
ECCCOs from the Black Box: Faithful Explanations through Energy-Constrained Conformal Counterfactuals

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Counterfactual Explanations offer an intuitive and straightforward way to explain
2 black-box models and offer Algorithmic Recourse to individuals. To address the
3 need for plausible explanations, existing work has primarily relied on surrogate
4 models to learn how the input data is distributed. This effectively reallocates
5 the task of learning realistic explanations for the data from the model itself to
6 the surrogate. Consequently, the generated explanations may seem plausible to
7 humans but need not necessarily describe the behaviour of the black-box model
8 faithfully. We formalise this notion of faithfulness through the introduction of a tai-
9 lored evaluation metric and propose a novel algorithmic framework for generating
10 **Energy-Constrained Conformal Counterfactuals (ECCCOs)** that are only as plausi-
11 ble as the model permits. Through extensive empirical studies, we demonstrate that
12 ECCCOs reconcile the need for faithfulness and plausibility. In particular, we show
13 that for models with gradient access, it is possible to achieve state-of-the-art perfor-
14 mance without the need for surrogate models. To do so, our framework relies solely
15 on properties defining the black-box model itself by leveraging recent advances in
16 Energy-Based Modelling and Conformal Prediction. To our knowledge, this is the
17 first venture in this direction for generating faithful Counterfactual Explanations.
18 Thus, we anticipate that ECCCOs can serve as a baseline for future research. We
19 believe that our work opens avenues for researchers and practitioners seeking tools
20 to better distinguish trustworthy from unreliable models.

21 1 Introduction

22 Counterfactual Explanations (CE) provide a powerful, flexible and intuitive way to not only explain
23 black-box models but also help affected individuals through the means of Algorithmic Recourse.
24 Instead of opening the Black Box, CE works under the premise of strategically perturbing model
25 inputs to understand model behaviour [1]. Intuitively speaking, we generate explanations in this
26 context by asking what-if questions of the following nature: ‘Our credit risk model currently predicts
27 that this individual is not credit-worthy. What if they reduced their monthly expenditures by 10%?’

28 This is typically implemented by defining a target outcome $\mathbf{y}^+ \in \mathcal{Y}$ for some individual $\mathbf{x} \in \mathcal{X} = \mathbb{R}^D$
29 described by D attributes, for which the model $M_\theta : \mathcal{X} \mapsto \mathcal{Y}$ initially predicts a different outcome:
30 $M_\theta(\mathbf{x}) \neq \mathbf{y}^+$. Counterfactuals are then searched by minimizing a loss function that compares the
31 predicted model output to the target outcome: $\text{yloss}(M_\theta(\mathbf{x}), \mathbf{y}^+)$. Since CE work directly with the
32 black-box model, valid counterfactuals always have full local fidelity by construction where fidelity is
33 defined as the degree to which explanations approximate the predictions of a black-box model [2, 3].

34 In situations where full fidelity is a requirement, CE offer a more appropriate solution to Explainable
 35 Artificial Intelligence (XAI) than other popular approaches like LIME [4] and SHAP [5], which
 36 involve local surrogate models. But even full fidelity is not a sufficient condition for ensuring
 37 that an explanation faithfully describes the behaviour of a model. That is because multiple very
 38 distinct explanations can all lead to the same model prediction, especially when dealing with heavily
 39 parameterized models like deep neural networks, which are typically underspecified by the data [6].

40 In the context of CE, the idea that no two explanations are the same arises almost naturally. A key
 41 focus in the literature has therefore been to identify those explanations and algorithmic recourses that
 42 are most appropriate based on a myriad of desiderata such as sparsity, actionability and plausibility.
 43 In this work, we draw closer attention to model faithfulness rather than fidelity as a desideratum for
 44 counterfactuals. Our key contributions are as follows:

- 45 • We show that fidelity is an insufficient evaluation metric for counterfactuals (Section 3) and
 46 propose a definition of faithfulness that gives rise to more suitable metrics (Section 4).
- 47 • We introduce a novel algorithmic approach for generating Energy-Constrained Conformal
 48 Counterfactuals (ECCCoS) in Section 5.
- 49 • We provide extensive empirical evidence demonstrating that ECCCoS faithfully explain
 50 model behaviour and attain plausibility only when appropriate (Section 6).

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

55 2 Background

56 While CE can also be generated for arbitrary regression models [7], existing work has primarily
 57 focused on classification problems. Let $\mathcal{Y} = (0, 1)^K$ denote the one-hot-encoded output domain
 58 with K classes. Then most counterfactual generators rely on gradient descent to optimize different
 59 flavours of the following counterfactual search objective:

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

60 Here $\text{yloss}(\cdot)$ denotes the primary loss function, $f(\cdot)$ is a function that maps from the counterfactual
 61 state space to the feature space and $\text{cost}(\cdot)$ is either a single penalty or a collection of penalties that
 62 are used to impose constraints through regularization. Equation 1 restates the baseline approach to
 63 gradient-based counterfactual search proposed by Wachter et al. [1] in general form as introduced
 64 by Altmeyer et al. [8]. To explicitly account for the multiplicity of explanations, $\mathbf{Z}' = \{\mathbf{z}_l\}_L$ denotes
 65 an L -dimensional array of counterfactual states.

66 The baseline approach, which we will simply refer to as *Wachter*, searches a single counterfactual
 67 directly in the feature space and penalises its distance to the original factual. In this case, $f(\cdot)$ is
 68 simply the identity function and \mathcal{Z} corresponds to the feature space itself. Many derivative works
 69 of Wachter et al. [1] have proposed new flavours of Equation 1, each of them designed to address
 70 specific *desiderata* that counterfactuals ought to meet in order to properly serve both AI practitioners
 71 and individuals affected by algorithmic decision-making systems. The list of desiderata includes but
 72 is not limited to the following: sparsity, proximity [1], actionability [9], diversity [2], plausibility [10,
 73 11, 12], robustness [13, 14, 8] and causality [15]. Different counterfactual generators addressing
 74 these needs have been extensively surveyed and evaluated in various studies [16, 17, 18, 19, 20].

75 Perhaps unsurprisingly, the different desiderata are often positively correlated. For example, Artelt
 76 et al. [19] find that plausibility typically also leads to improved robustness. Similarly, plausibility has
 77 also been connected to causality in the sense that plausible counterfactuals respect causal relation-
 78 ships [21]. Consequently, the plausibility of counterfactuals has been among the primary concerns
 79 for researchers. Achieving plausibility is equivalent to ensuring that the generated counterfactuals
 80 comply with the true and unobserved data-generating process (DGP). We define plausibility formally
 81 in this work as follows:

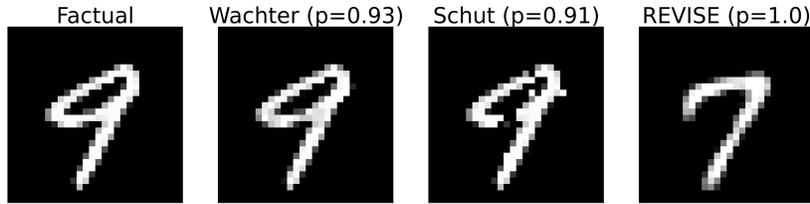


Figure 1: Counterfactuals for turning a 9 (nine) into a 7 (seven): original image (left); then from left to right the counterfactuals generated using *Wachter*, *Schut* and *REVISE*.

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

85 To generate plausible counterfactuals, we need to be able to quantify the DGP: $\mathcal{X}|\mathbf{y}^+$. One straight-
 86 forward way to do this is to use surrogate models for the task. Joshi et al. [10], for example, suggest
 87 that instead of searching counterfactuals in the feature space \mathcal{X} , we can instead traverse a latent
 88 embedding \mathcal{Z} (Equation 1) that implicitly codifies the DGP. To learn the latent embedding, they
 89 propose using a generative model such as a Variational Autoencoder (VAE). Provided the surrogate
 90 model is well-specified, their proposed approach called *REVISE* can yield plausible explanations.
 91 Others have proposed similar approaches: Dombrowski et al. [22] traverse the base space of a
 92 normalizing flow to solve Equation 1; Poyiadzi et al. [11] use density estimators ($\hat{p} : \mathcal{X} \mapsto [0, 1]$) to
 93 constrain the counterfactuals to dense regions in the feature space; and, finally, Karimi et al. [15]
 94 assume knowledge about the structural causal model that generates the data.

95 A competing approach towards plausibility that is also closely related to this work instead relies on
 96 the black-box model itself. Schut et al. [12] show that to meet the plausibility objective we need not
 97 explicitly model the input distribution. Pointing to the undesirable engineering overhead induced by
 98 surrogate models, they propose that we rely on the implicit minimisation of predictive uncertainty
 99 instead. Their proposed methodology, which we will refer to as *Schut*, solves Equation 1 by greedily
 100 applying Jacobian-Based Saliency Map Attacks (JSMA) in the feature space with cross-entropy loss
 101 and no penalty at all. The authors demonstrate theoretically and empirically that their approach yields
 102 counterfactuals for which the model M_θ predicts the target label \mathbf{y}^+ with high confidence. Provided
 103 the model is well-specified, these counterfactuals are plausible. This idea hinges on the assumption that
 104 the black-box model provides well-calibrated predictive uncertainty estimates.

105 3 Why Fidelity is not Enough

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

110 To demonstrate this with an example, we have trained a simple image classifier M_θ on the well-
 111 known *MNIST* dataset [23]: a Multi-Layer Perceptron (*MLP*) with above 90 percent test accuracy. No
 112 measures have been taken to improve the model’s adversarial robustness or its capacity for predictive
 113 uncertainty quantification. The far left panel of Figure 1 shows a random sample drawn from the
 114 dataset. The underlying classifier correctly predicts the label ‘nine’ for this image. For the given
 115 factual image and model, we have used *Wachter*, *Schut* and *REVISE* to generate one counterfactual
 116 each in the target class ‘seven’. The perturbed images are shown next to the factual image from left
 117 to right in Figure 1. Captions on top of the individual images indicate the generator along with the
 118 predicted probability that the image belongs to the target class. In all three cases that probability is
 119 above 90 percent and yet the counterfactuals look very different from each other.

120 Since *Wachter* is only concerned with proximity, the generated counterfactual is almost indistinguish-
 121 able from the factual. The approach by Schut et al. [12] expects a well-calibrated model that can
 122 generate predictive uncertainty estimates. Since this is not the case, the generated counterfactual
 123 looks like an adversarial example. Finally, the counterfactual generated by *REVISE* looks much more
 124 plausible than the other two. But is it also more faithful to the behaviour of our *MNIST* classifier?

125 That is much less clear because the surrogate used by *REVISE* introduces friction: the generated
 126 explanations no longer depend exclusively on the black-box model itself.

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

130 4 A New Notion of Faithfulness

131 Considering the limitations of fidelity as demonstrated in the previous section, analogous to Defini-
 132 tion 2.1, we introduce a new notion of faithfulness in the context of CE:

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

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

138 4.1 Quantifying the Model’s Generative Property

139 To assess counterfactuals with respect to Definition 4.1, we need a way to quantify the posterior
 140 conditional distribution $p_\theta(\mathbf{x}|y^+)$. To this end, we draw on recent advances in Energy-Based
 141 Modelling (EBM), a subdomain of machine learning that is concerned with generative or hybrid
 142 modelling [24, 25]. In particular, note that if we fix \mathbf{y} to our target value \mathbf{y}^+ , we can conditionally
 143 draw from $p_\theta(\mathbf{x}|y^+)$ by randomly initializing \mathbf{x}_0 and then using Stochastic Gradient Langevin
 144 Dynamics (SGLD) as follows,

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

145 where $\mathbf{r}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the stochastic term and the step-size ϵ is typically polynomially decayed [26].
 146 The term $\mathcal{E}(\mathbf{x}_j|y^+)$ denotes the model energy conditioned on the target class label y^+ which we
 147 specify as the negative logit corresponding to the target class label y^* . To allow for faster sampling,
 148 we follow the common practice of choosing the step-size ϵ and the standard deviation of \mathbf{r}_j separately.
 149 While \mathbf{x}_j is only guaranteed to distribute as $p_\theta(\mathbf{x}|y^*)$ if $\epsilon \rightarrow 0$ and $J \rightarrow \infty$, the bias introduced for
 150 a small finite ϵ is negligible in practice [27, 24]. Appendix A provides additional implementation
 151 details for any tasks related to energy-based modelling.

152 Generating multiple samples using SGLD thus yields an empirical distribution $\hat{\mathbf{X}}_{\theta, y^+}$ that approxi-
 153 mates what the model has learned about the input data. While in the context of EBM, this is usually
 154 done during training, we propose to repurpose this approach during inference in order to evaluate and
 155 generate faithful model explanations.

156 4.2 Evaluating Plausibility and Faithfulness

157 The parallels between our definitions of plausibility and faithfulness imply that we can also use
 158 similar evaluation metrics in both cases. Since existing work has focused heavily on plausibility,
 159 it offers a useful starting point. In particular, Guidotti [20] have proposed an implausibility metric
 160 that measures the distance of the counterfactual from its nearest neighbour in the target class. As
 161 this distance is reduced, counterfactuals get more plausible under the assumption that the nearest
 162 neighbour itself is plausible in the sense of Definition 2.1. In this work, we use the following adapted
 163 implausibility metric,

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

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

167 Equation 3 gives rise to a similar evaluation metric for unfaithfulness. We merely swap out the
 168 subsample of individuals in the target class for a subset $\hat{\mathbf{X}}_{\theta, \mathbf{y}^+}^{n_E}$ of the generated conditional samples:

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

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

170 5 Energy-Constrained Conformal Counterfactuals

171 In this section, we describe *ECCCo*, our proposed framework for generating Energy-Constrained
 172 Conformal Counterfactuals (ECCCOs). It is based on the premise that counterfactuals should first
 173 and foremost be faithful. Plausibility, as a secondary concern, is then still attainable, but only to the
 174 degree that the black-box model itself has learned plausible explanations for the underlying data.

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

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

177 The first penalty term involving λ_1 induces proximity like in Wachter et al. [1]. Our default choice
 178 for $\text{dist}(\cdot)$ is the L1 Norm due to its sparsity-inducing properties. The second penalty term involving
 179 λ_2 induces faithfulness by constraining the energy of the generated counterfactual where $\text{unfaith}(\cdot)$
 180 corresponds to the metric defined in Equation 4. The third and final penalty term involving λ_3
 181 introduces a new concept: it ensures that the generated counterfactual is associated with low predictive
 182 uncertainty. As mentioned above, Schut et al. [12] have shown that plausible counterfactuals can
 183 be generated implicitly through predictive uncertainty minimization. Unfortunately, this relies on
 184 the assumption that the model itself can provide predictive uncertainty estimates, which may be too
 185 restrictive in practice.

186 To relax this assumption, we leverage recent advances in Conformal Prediction (CP), an approach to
 187 predictive uncertainty quantification that has recently gained popularity [28, 29]. Crucially for our
 188 intended application, CP is model-agnostic and can be applied during inference without placing any
 189 restrictions on model training. Intuitively, CP works under the premise of turning heuristic notions of
 190 uncertainty into rigorous uncertainty estimates by repeatedly sifting through the training data or a
 191 dedicated calibration dataset. Conformal classifiers produce prediction sets for individual inputs that
 192 include all output labels that can be reasonably attributed to the input. These sets tend to be larger for
 193 inputs that do not conform with the training data and are characterized by high predictive uncertainty.

194 In order to generate counterfactuals that are associated with low predictive uncertainty, we use a
 195 smooth set size penalty introduced by Stutz et al. [30] in the context of conformal training:

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

196 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
 197 \mathbf{y} being included in the prediction set. In order to compute this penalty for any black-box model
 198 we merely need to perform a single calibration pass through a holdout set \mathcal{D}_{cal} . Arguably, data is
 199 typically abundant and in most applications, practitioners tend to hold out a test data set anyway.
 200 Consequently, CP removes the restriction on the family of predictive models, at the small cost of
 201 reserving a subset of the available data for calibration. This particular case of conformal prediction is
 202 referred to as Split Conformal Prediction (SCP) as it involves splitting the training data into a proper
 203 training dataset and a calibration dataset. In addition to the smooth set size penalty, we have also
 204 experimented with the use of a tailored function for $\text{yloss}(\cdot)$ that enforces that only the target label
 205 \mathbf{y}^+ is included in the prediction set Stutz et al. [30]. Further details are provided in Appendix B.

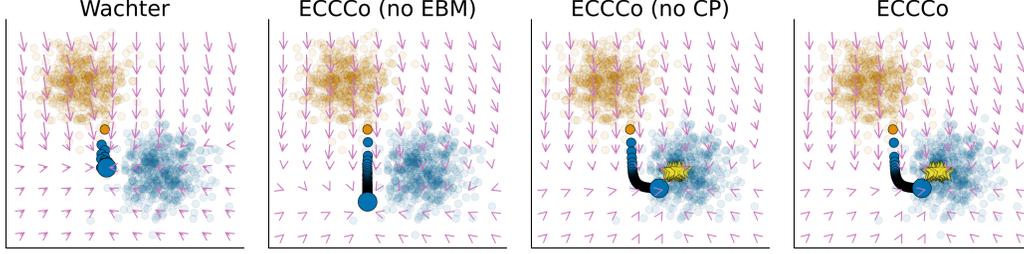


Figure 2: Gradient fields and counterfactual paths for different generators. The objective is to generate a counterfactual in the ‘blue’ class for a sample from the ‘orange’ class. Bright yellow stars indicate conditional samples generated through SGLD. The underlying classifier is a Joint Energy Model.

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 $\hat{\mathbf{X}}_{\theta, \mathbf{y}^+}^{n_E} \leftarrow \{\hat{\mathbf{x}}_{\theta, \mathbf{y}^+}\}_{n_B}$ ▷ 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 Appendix C.
 - 7: $\mathbf{z}' \leftarrow \mathbf{z}' - \eta \nabla_{\mathbf{z}'} \mathcal{L}(\mathbf{z}', \mathbf{y}^+, \hat{\mathbf{X}}_{\theta, \mathbf{y}^+}^{n_E}; \Lambda, \alpha)$ ▷ Take gradient step of size η .
 - 8: $t \leftarrow t + 1$
 - 9: **end while**
 - 10: $\mathbf{x}' \leftarrow f(\mathbf{z}')$ ▷ Map back to feature space.
-

206 To provide some further intuition about our objective defined in Equation 5, Figure 2 illustrates how
 207 the different components affect the counterfactual search for a synthetic dataset. The underlying
 208 classifier is a Joint Energy Model (*JEM*) that was trained to predict the output class (‘blue’ or
 209 ‘orange’) and generate class-conditional samples [24]. We have used four different generator flavours
 210 to produce a counterfactual in the ‘blue’ class for a sample from the ‘orange’ class: *Wachter*, which
 211 only uses the first penalty ($\lambda_2 = \lambda_3 = 0$); *ECCCo (no EBM)*, which does not constrain energy
 212 ($\lambda_2 = 0$); *ECCCo (no CP)*, which involves no set size penalty ($\lambda_3 = 0$); and, finally, *ECCCo*, which
 213 involves all penalties defined in Equation 5. Arrows indicate (negative) gradients with respect to the
 214 objective function at different points in the feature space.

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

222 Algorithm 1 describes how exactly *ECCCo* works. For the sake of simplicity and without loss of
 223 generality, we limit our attention to generating a single counterfactual $\mathbf{x}' = f(\mathbf{z}')$. The counterfactual
 224 state \mathbf{z}' is initialized by passing the factual \mathbf{x} through a simple feature transformer f^{-1} . Next, we
 225 generate n_B conditional samples $\hat{\mathbf{x}}_{\theta, \mathbf{y}^+}$ using SGLD (Equation 2) and store the n_E instances with
 226 the lowest energy. We then calibrate the model M_θ through Split Conformal Prediction. Finally,
 227 we search counterfactuals through gradient descent where $\mathcal{L}(\mathbf{z}', \mathbf{y}^+, \hat{\mathbf{X}}_{\theta, \mathbf{y}^+}^{n_E}; \Lambda, \alpha)$ denotes our loss
 228 function defined in Equation 5. The search terminates once the convergence criterium is met or the
 229 maximum number of iterations T has been exhausted. Note that the choice of convergence criterium
 230 has important implications on the final counterfactual which we explain in Appendix C.

231 6 Empirical Analysis

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

233 **Research Question 6.1** (Faithfulness). *Are ECCCoS more faithful than counterfactuals produced by*
234 *our benchmark generators?*

235 **Research Question 6.2** (Balancing Objectives). *Compared to our benchmark generators, how do*
236 *ECCCoS balance the two key objectives of faithfulness and plausibility?*

237 The second question is motivated by the intuition that faithfulness and plausibility should coincide
238 for models that have learned plausible explanations of the data. Next, we first briefly describe our
239 experimental setup before presenting our main results.

240 6.1 Experimental Setup

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

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

252 As for real-world data, we follow Schut et al. [12] and use the *MNIST* [23] dataset containing images
253 of handwritten digits such as the example shown above in Figure 1. From the social sciences domain,
254 we include Give Me Some Credit (*GMSC*) [31]: a tabular dataset that has been studied extensively in
255 the literature on Algorithmic Recourse [18]. It consists of 11 numeric features that can be used to
256 predict the binary outcome variable indicating whether retail borrowers experience financial distress.

257 For the predictive modelling tasks, we use simple neural networks (*MLP*) and Joint Energy Models
258 (*JEM*). For the more complex real-world datasets we also use ensembling in each case. Both joint-
259 energy modelling and ensembling have been associated with improved generative properties and
260 adversarial robustness [24, 32], so we expect this to be positively correlated with the plausibility
261 of ECCCoS. To account for stochasticity, we generate multiple counterfactuals for each target
262 class, generator, model and dataset. Specifically, we randomly sample n^- times from the subset
263 of individuals for which the given model predicts the non-target class y^- given the current target.
264 We set $n^- = 25$ for all of our synthetic datasets, $n^- = 10$ for *GMSC* and $n^- = 5$ for *MNIST*. Full
265 details concerning our parameter choices, training procedures and model performance can be found
266 in Appendix D.

267 6.2 Results for Synthetic Data

268 Table 1 shows the key results for the synthetic datasets separated by model (first column) and generator
269 (second column). The numerical columns show sample averages and standard deviations of our key
270 evaluation metrics computed across all counterfactuals. We have highlighted the best outcome for
271 each model and metric in bold. To provide some sense of effect sizes, we have added asterisks to
272 indicate that a given value is at least one (*) or two (**) standard deviations lower than the baseline
273 (*Wachter*).

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

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

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

291 For the *Circles* data, it appears that *REVISE* performs well, but we note that it generates valid
292 counterfactuals only half of the time (see Appendix E for a complete overview including additional
293 common evaluation metrics). The underlying VAE with default parameters has not adequately learned
294 the data-generating process. Of course, it is possible to improve generative performance through
295 hyperparameter tuning but this example serves to illustrate that *REVISE* depends on the quality of its
296 surrogate. Independent of the outcome for *REVISE*, however, the results do not seem to indicate that
297 *ECCCo* substantially improves faithfulness and plausibility for the *Circles* data. We think this points
298 to a limitation of our evaluation metrics rather than *ECCCo* itself: computing average distances fails
299 to account for the ‘wraparound’ effect associated with circular data [33].

300 6.3 Results for Real-World Data

301 The results for our real-world datasets are shown in Table 2. Once again the findings indicate that the
302 plausibility of *ECCCo*s is positively correlated with the capacity of the black-box model to distinguish
303 plausible from implausible inputs. The case is very clear for *MNIST*: *ECCCo*s are consistently more
304 faithful than the counterfactuals produced by our benchmark generators and their plausibility gradually
305 improves through ensembling and joint-energy modelling. Interestingly, faithfulness also gradually
306 improves for *REVISE*. This indicates that as our models improve, their generative capacity approaches
307 that of the surrogate VAE used by *REVISE*. The VAE still outperforms our classifiers in this regard,
308 as evident from the fact that *ECCCo* never quite reaches the same level of plausibility as *REVISE*.
309 With reference to Appendix E we note that the results for *Schut* need to be discounted as it rarely
310 produces valid counterfactuals for *MNIST*. Relatedly, we find that *ECCCo* is the only generator that
311 consistently achieves full validity. Finally, it is worth noting that *ECCCo* produces counterfactual
312 images with the lowest average predictive uncertainty for all models.

313 For the tabular credit dataset (*GMSC*) it is inherently challenging to use deep neural networks in order
314 to achieve good discriminative performance [34, 35] and generative performance [36], respectively. In
315 order to achieve high plausibility, *ECCCo* effectively requires classifiers to achieve good performance
316 for both tasks. Since this is a challenging task even for Joint Energy Models, it is not surprising to
317 find that even though *ECCCo* once again achieves state-of-the-art faithfulness, it is outperformed by
318 *REVISE* and *Schut* with respect to plausibility.

Table 2: Results for real-world datasets: sample averages +/- 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 ↓	Implausibility ↓	Unfaithfulness ↓	Implausibility ↓
JEM	ECCCo	19.28 ± 5.01**	314.76 ± 32.36*	79.16 ± 11.67**	18.26 ± 4.92**
	REVISE	188.70 ± 26.18*	255.26 ± 41.50**	186.40 ± 28.06	5.34 ± 2.38**
	Schut	211.00 ± 27.21	286.61 ± 39.85*	200.98 ± 28.49	6.50 ± 2.01**
	Wachter	222.90 ± 26.56	361.88 ± 39.74	214.08 ± 45.35	61.04 ± 2.58
JEM Ensemble	ECCCo	15.99 ± 3.06**	294.72 ± 30.75**	83.28 ± 13.26**	17.21 ± 4.46**
	REVISE	173.59 ± 20.65**	246.32 ± 37.46**	194.24 ± 35.41	4.95 ± 1.26**
	Schut	205.33 ± 24.07	287.39 ± 39.33*	208.45 ± 34.60	6.12 ± 1.91**
	Wachter	217.67 ± 23.78	363.23 ± 39.24	186.19 ± 33.88	60.70 ± 44.32
MLP	ECCCo	41.95 ± 6.50**	591.58 ± 36.24	75.93 ± 14.27**	17.20 ± 3.15**
	REVISE	365.82 ± 15.35*	249.49 ± 41.55**	196.75 ± 41.25	4.84 ± 0.60**
	Schut	382.44 ± 17.81	285.98 ± 42.48*	212.00 ± 41.15	6.44 ± 1.34**
	Wachter	386.05 ± 16.60	361.83 ± 42.18	218.34 ± 53.26	45.84 ± 39.39
MLP Ensemble	ECCCo	31.43 ± 3.91**	490.88 ± 27.19	73.86 ± 14.63**	17.92 ± 4.17**
	REVISE	337.74 ± 11.89*	247.67 ± 38.36**	207.21 ± 43.20	5.78 ± 2.10**
	Schut	359.54 ± 14.52	283.99 ± 41.08*	205.36 ± 32.11	7.00 ± 2.15**
	Wachter	360.79 ± 14.39	357.73 ± 42.55	213.71 ± 54.17	73.09 ± 64.50

319 6.4 Key Takeways

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

325 7 Limitations

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

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

333 While our approach is readily applicable to models with gradient access like deep neural networks,
 334 more work is needed to generalise it to other machine learning models such as decision trees.
 335 Relatedly, common challenges associated with Energy-Based Modelling including sensitivity to scale,
 336 training instabilities and sensitivity to hyperparameters also apply to *ECCCo*.

337 8 Conclusion

338 This work leverages recent advances in Energy-Based Modelling and Conformal Prediction in the
 339 context of Explainable Artificial Intelligence. We have proposed a new way to generate counterfactuals
 340 that are maximally faithful to the black-box model they aim to explain. Our proposed generator,
 341 *ECCCo*, produces plausible counterfactuals if and only if the black-box model itself has learned
 342 realistic explanations for the data, which we have demonstrated through rigorous empirical analysis.
 343 This should enable researchers and practitioners to use counterfactuals in order to discern trustworthy
 344 models from unreliable ones. While the scope of this work limits its generalizability, we believe that
 345 *ECCCo* offers a solid baseline for future work on faithful Counterfactual Explanations.

References

- [1] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31:841, 2017.
- [2] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.
- [3] Christoph Molnar. *Interpretable Machine Learning*. Lulu. com, 2020.
- [4] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.
- [5] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4768–4777, 2017.
- [6] Andrew Gordon Wilson. The case for Bayesian deep learning. 2020.
- [7] Thomas Spooner, Danial Dervovic, Jason Long, Jon Shepard, Jiahao Chen, and Daniele Magazzeni. Counterfactual Explanations for Arbitrary Regression Models. 2021.
- [8] Patrick Altmeyer, Giovan Angela, Aleksander Buszydluk, Karol Dobiczek, Arie van Deursen, and Cynthia Liem. Endogenous Macrodynamics in Algorithmic Recourse. In *First IEEE Conference on Secure and Trustworthy Machine Learning, 2023*.
- [9] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19, 2019.
- [10] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. 2019.
- [11] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. FACE: Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, 2020.
- [12] Lisa Schut, Oscar Key, Rory Mc Grath, Luca Costabello, Bogdan Sacaleanu, Yarin Gal, et al. Generating Interpretable Counterfactual Explanations By Implicit Minimisation of Epistemic and Aleatoric Uncertainties. In *International Conference on Artificial Intelligence and Statistics*, pages 1756–1764. PMLR, 2021.
- [13] Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. Towards Robust and Reliable Algorithmic Recourse. 2021.
- [14] Martin Pawelczyk, Teresa Datta, Johannes van-den Heuvel, Gjergji Kasneci, and Himabindu Lakkaraju. Probabilistically Robust Recourse: Navigating the Trade-offs between Costs and Robustness in Algorithmic Recourse. *arXiv preprint arXiv:2203.06768*, 2022.
- [15] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: From counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 353–362, 2021.
- [16] Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. 2020.
- [17] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: Definitions, formulations, solutions, and prospects. 2020.
- [18] Martin Pawelczyk, Sascha Bielawski, Johannes van den Heuvel, Tobias Richter, and Gjergji Kasneci. Carla: A python library to benchmark algorithmic recourse and counterfactual explanation algorithms. 2021.

- 393 [19] André Artelt, Valerie Vaquet, Riza Velioglu, Fabian Hinder, Johannes Brinkrolf, Malte Schilling,
394 and Barbara Hammer. Evaluating Robustness of Counterfactual Explanations. Technical report,
395 arXiv. URL <http://arxiv.org/abs/2103.02354>. arXiv:2103.02354 [cs] type: article.
- 396 [20] Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and
397 benchmarking. ISSN 1573-756X. doi: 10.1007/s10618-022-00831-6. URL <https://doi.org/10.1007/s10618-022-00831-6>.
- 399 [21] Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving Causal Constraints in Counter-
400 factual Explanations for Machine Learning Classifiers. Technical report, arXiv. URL
401 <http://arxiv.org/abs/1912.03277>. arXiv:1912.03277 [cs, stat] type: article.
- 402 [22] Ann-Kathrin Dombrowski, Jan E Gerken, and Pan Kessel. Diffeomorphic explanations with
403 normalizing flows. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and*
404 *Explicit Likelihood Models*, 2021.
- 405 [23] Yann LeCun. The MNIST database of handwritten digits. 1998.
- 406 [24] Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad
407 Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should
408 treat it like one. March 2020. URL <https://openreview.net/forum?id=HkxzxONtDB>.
- 409 [25] Yilun Du and Igor Mordatch. Implicit Generation and Generalization in Energy-Based Models.
410 Technical report, arXiv. URL <http://arxiv.org/abs/1903.08689>. arXiv:1903.08689 [cs,
411 stat] type: article.
- 412 [26] M. Welling and Y. Teh. Bayesian Learning via Stochastic Gradient
413 Langevin Dynamics. URL [https://www.semanticscholar.org/paper/
414 Bayesian-Learning-via-Stochastic-Gradient-Langevin-Welling-Teh/
415 aeed631d6a84100b5e9a021ec1914095c66de415](https://www.semanticscholar.org/paper/Bayesian-Learning-via-Stochastic-Gradient-Langevin-Welling-Teh/aeed631d6a84100b5e9a021ec1914095c66de415).
- 416 [27] Kevin P. Murphy. *Probabilistic machine learning: Advanced topics*. MIT Press.
- 417 [28] Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction
418 and distribution-free uncertainty quantification. 2021.
- 419 [29] Valery Manokhin. Awesome conformal prediction.
- 420 [30] David Stutz, Krishnamurthy Dj Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning
421 Optimal Conformal Classifiers. May 2022. URL [https://openreview.net/forum?id=
422 t80-4LKFVx](https://openreview.net/forum?id=t80-4LKFVx).
- 423 [31] Kaggle. Give me some credit, Improve on the state of the art in credit scoring by predicting the
424 probability that somebody will experience financial distress in the next two years., 2011. URL
425 <https://www.kaggle.com/c/GiveMeSomeCredit>.
- 426 [32] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable
427 predictive uncertainty estimation using deep ensembles. 2016.
- 428 [33] Jeff Gill and Dominik Hangartner. Circular Data in Political Science and How to
429 Handle It. 18(3):316–336. ISSN 1047-1987, 1476-4989. doi: 10.1093/pan/mpq009.
430 URL [https://www.cambridge.org/core/journals/political-analysis/
431 article/circular-data-in-political-science-and-how-to-handle-it/
432 6DF2D9DA60C455E6A48FFB0FF011F747](https://www.cambridge.org/core/journals/political-analysis/article/circular-data-in-political-science-and-how-to-handle-it/6DF2D9DA60C455E6A48FFB0FF011F747).
- 433 [34] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and
434 Gjergji Kasneci. Deep neural networks and tabular data: A survey. 2021.
- 435 [35] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still
436 outperform deep learning on tabular data? 2022.
- 437 [36] Tennison Liu, Zhaozhi Qian, Jeroen Berrevoets, and Mihaela van der Schaar. GOGGLE:
438 Generative Modelling for Tabular Data by Learning Relational Structure. URL [https://
439 openreview.net/forum?id=fPVRcJqspu](https://openreview.net/forum?id=fPVRcJqspu).
- 440 [37] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. Technical
441 report, arXiv. URL <http://arxiv.org/abs/1412.6980>. arXiv:1412.6980 [cs] type: article.

442 **Appendices**

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

450 **A Energy-Based Modelling**

451 Since we were not able to identify any existing open-source software for Energy-Based Modelling
 452 that would be flexible enough to cater to our needs, we have developed a Julia package from scratch.
 453 The package has been open-sourced, but to avoid compromising the double-blind review process, we
 454 refrain from providing more information at this stage. In our development we have heavily drawn on
 455 the existing literature: Du and Mordatch [25] describe best practices for using EBM for generative
 456 modelling; Grathwohl et al. [24] explain how EBM can be used to train classifiers jointly for the
 457 discriminative and generative tasks. We have used the same package for training and inference, but
 458 there are some important differences between the two cases that are worth highlighting here.

459 **A.1 Training: Joint Energy Models**

460 To train our Joint Energy Models we broadly follow the approach outlined in Grathwohl et al. [24].
 461 These models are trained to optimize a hybrid objective that involves a standard classification loss
 462 component $L_{\text{clf}}(\theta) = -\log p_{\theta}(\mathbf{y}|\mathbf{x})$ (e.g. cross-entropy loss) as well as a generative loss component
 463 $L_{\text{gen}}(\theta) = -\log p_{\theta}(\mathbf{x})$.

464 To draw samples from $p_{\theta}(\mathbf{x})$, we rely exclusively on the conditional sampling approach described
 465 in Grathwohl et al. [24] for both training and inference: we first draw $\mathbf{y} \sim p(\mathbf{y})$ and then sample
 466 $\mathbf{x} \sim p_{\theta}(\mathbf{x}|\mathbf{y})$ [24] 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
 467 the linear predictions (logits) of our classifier M_{θ} . While our package also supports unconditional
 468 sampling, we found conditional sampling to work well. It is also well aligned with CE, since in this
 469 context we are interested in conditioning on the target class.

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

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

473 Consistent with Grathwohl et al. [24], we have specified $\epsilon = 2$ and $\sigma = 0.01$ as the default values for
 474 all of our experiments. The number of total SGLD steps J varies by dataset (Table 3). Following best
 475 practices, we initialize \mathbf{x}_0 randomly in 5% of all cases and sample from a buffer in all other cases.
 476 The buffer itself is randomly initialised and gradually grows to a maximum of 10,000 samples during
 477 training as $\hat{\mathbf{x}}_J$ is stored in each epoch [25, 24].

478 It is important to realise that sampling is done during each training epoch, which makes training Joint
 479 Energy Models significantly harder than conventional neural classifiers. In each epoch the generated
 480 (batch of) sample(s) $\hat{\mathbf{x}}_J$ is used as part of the generative loss component, which compares its energy
 481 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
 482 summarized as follows,

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

483 where $L_{\text{reg}}(\theta)$ is a Ridge penalty (L2 norm) that regularises energy magnitudes for both observed and
 484 generated samples [25]. We have used varying degrees of regularization depending on the dataset (λ
 485 in Table 3).

486 Contrary to existing work, we have not typically used the entire minibatch of training data for the
 487 generative loss component but found that using a subset of the minibatch was often sufficient in

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

JEM Ensemble

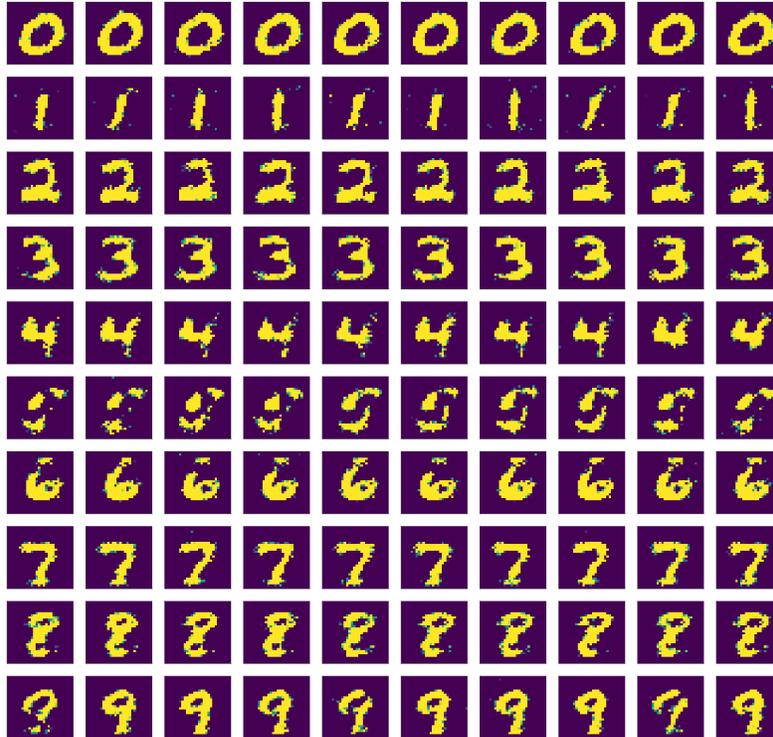


Figure 3: Conditionally generated *MNIST* images for our JEM Ensemble.

488 attaining decent generative performance (Table 3). This has helped to reduce the computational
 489 burden for our models, which should make it easier for others to reproduce our findings. Figures 3
 490 and 4 show generated samples for our *MNIST* and *Moons* data, to provide a sense of their generative
 491 property.

492 A.2 Inference: Quantifying Models’ Generative Property

493 At inference time, we assume no prior knowledge about the model’s generative property. This means
 494 that we do not tap into the existing buffer of generated samples for our Joint Energy Models, but
 495 instead generate conditional samples from scratch. While we have relied on the default values $\epsilon = 2$
 496 and $\sigma = 0.01$ also during inference, the number of total SGLD steps was set to $J = 500$ in all cases,
 497 so significantly higher than during training. For all of our synthetic datasets and models, we generated
 498 50 conditional samples and then formed subsets containing the $n_E = 25$ lowest-energy samples.
 499 While in practice it would be sufficient to do this once for each model and dataset, we have chosen
 500 to perform sampling separately for each individual counterfactual in our experiments to account for
 501 stochasticity. To help reduce the computational burden for our real-world datasets we have generated
 502 only 10 conditional samples each time and used all of them in our counterfactual search. Using more
 503 samples, as we originally did, had no substantial impact on our results.

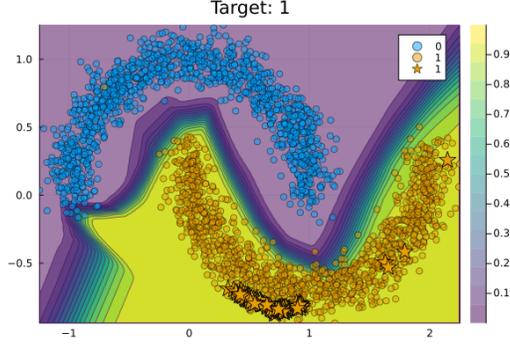


Figure 4: Conditionally generated samples (stars) for our *Moons* data using a JEM.

504 B Conformal Prediction

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

507 B.1 Background on CP

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

514 Specifically, SCP involves splitting the data $\mathcal{D}_n = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1, \dots, n}$ into a proper training set $\mathcal{D}_{\text{train}}$
 515 and a calibration set \mathcal{D}_{cal} . The former is used to train the classifier in any conventional fashion.
 516 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
 517 $s : (\mathcal{X}, \mathcal{Y}) \mapsto \mathbb{R}$ is referred to as *score function*. In the context of classification, a common choice for
 518 the score function is just $s_i = 1 - M_\theta(\mathbf{x}_i)[\mathbf{y}_i]$, that is one minus the softmax output corresponding
 519 to the observed label \mathbf{y}_i [28].

520 Finally, classification sets are formed as follows,

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

521 where \hat{q} denotes the $(1 - \alpha)$ -quantile of \mathcal{S} and α is a predetermined error rate. As the size of the
 522 calibration set increases, the probability that the classification set $C(\mathbf{x}_{\text{test}})$ for a newly arrived sample
 523 \mathbf{x}_{test} does not cover the true test label \mathbf{y}_{test} approaches α [28].

524 Observe from Equation 9 that Conformal Prediction works on an instance-level basis, much like CE
 525 are local. The prediction set for an individual instance \mathbf{x}_i depends only on the characteristics of that
 526 sample and the specified error rate. Intuitively, the set is more likely to include multiple labels for
 527 samples that are difficult to classify, so the set size is indicative of predictive uncertainty. To see why
 528 this effect is exacerbated by small choices for α consider the case of $\alpha = 0$, which requires that the
 529 true label is covered by the prediction set with probability equal to 1.

530 B.2 Differentiability

531 The fact that conformal classifiers produce set-valued predictions introduces a challenge: it is not
 532 immediately obvious how to use such classifiers in the context of gradient-based counterfactual
 533 search. Put differently, it is not clear how to use prediction sets in Equation 1. Fortunately, Stutz et al.
 534 [30] have recently proposed a framework for Conformal Training that also hinges on differentiability.
 535 Specifically, they show how Stochastic Gradient Descent can be used to train classifiers not only
 536 for the discriminative task but also for additional objectives related to Conformal Prediction. One
 537 such objective is *efficiency*: for a given target error rate α , the efficiency of a conformal classifier

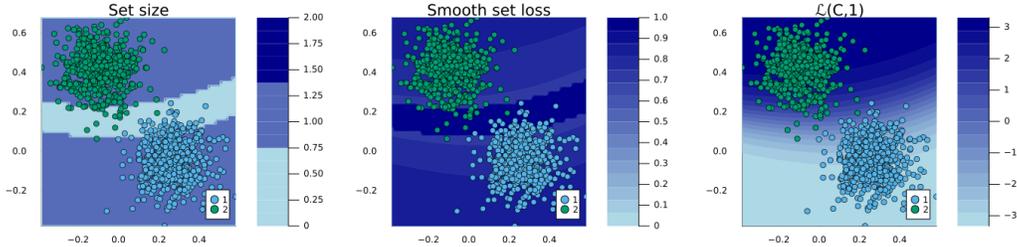


Figure 5: Prediction set size (left), smooth set size loss (centre) and configurable classification loss (right) for a JEM trained on our *Linearly Separable* data.

538 improves as its average prediction set size decreases. To this end, the authors introduce a smooth set
 539 size penalty defined in Equation 6 in the body of this paper. Formally, it is defined as $C_{\theta, \mathbf{y}}(\mathbf{x}_i; \alpha) :=$
 540 $\sigma((s(\mathbf{x}_i, \mathbf{y}) - \alpha)T^{-1})$ for $\mathbf{y} \in \mathcal{Y}$, where σ is the sigmoid function and T is a hyper-parameter used
 541 for temperature scaling [30].

542 In addition to the smooth set size penalty, Stutz et al. [30] also propose a configurable classification
 543 loss function, that can be used to enforce coverage. For *MNIST* data, we found that using this
 544 function generally improved the visual quality of the generated counterfactuals, so we used it
 545 in our experiments involving real-world data. For the synthetic dataset, visual inspection of the
 546 counterfactuals showed that using the configurable loss function sometimes led to overshooting:
 547 counterfactuals would end up deep inside the target domain but far away from the observed samples.
 548 For this reason, we instead relied on standard cross-entropy loss for our synthetic datasets. As we have
 549 noted in the body of the paper, more experimental work is certainly needed in this context. Figure 5
 550 shows the prediction set size (left), smooth set size loss (centre) and configurable classification loss
 551 (right) for a JEM trained on our *Linearly Separable* data.

552 C ECCCo

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

555 C.1 A Note on Convergence

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

566 C.2 ECCCo.jl

567 The core part of our code base is integrated into a larger ecosystem of Julia packages that we
 568 are actively developing and maintaining. To avoid compromising the double-blind review process,
 569 we only provide a link to an anonymized repository at this stage: <https://anonymous.4open.science/r/ECCCo-1252/README.md>.
 570

571 D Experimental Setup

572 Table 4 provides an overview of all parameters related to our experiments. The *GMSC* data were
 573 randomly undersampled for balancing purposes and all features were standardized. *MNIST* data was
 574 also randomly undersampled for reasons outlined below. Pixel values were preprocessed to fall in the
 575 range of $[-1, 1]$ and a small Gaussian noise component ($\sigma = 0.03$) was added to training samples

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

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

576 following common practice in the EBM literature. All of our models were trained through mini-batch
 577 training using the Adam optimiser (Kingma and Ba [37]). Table 5 shows standard evaluation metrics
 578 measuring the predictive performance of our different models grouped by dataset. These measures
 579 were computed on test data.

580 Table 6 summarises our hyperparameter choices for the counterfactual generators where η denotes
 581 the learning rate used for Stochastic Gradient Descent (SGD) and $\lambda_1, \lambda_2, \lambda_3$ represent the chosen
 582 penalty strengths (Equations 1 and 5). Here λ_1 also refers to the chosen penalty for the distance from
 583 factual values that applies to both *Wachter* and *REVISE*, but not *Schut* which is penalty-free. *Schut* is
 584 also the only generator that uses JSMA instead of SGD for optimization.

585 D.1 Compute

586 To enable others to easily replicate our experiments, we have chosen to work with small neural
 587 network architectures and randomly undersampled the *MNIST* dataset (maintaining class balance).
 588 All of our experiments could then be run locally on a personal machine. The longest runtimes we

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

589 experienced for model training and counterfactual benchmarking were on the order of 8-12 hours
590 (*MNIST* data). For the synthetic data, all experiments could be completed in less than an hour.

591 We have summarised our system information below:

592 **Software:**

- 593 • System Version: macOS 13.3.1
- 594 • Kernel Version: Darwin 22.4.0

595 **Hardware:**

- 596 • Model Name: MacBook Pro
- 597 • Model Identifier: MacBookPro16,1
- 598 • Processor Name: 8-Core Intel Core i9
- 599 • Processor Speed: 2.3 GHz
- 600 • Number of Processors: 1
- 601 • Total Number of Cores: 8
- 602 • L2 Cache (per Core): 256 KB
- 603 • L3 Cache: 16 MB
- 604 • Hyper-Threading Technology: Enabled
- 605 • Memory: 32 GB

606 **E Results**

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

610 Table 7 reports all of the evaluation metrics we have computed. Table 8 reports the same metrics
611 for the subset of valid counterfactuals. The ‘Unfaithfulness’ and ‘Implausibility’ metrics have been
612 discussed extensively in the body of the paper. The ‘Cost’ metric relates to the distance between
613 the factual and the counterfactual. The ‘Redundancy’ metric measures sparsity in is defined as the
614 percentage of features that remain unperturbed (higher is better). The ‘Uncertainty’ metric is just
615 the average value of the smooth set size penalty (Equation 6). Finally, ‘Validity’ is the percentage of
616 valid counterfactuals.

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	ECCCo	0.74 ± 0.21	0.52 ± 0.36	1.22 ± 0.46	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00**	
		ECCCo (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**	
		ECCCo (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	ECCCo	0.67 ± 0.19	0.65 ± 0.53	1.17 ± 0.41	0.00 ± 0.00	0.09 ± 0.19**	1.00 ± 0.00	
		ECCCo (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	
		ECCCo (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	JEM	ECCCo	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	ECCCo	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	ECCCo	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	ECCCo	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	ECCCo	0.75 ± 0.17	0.03 ± 0.06**	0.20 ± 0.08**	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
			ECCCo (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
			ECCCo (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	ECCCo	0.95 ± 0.16	0.29 ± 0.05**	0.23 ± 0.06**	0.00 ± 0.00	0.00 ± 0.00**	1.00 ± 0.00
			ECCCo (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
ECCCo (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		JEM	ECCCo	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	ECCCo	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	ECCCo	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	ECCCo	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	ECCCo	1.56 ± 0.44	0.31 ± 0.30*	1.20 ± 0.15**	0.00 ± 0.00	0.00 ± 0.00**	1.00 ± 0.00**
			ECCCo (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**
			ECCCo (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	ECCCo	2.18 ± 1.05	0.80 ± 0.62	1.69 ± 0.40	0.00 ± 0.00	0.15 ± 0.24*	1.00 ± 0.00
			ECCCo (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
ECCCo (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	ECCCo	0.74 ± 0.21	0.52 ± 0.36	1.22 ± 0.46	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00	
		ECCCo (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	
		ECCCo (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	ECCCo	0.67 ± 0.19	0.65 ± 0.53	1.17 ± 0.41	0.00 ± 0.00	0.09 ± 0.19**	1.00 ± 0.00	
		ECCCo (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	
		ECCCo (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	ECCCo	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	ECCCo	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	ECCCo	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	ECCCo	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	ECCCo	0.75 ± 0.17	0.03 ± 0.06**	0.20 ± 0.08**	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
			ECCCo (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
ECCCo (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		ECCCo	0.95 ± 0.16	0.29 ± 0.05**	0.23 ± 0.06**	0.00 ± 0.00	0.00 ± 0.00**	1.00 ± 0.00	
		ECCCo (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	
		ECCCo (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	ECCCo	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	ECCCo	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	ECCCo	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	ECCCo	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	ECCCo	1.56 ± 0.44	0.31 ± 0.30*	1.20 ± 0.15**	0.00 ± 0.00	0.00 ± 0.00**	1.00 ± 0.00	
		ECCCo (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	
		ECCCo (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	ECCCo	2.18 ± 1.05	0.80 ± 0.62	1.69 ± 0.40	0.00 ± 0.00	0.15 ± 0.24*	1.00 ± 0.00	
		ECCCo (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	
		ECCCo (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	

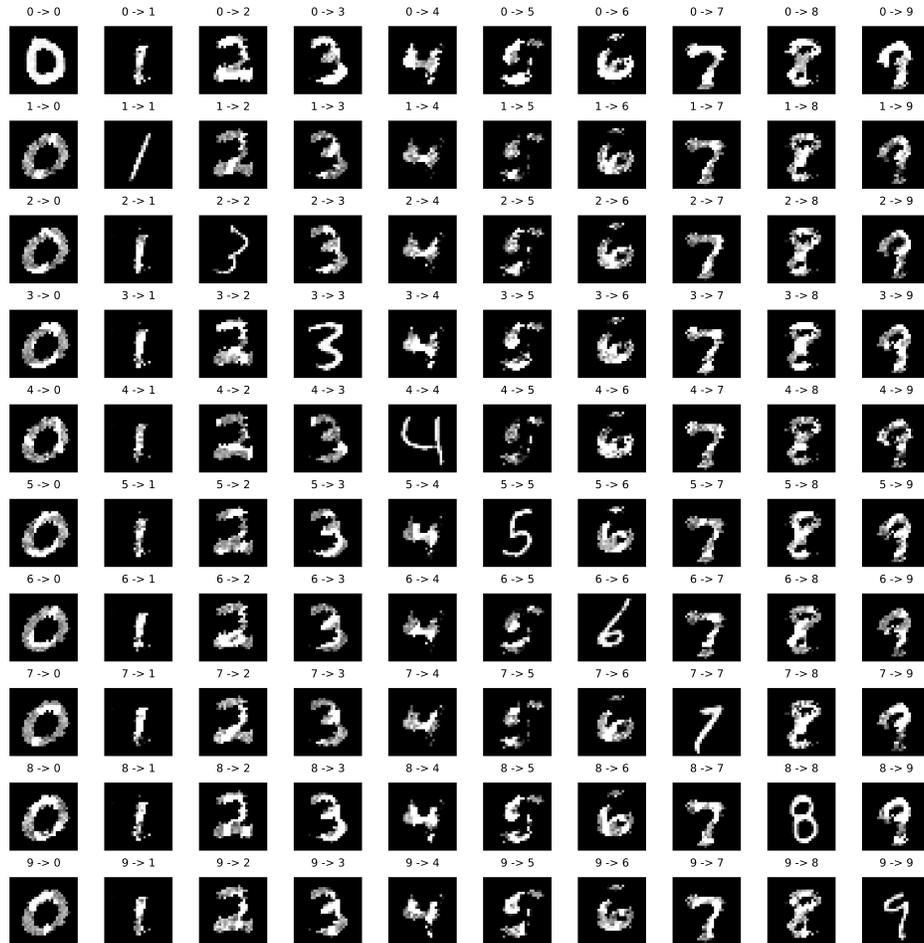


Figure 6: Counterfactuals for *MNIST* data and our *JEM Ensemble*. Original images are shown on the diagonal with the corresponding counterfactuals plotted across rows.