.78 ± 1.02**	205.33 ± 24.07	287.39 ± 39.33*	0.99 ± 0.00**	0.32 ± 0.94**	0.11 ± 0.31
		Wachter	135.07 ± 16.79	217.67 ± 23.78	363.23 ± 39.24	0.00 ± 0.00	2.93 ± 0.77	0.94 ± 0.23
	MLP	ECCTCo	605.17 ± 44.78	41.95 ± 6.50**	591.58 ± 36.24	0.00 ± 0.00	0.57 ± 0.00**	1.00 ± 0.00**
		REVISE	146.61 ± 36.96	365.82 ± 15.35*	249.49 ± 41.55**	0.00 ± 0.00	0.62 ± 0.30	0.87 ± 0.34
		Schut	9.95 ± 0.37**	382.44 ± 17.81	285.98 ± 42.48*	0.99 ± 0.00**	0.05 ± 0.19**	0.06 ± 0.24
		Wachter	136.08 ± 16.09	386.05 ± 16.60	361.83 ± 42.18	0.00 ± 0.00	0.68 ± 0.36	0.84 ± 0.36
	MLP Ensemble	ECCTCo	525.87 ± 34.00	31.43 ± 3.91**	490.88 ± 27.19	0.00 ± 0.00	0.29 ± 0.00**	1.00 ± 0.00**
		REVISE	146.60 ± 35.64	337.74 ± 11.89*	247.67 ± 38.36**	0.00 ± 0.00	0.39 ± 0.22	0.85 ± 0.36
		Schut	9.98 ± 0.25**	359.54 ± 14.52	283.99 ± 41.08*	0.99 ± 0.00**	0.03 ± 0.14**	0.06 ± 0.24
		Wachter	137.53 ± 18.95	360.79 ± 14.39	357.73 ± 42.55	0.00 ± 0.00	0.47 ± 0.64	0.80 ± 0.40
Moons	JEM	ECCTCo	1.56 ± 0.44	0.31 ± 0.30*	1.20 ± 0.15**	0.00 ± 0.00	0.00 ± 0.00**	1.00 ± 0.00**
		ECCTCo (no CP)	1.56 ± 0.46	0.37 ± 0.30*	1.21 ± 0.17**	0.00 ± 0.00	0.00 ± 0.00**	1.00 ± 0.00**
		ECCTCo (no EBM)	0.80 ± 0.25	0.91 ± 0.32	1.71 ± 0.25	0.00 ± 0.00	0.00 ± 0.00**	1.00 ± 0.00**
		REVISE	1.04 ± 0.43	0.78 ± 0.23	1.57 ± 0.26	0.00 ± 0.00	0.00 ± 0.00**	1.00 ± 0.00**
		Schut	1.12 ± 0.31	0.67 ± 0.27	1.50 ± 0.22*	0.08 ± 0.19	0.00 ± 0.00**	0.98 ± 0.14
		Wachter	0.72 ± 0.24	0.80 ± 0.27	1.78 ± 0.24	0.00 ± 0.00	0.02 ± 0.10	0.98 ± 0.14
	MLP	ECCTCo	2.18 ± 1.05	0.80 ± 0.62	1.69 ± 0.40	0.00 ± 0.00	0.15 ± 0.24*	1.00 ± 0.00
		ECCTCo (no CP)	2.07 ± 1.15	0.79 ± 0.62	1.68 ± 0.42	0.00 ± 0.00	0.15 ± 0.24*	1.00 ± 0.00
		ECCTCo (no EBM)	1.25 ± 0.92	1.34 ± 0.47	1.68 ± 0.47	0.00 ± 0.00	0.43 ± 0.18	1.00 ± 0.00
		REVISE	0.79 ± 0.19*	1.45 ± 0.44	1.64 ± 0.31	0.00 ± 0.00	0.40 ± 0.22	1.00 ± 0.00
		Schut	0.73 ± 0.25*	1.45 ± 0.55	1.73 ± 0.48	0.31 ± 0.28*	0.00 ± 0.00**	0.90 ± 0.30
		Wachter	1.08 ± 0.83	1.32 ± 0.41	1.69 ± 0.32	0.00 ± 0.00	0.52 ± 0.08	1.00 ± 0.00

Table 8: All results for all datasets: 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	Data	Generator	Cost ↓	Unfaithfulness ↓	Implausibility ↓	Redundancy ↑	Uncertainty ↓	Validity ↑
Circles	JEM	ECCTCo	0.74 ± 0.21	0.52 ± 0.36	1.22 ± 0.46	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
		ECCTCo (no CP)	0.72 ± 0.21	0.54 ± 0.39	1.21 ± 0.46	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
		ECCTCo (no EBM)	0.52 ± 0.15	0.70 ± 0.33	1.30 ± 0.37	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
		REVISE	1.28 ± 0.14	0.33 ± 0.01**	0.64 ± 0.00**	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
		Schut	1.06 ± 0.43	0.54 ± 0.43	1.28 ± 0.53	0.26 ± 0.25*	0.00 ± 0.00	1.00 ± 0.00
		Wachter	0.45 ± 0.15	0.68 ± 0.34	1.33 ± 0.32	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
	MLP	ECCTCo	0.67 ± 0.19	0.65 ± 0.53	1.17 ± 0.41	0.00 ± 0.00	0.09 ± 0.19**	1.00 ± 0.00
		ECCTCo (no CP)	0.71 ± 0.16	0.49 ± 0.35	1.19 ± 0.44	0.00 ± 0.00	0.05 ± 0.16**	1.00 ± 0.00
		ECCTCo (no EBM)	0.45 ± 0.11	0.84 ± 0.51	1.23 ± 0.31	0.00 ± 0.00	0.15 ± 0.23*	1.00 ± 0.00
		REVISE	1.24 ± 0.15	0.06 ± 0.01**	0.64 ± 0.00**	0.00 ± 0.00	0.00 ± 0.00**	1.00 ± 0.00
		Schut	0.57 ± 0.11	0.58 ± 0.37	1.23 ± 0.43	0.43 ± 0.18**	0.00 ± 0.00**	1.00 ± 0.00
		Wachter	0.40 ± 0.09	0.83 ± 0.50	1.24 ± 0.29	0.00 ± 0.00	0.53 ± 0.01	1.00 ± 0.00
GMSC	JEM	ECCTCo	17.45 ± 2.92**	79.16 ± 11.67**	18.26 ± 4.92**	0.00 ± 0.00	0.10 ± 0.01	1.00 ± 0.00
		REVISE	3.43 ± 1.67**	186.40 ± 28.06	5.34 ± 2.38**	0.00 ± 0.00	0.51 ± 0.22	1.00 ± 0.00
		Schut	1.27 ± 0.33**	200.98 ± 28.49	6.50 ± 2.01**	0.77 ± 0.07**	0.07 ± 0.00	1.00 ± 0.00
		Wachter	57.71 ± 0.47	214.08 ± 45.35	61.04 ± 2.58	0.00 ± 0.00	0.07 ± 0.00	1.00 ± 0.00
	JEM Ensemble	ECCTCo	17.43 ± 3.04**	83.28 ± 13.26**	17.21 ± 4.46**	0.00 ± 0.00	0.16 ± 0.11	1.00 ± 0.00
		REVISE	2.94 ± 1.13**	194.24 ± 35.41	4.95 ± 1.26**	0.00 ± 0.00	0.51 ± 0.29	1.00 ± 0.00
		Schut	1.03 ± 0.20**	208.45 ± 34.60	6.12 ± 1.91**	0.85 ± 0.05**	0.09 ± 0.04	1.00 ± 0.00
		Wachter	56.79 ± 44.68	186.19 ± 33.88	60.70 ± 44.32	0.00 ± 0.00	0.07 ± 0.00	1.00 ± 0.00
	MLP	ECCTCo	17.05 ± 2.87	75.93 ± 14.27**	17.20 ± 3.15	0.00 ± 0.00	0.19 ± 0.08	1.00 ± 0.00
		REVISE	2.93 ± 1.24*	196.75 ± 41.25	4.84 ± 0.60**	0.00 ± 0.00	0.38 ± 0.18	1.00 ± 0.00
		Schut	1.49 ± 0.87**	212.00 ± 41.15	6.44 ± 1.34	0.77 ± 0.13**	0.12 ± 0.01	1.00 ± 0.00
		Wachter	4.48 ± 0.18	184.03 ± 48.16	7.49 ± 0.89	0.00 ± 0.00	0.12 ± 0.00	1.00 ± 0.00
	MLP Ensemble	ECCTCo	16.63 ± 2.62	73.86 ± 14.63**	17.92 ± 4.17	0.00 ± 0.00	0.23 ± 0.07	1.00 ± 0.00
		REVISE	3.73 ± 2.36	207.21 ± 43.20	5.78 ± 2.10**	0.00 ± 0.00	0.33 ± 0.19	1.00 ± 0.00
		Schut	1.20 ± 0.47**	205.36 ± 32.11	7.00 ± 2.15*	0.79 ± 0.09**	0.12 ± 0.01	1.00 ± 0.00
		Wachter	4.97 ± 0.47	177.20 ± 25.86	10.27 ± 3.21	0.00 ± 0.00	0.11 ± 0.00	1.00 ± 0.00
Linearly Separable	JEM	ECCTCo	0.75 ± 0.17	0.03 ± 0.06**	0.20 ± 0.08**	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
		ECCTCo (no CP)	0.75 ± 0.17	0.03 ± 0.06**	0.20 ± 0.08**	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
		ECCTCo (no EBM)	0.70 ± 0.16	0.16 ± 0.11	0.34 ± 0.19	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
		REVISE	0.41 ± 0.14	0.15 ± 0.00**	0.41 ± 0.01**	0.00 ± 0.00	0.72 ± 0.02	1.00 ± 0.00
		Schut	1.15 ± 0.35	0.39 ± 0.07	0.73 ± 0.17	0.25 ± 0.25	0.00 ± 0.00	1.00 ± 0.00
		Wachter	0.50 ± 0.13	0.18 ± 0.10	0.44 ± 0.17	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
	MLP	ECCTCo	0.95 ± 0.16	0.29 ± 0.05**	0.23 ± 0.06**	0.00 ± 0.00	0.00 ± 0.00**	1.00 ± 0.00
		ECCTCo (no CP)	0.94 ± 0.16	0.29 ± 0.05**	0.23 ± 0.07**	0.00 ± 0.00	0.00 ± 0.00**	1.00 ± 0.00
		ECCTCo (no EBM)	0.60 ± 0.15	0.46 ± 0.05	0.28 ± 0.04**	0.00 ± 0.00	0.02 ± 0.10**	1.00 ± 0.00
		REVISE	0.39 ± 0.15	0.52 ± 0.04	0.41 ± 0.01	0.00 ± 0.00	0.98 ± 0.00	1.00 ± 0.00
		Schut	0.77 ± 0.17	0.43 ± 0.06*	0.47 ± 0.36	0.20 ± 0.25	0.00 ± 0.00**	1.00 ± 0.00
		Wachter	0.51 ± 0.15	0.51 ± 0.04	0.40 ± 0.08	0.00 ± 0.00	0.59 ± 0.02	1.00 ± 0.00
MNIST	JEM	ECCTCo	334.98 ± 46.54	19.27 ± 5.02**	314.54 ± 32.54*	0.00 ± 0.00	4.50 ± 0.00**	1.00 ± 0.00
		REVISE	170.06 ± 62.45	188.54 ± 26.22*	254.32 ± 41.55**	0.00 ± 0.00	4.57 ± 0.14	1.00 ± 0.00
		Schut	7.63 ± 2.55**	199.70 ± 28.43	273.01 ± 39.60**	0.99 ± 0.00**	4.56 ± 0.13	1.00 ± 0.00
		Wachter	128.13 ± 14.81	222.81 ± 26.22	361.38 ± 39.55	0.00 ± 0.00	4.58 ± 0.16	1.00 ± 0.00
	JEM Ensemble	ECCTCo	342.64 ± 41.14	15.99 ± 3.06**	294.72 ± 30.75**	0.00 ± 0.00	2.07 ± 0.06**	1.00 ± 0.00
		REVISE	171.95 ± 58.81	173.05 ± 20.38**	246.20 ± 37.74**	0.00 ± 0.00	2.76 ± 0.45	1.00 ± 0.00
		Schut	7.96 ± 2.49**	186.91 ± 22.98*	264.68 ± 37.58**	0.99 ± 0.00**	3.02 ± 0.26	1.00 ± 0.00
		Wachter	134.98 ± 16.95	217.37 ± 23.93	362.91 ± 39.40	0.00 ± 0.00	3.10 ± 0.31	1.00 ± 0.00
	MLP	ECCTCo	605.17 ± 44.78	41.95 ± 6.50**	591.58 ± 36.24	0.00 ± 0.00	0.57 ± 0.00**	1.00 ± 0.00
		REVISE	146.76 ± 37.07	365.69 ± 14.90*	245.36 ± 39.69**	0.00 ± 0.00	0.72 ± 0.18	1.00 ± 0.00
		Schut	9.25 ± 1.31**	371.12 ± 19.99	245.11 ± 35.72**	0.99 ± 0.00**	0.75 ± 0.23	1.00 ± 0.00
		Wachter	135.08 ± 15.68	384.76 ± 16.52	359.21 ± 42.03	0.00 ± 0.00	0.81 ± 0.22	1.00 ± 0.00
	MLP Ensemble	ECCTCo	525.87 ± 34.00	31.43 ± 3.91**	490.88 ± 27.19	0.00 ± 0.00	0.29 ± 0.00**	1.00 ± 0.00
		REVISE	146.38 ± 35.18	337.21 ± 11.68*	244.84 ± 37.17**	0.00 ± 0.00	0.45 ± 0.16	1.00 ± 0.00
		Schut	9.75 ± 1.00**	344.60 ± 13.64*	252.53 ± 37.92**	0.99 ± 0.00**	0.55 ± 0.21	1.00 ± 0.00
		Wachter	134.48 ± 17.69	358.51 ± 13.18	352.63 ± 39.93	0.00 ± 0.00	0.58 ± 0.67	1.00 ± 0.00
Moons	JEM	ECCTCo	1.56 ± 0.44	0.31 ± 0.30*	1.20 ± 0.15**	0.00 ± 0.00	0.00 ± 0.00**	1.00 ± 0.00
		ECCTCo (no CP)	1.56 ± 0.46	0.37 ± 0.30*	1.21 ± 0.17**	0.00 ± 0.00	0.00 ± 0.00**	1.00 ± 0.00
		ECCTCo (no EBM)	0.80 ± 0.25	0.91 ± 0.32	1.71 ± 0.25	0.00 ± 0.00	0.00 ± 0.00**	1.00 ± 0.00
		REVISE	1.04 ± 0.43	0.78 ± 0.23	1.57 ± 0.26	0.00 ± 0.00	0.00 ± 0.00**	1.00 ± 0.00
		Schut	1.13 ± 0.29	0.66 ± 0.25	1.47 ± 0.10**	0.07 ± 0.18	0.00 ± 0.00**	1.00 ± 0.00
		Wachter	0.73 ± 0.24	0.78 ± 0.23	1.75 ± 0.19	0.00 ± 0.00	0.02 ± 0.11	1.00 ± 0.00
	MLP	ECCTCo	2.18 ± 1.05	0.80 ± 0.62	1.69 ± 0.40	0.00 ± 0.00	0.15 ± 0.24*	1.00 ± 0.00
		ECCTCo (no CP)	2.07 ± 1.15	0.79 ± 0.62	1.68 ± 0.42	0.00 ± 0.00	0.15 ± 0.24*	1.00 ± 0.00
		ECCTCo (no EBM)	1.25 ± 0.92	1.34 ± 0.47	1.68 ± 0.47	0.00 ± 0.00	0.43 ± 0.18	1.00 ± 0.00
		REVISE	0.79 ± 0.19*	1.45 ± 0.44	1.64 ± 0.31	0.00 ± 0.00	0.40 ± 0.22	1.00 ± 0.00
		Schut	0.78 ± 0.17*	1.39 ± 0.50	1.59 ± 0.26	0.28 ± 0.25*	0.00 ± 0.00**	1.00 ± 0.00
		Wachter	1.08 ± 0.83	1.32 ± 0.41	1.69 ± 0.32	0.00 ± 0.00	0.52 ± 0.08	1.00 ± 0.00