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 [25] describe best practices for using EBM for generative modelling; Grathwohl et al. [24] 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.

A.1 Training: Joint Energy Models

To train our Joint Energy Models we broadly follow the approach outlined in Grathwohl et al. [24]. 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. [24] for both training and inference: we first draw $\mathbf{y} \sim p(\mathbf{y})$ and then sample $\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{y})$ [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 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:

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

Consistent with Grathwohl et al. [24], 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 [25, 24].

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,

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

where $L_{\text{reg}}(\theta)$ is a Ridge penalty (L2 norm) that regularises energy magnitudes for both observed and generated samples [25]. 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

Table 3: EBM hyperparemeter 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

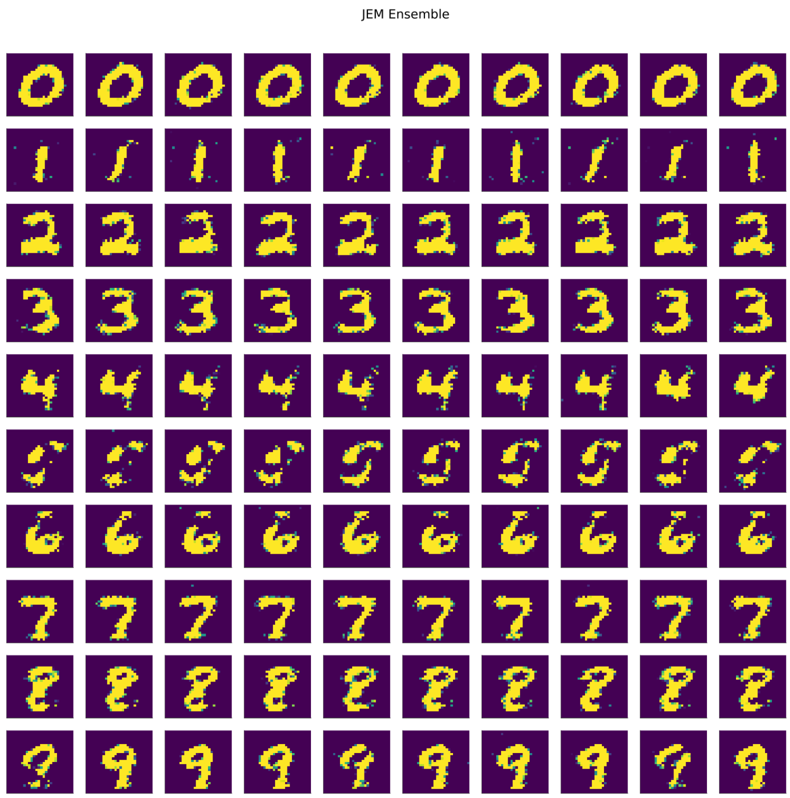

Figure 3: Conditionally generated *MNIST* images for our JEM Ensemble.

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.

A.2 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 tab 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.

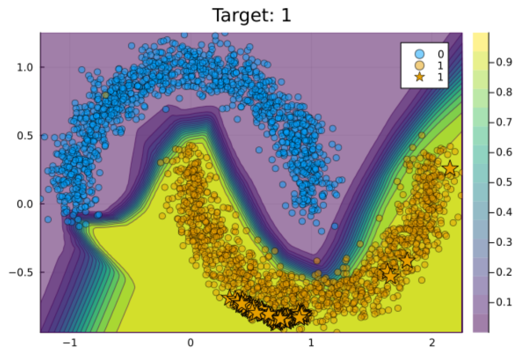

Figure 4: Conditionally generated samples (stars) for our *Moons* data using a JEM.

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.

B.1 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 $\mathcal{D}_{\text{cal}}$. 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 $\mathbf{y}_i$ [28].

Finally, classification sets are formed as follows,

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

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 $\mathbf{x}_{\text{test}}$ does not cover the true test label $\mathbf{y}_{\text{test}}$ approaches α [28].

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.

B.2 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. [30] 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

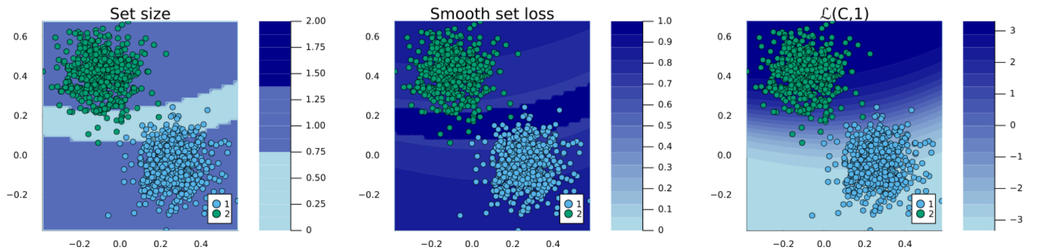

Figure 5: Prediction set size (left), smooth set size loss (centre) and configurable classification loss (right) for a JEM trained on our *Linearly Separable* data.

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 $C_{\theta,\mathbf{y}}(\mathbf{x}_i;\alpha) := \sigma\left((s(\mathbf{x}_i,\mathbf{y})-\alpha)T^{-1}\right)$ for $\mathbf{y} \in \mathcal{Y}$, where σ is the sigmoid function and T is a hyper-parameter used for temperature scaling [30].

In addition to the smooth set size penalty, Stutz et al. [30] 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 ECCCo

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

C.1 A Note on Convergence

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

C.2 `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

Table 4: Paremeter choices for our experiments.

		Network Architecture				Training	
Dataset	Sample Size	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.

		Performance Metrics		
Dataset	Model	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

following common practice in the EBM literature. All of our models were trained through mini-batch training using the Adam optimiser (Kingma and Ba [37]). 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 λ_1, λ_2, λ_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.

D.1 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

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

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:

- 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.

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

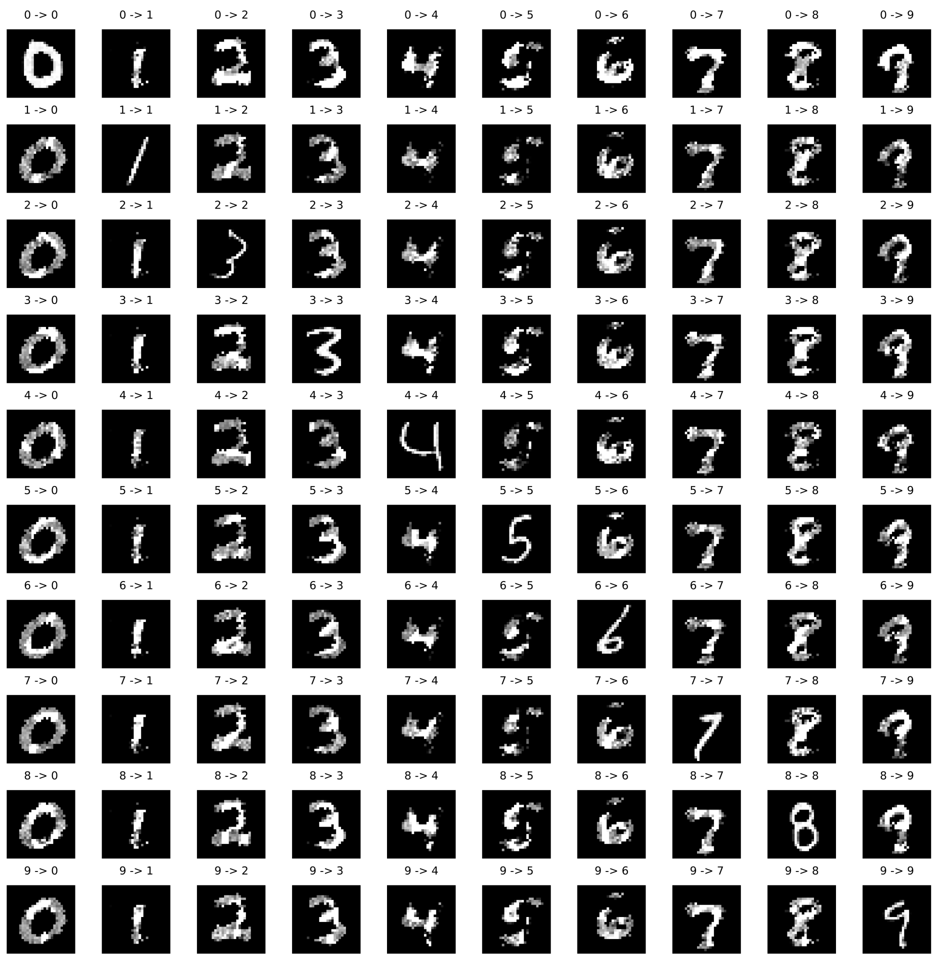

Figure 6: Counterfactuals for *MNIST* data and our *JEM Ensemble*. Original images are shown on the diagonal with the corresponding counterfactuals plotted across rows.