

---

# COUNTERFACTUAL TRAINING: TEACHING MODELS PLAUSIBLE AND ACTIONABLE EXPLANATIONS

---

A PREPRINT

**Patrick Altmeyer** 

Faculty of Electrical Engineering, Mathematics and Computer Science  
Delft University of Technology

[p.altmeyer@tudelft.nl](mailto:p.altmeyer@tudelft.nl)

**Arie van Deursen**

Faculty of Electrical Engineering, Mathematics and Computer Science  
Delft University of Technology

**Cynthia C. S. Liem**

Faculty of Electrical Engineering, Mathematics and Computer Science  
Delft University of Technology

February 19, 2025

## ABSTRACT

Counterfactual Explanations have emerged as a popular tool to explain predictions made by opaque machine learning models: they explain how factual inputs need to change in order for some fitted model to produce some desired output. Much existing research has focused on identifying explanations that are not only valid but also deemed plausible and desirable with respect to the underlying data and stakeholder requirements. Recent work has shown that under this premise, the task of learning plausible explanations is effectively reassigned from the model itself to the (post-hoc) counterfactual explainer. Building on that work, we propose a novel model objective that leverages counterfactuals during the training phase (ad-hoc) in order to minimize the divergence between learned representations and plausible explanations. Through extensive experiments, we demonstrate that our proposed methodology facilitates training models that inherently deliver plausible explanations while maintaining high predictive performance.

**Keywords** Counterfactual Explanations • Explainable AI • Representation Learning

## 1 Introduction

Today's prominence of artificial intelligence (AI) has largely been driven by advances in **representation learning**: instead of relying on features and rules that are carefully hand-crafted by humans, modern machine learning (ML) models are tasked with learning these representations from scratch, guided by narrow objectives such as predictive accuracy (I. Goodfellow, Bengio, and Courville 2016). Modern advances in computing have made it possible to provide such models with ever greater degrees of freedom to achieve that task, which has often led them to outperform traditionally more parsimonious models. Unfortunately, in doing so they also learn increasingly complex and highly sensitive representations that we can no longer easily interpret.

This trend towards complexity for the sake of performance has come under serious scrutiny in recent years. At the very cusp of the deep learning revolution, Szegedy et al. (2013) showed that artificial neural networks (ANN) are sensitive

23 to adversarial examples: counterfactuals of model inputs that yield vastly different model predictions despite being  
 24 “imperceptible” in that they are semantically indifferent from their factual counterparts. Despite partially effective  
 25 mitigation strategies such as **adversarial training** (I. J. Goodfellow, Shlens, and Szegedy 2014), truly robust deep  
 26 learning (DL) remains unattainable even for models that are considered shallow by today’s standards (Kolter 2023).

27 Part of the problem is that high degrees of freedom provide room for many solutions that are locally optimal with  
 28 respect to narrow objectives (Wilson 2020)<sup>1</sup>. Based purely on predictive performance, these solutions may seem to  
 29 provide compelling explanations for the data, when in fact they are based on purely associative, semantically mean-  
 30 ingless patterns. This poses two related challenges: firstly, it makes these models inherently opaque, since humans  
 31 cannot simply interpret what type of explanation the complex learned representations correspond to; secondly, even  
 32 if we could resolve the first challenge, it is not obvious how to mitigate models from learning representations that  
 33 correspond to meaningless and implausible explanations.

34 The first challenge has attracted an abundance of research on **explainable AI** (XAI) which aims to develop tools to  
 35 derive explanations from complex model representations. This can mitigate a scenario in which we deploy opaque  
 36 models and blindly rely on their predictions. On countless occasions, this scenario has already occurred in practice  
 37 and caused real harm to people who were affected adversely and often unfairly by automated decision-making systems  
 38 (ADMS) involving opaque models (O’Neil 2016). Effective XAI tools can aide us in monitoring models and providing  
 39 recourse to individuals to turn adverse outcomes (e.g. “loan application rejected”) into positive ones (“application  
 40 accepted”). Wachter, Mittelstadt, and Russell (2017) propose **counterfactual explanations** as an effective approach  
 41 to achieve this: they explain how factual inputs need to change in order for some fitted model to produce some desired  
 42 output, typically involving minimal perturbations.

43 To our surprise, the second challenge has not yet attracted any consolidated research effort. Specifically, there has  
 44 been no concerted effort towards improving model **explainability**, which we define here as the degree to which learned  
 45 representations correspond to explanations that are interpretable and deemed **plausible** by humans (see Definition 3.1).  
 46 Instead, the choice has typically been to improve the capacity of XAI tools to identify the subset explanations that are  
 47 both plausible and valid for any given model, independent of whether the learned representations are also compatible  
 48 with implausible explanations (Altmeyer et al. 2024). Fortunately, recent findings indicate that explainability can arise  
 49 as byproduct of regularization techniques aimed at other objectives such as robustness, generalization and generative  
 50 capacity Altmeyer et al. (2024).

51 Building on these findings, we introduce **counterfactual training**: a novel regularization technique geared explicitly  
 52 towards aligning model representations with plausible explanations. Our contributions are as follows:

- 53 • We discuss existing related work on improving models and consolidate it through the lens of counterfactual  
 54 explanations (Section 2).
- 55 • We present our proposed methodological framework that leverages faithful counterfactual explanations during  
 56 the training phase of models to achieve the explainability objective (Section 3).
- 57 • Through extensive experiments we demonstrate the counterfactual training improve model explainability  
 58 while maintaining high predictive performance. We run ablation studies and grid searches to understand  
 59 how the underlying model components and hyperparameters affect outcomes. (Section 4).

60 Despite limitations of our approach discussed in Section 5, we conclude that counterfactual training provides a practi-  
 61 cal framework for researchers and practitioners interested in making opaque models more trustworthy Section 6. We  
 62 also believe that this work serves as an opportunity for XAI researchers to reevaluate the premise of improving XAI  
 63 tools without improving models.

## 64 2 Related Literature

65 To the best of our knowledge, our proposed framework for counterfactual training represents the first attempt to use  
 66 counterfactual explanations during training to improve model explainability. In high-level terms, we define model  
 67 explainability as the extent to which valid explanations derived for an opaque model are also deemed plausible with  
 68 respect to the underlying data and stakeholder requirements. To make this more concrete, we follow Augustin, Meinke,  
 69 and Hein (2020) in tying the concept of explainability to the quality of counterfactual explanations that we can  
 70 generate for a given model. The authors show that counterfactual explanations—understood here as minimal input  
 71 perturbations that yield some desired model prediction—are generally more meaningful if the underlying model is  
 72 more robust to adversarial examples. We can make intuitive sense of this finding when looking at adversarial training  
 73 (AT) through the lens of representation learning with high degrees of freedom: by inducing models to “unlearn”

---

<sup>1</sup>For clarity: we follow standard ML convention in using “degrees of freedom” to refer to the number of parameters estimated from data.

74 representations that are susceptible to worst-case counterfactuals (i.e. adversarial examples), AT effectively removes  
 75 some implausible explanations from the solution space.

## 76 2.1 Adversarial Examples are Counterfactual Explanations

77 This interpretation of the link between explainability through counterfactuals on one side, and robustness to adversarial  
 78 examples on the other, is backed by empirical evidence. Sauer and Geiger (2021) demonstrate that using counterfactual  
 79 images during classifier training improves model robustness. Similarly, Abbasnejad et al. (2020) argue that counter-  
 80 factuals represent potentially useful training data in machine learning, especially in supervised settings where inputs  
 81 may be reasonably mapped to multiple outputs. They, too, demonstrate the augmenting the training data of image  
 82 classifiers can improve generalization. Teney, Abbasnejad, and Hengel (2020) propose an approach using counterfac-  
 83 tuals in training that does not rely on data augmentation: they argue that counterfactual pairs typically already exist in  
 84 training datasets. Specifically, their approach relies on, firstly, identifying similar input samples with different annota-  
 85 tions and, secondly, ensuring that the gradient of the classifier aligns with the vector between pairs of counterfactual  
 86 inputs using the cosine distance as a loss function.

87 In the natural language processing (NLP) domain, counterfactuals have similarly been used to improve models through  
 88 data augmentation: Wu et al. (2021), propose *Polyjuice*, a general-purpose counterfactual generator for language mod-  
 89 els. They demonstrate empirically that augmenting training data through *Polyjuice* counterfactuals improves robust-  
 90 ness in a number of NLP tasks. Balashankar et al. (2023) also use *Polyjuice* to augment NLP datasets through diverse  
 91 counterfactuals and show that classifier robustness improves up to 20%. Finally, Luu and Inoue (2023) introduce  
 92 Counterfactual Adversarial Training (CAT), which also aims at improving generalization and robustness of language  
 93 models. Specifically, they propose to proceed as follows: firstly, they identify training samples that are subject to  
 94 high predictive uncertainty; secondly, they generate counterfactual explanations for those samples; and, finally, they  
 95 fine-tune the given language model on the augmented dataset that includes the generated counterfactuals.

96 There have also been several attempts at formalizing the relationship between counterfactual explanations (CE) and  
 97 adversarial examples (AE). Pointing to clear similarities in how CE and AE are generated, Freiesleben (2022) makes  
 98 the case for jointly studying the opaqueness and robustness problem in representation learning. Formally, AE can  
 99 be seen as the subset of CE, for which misclassification is achieved (Freiesleben 2022). Similarly, Pawelczyk et  
 100 al. (2022) show that CE and AE are equivalent under certain conditions and derive theoretical upper bounds on the  
 101 distances between them.

102 Two recent works are closely related to ours in that they use counterfactuals during training with the explicit goal  
 103 of affecting certain properties of post-hoc counterfactual explanations. Firstly, Ross, Lakkaraju, and Bastani (2024)  
 104 propose a way to train models that are guaranteed to provide recourse for individuals to move from an adverse outcome  
 105 to some positive target class with high probability. The approach proposed by Ross, Lakkaraju, and Bastani (2024)  
 106 builds on adversarial training, where in this context susceptibility to targeted adversarial examples for the positive  
 107 class is explicitly induced. The proposed method allows for imposing a set of actionability constraints ex-ante: for  
 108 example, users can specify that certain features (e.g. *age*, *gender*, ...) are immutable. Secondly, Guo, Nguyen, and  
 109 Yadav (2023) are the first to propose an end-to-end training pipeline that includes counterfactual explanations as part  
 110 of the training procedure. In particular, they propose a specific network architecture that includes a predictor and CE  
 111 generator network, where the parameters of the CE generator network are learnable. Counterfactuals are generated  
 112 during each training iteration and fed back to the predictor network. In contrast to Guo, Nguyen, and Yadav (2023),  
 113 we impose no restrictions on the neural network architecture at all.

## 114 2.2 Beyond Robustness

115 Improving the adversarial robustness of models is not the only path towards aligning representations with plausible  
 116 explanations. In a work closely related to this one, Altmeyer et al. (2024) show that explainability can be improved  
 117 through model averaging and refined model objectives. The authors propose a way to generate counterfactuals that  
 118 are maximally **faithful** to the model in that they are consistent with what the model has learned about the underlying  
 119 data. Formally, they rely on tools from energy-based modelling to minimize the divergence between the distribution  
 120 of counterfactuals and the conditional posterior over inputs learned by the model. Their proposed counterfactual  
 121 explainer, *ECCCo*, yields plausible explanations if and only if the underlying model has learned representations that  
 122 align with them. They find that both deep ensembles (Lakshminarayanan, Pritzel, and Blundell 2017) and joint energy-  
 123 based models (JEMs) (Grathwohl et al. 2020) tend to do well in this regard.

124 Once again it helps to look at these findings through the lens of representation learning with high degrees of freedom.  
 125 Deep ensembles are approximate Bayesian model averages, which are most called for when models are underspecified  
 126 by the available data (Wilson 2020). Averaging across solutions mitigates the aforementioned risk of relying on a  
 127 single locally optimal representations that corresponds to semantically meaningless explanations for the data. Previous  
 128 work by Schut et al. (2021) similarly found that generating plausible (“interpretable”) counterfactual explanations is

129 almost trivial for deep ensembles that have also undergone adversarial training. The case for JEMs is even clearer:  
 130 they involve a hybrid objective that induces both high predictive performance and generative capacity (Grathwohl et al.  
 131 2020). This is closely related to the idea of aligning models with plausible explanations and has inspired our proposed  
 132 counterfactual training objective, as we explain in Section 3.

### 133 3 Counterfactual Training

134 Counterfactual training combines ideas from adversarial training, energy-based modelling and counterfactuals expla-  
 135 nations with the explicit objective of aligning representations with plausible explanations that comply with user re-  
 136 quirements. In the context of CE, plausibility has broadly been defined as the degree to which counterfactuals comply  
 137 with the underlying data generating process (Poyiadzi et al. 2020; Guidotti 2022; Altmeyer et al. 2024). Plausibility  
 138 is a necessary but insufficient condition for using CE to provide algorithmic recourse (AR) to individuals affected by  
 139 opaque models in practice. This is because for recourse recommendations to be **actionable**, they need to not only  
 140 result in plausible counterfactuals but also be attainable. A plausible CE for a rejected 20-year-old loan applicant, for  
 141 example, might reveal that their application would have been accepted, if only they were 20 years older. Ignoring all  
 142 other features, this complies with the definition of plausibility if 40-year-old individuals are in fact more credit-worthy  
 143 on average than young adults. But of course this CE does not qualify for providing actionable recourse to the applicant  
 144 since *age* is not a mutable feature. For our intents and purposes, counterfactual training aims at improving model ex-  
 145 plainability by aligning models with counterfactuals that meet both desiderata, plausibility and actionability. Formally,  
 146 we define explainability as follows:

147 **Definition 3.1** (Model Explainability). Let  $\mathbf{M}_\theta : \mathcal{X} \mapsto \mathcal{Y}$  denote a supervised classification model that maps from the  
 148  $D$ -dimensional input space  $\mathcal{X}$  to representations  $\phi(\mathbf{x}; \theta)$  and finally to the  $K$ -dimensional output space  $\mathcal{Y}$ . Assume  
 149 that for any given input-output pair  $\{\mathbf{x}, \mathbf{y}\}_i$  there exists a counterfactual  $\mathbf{x}' = \mathbf{x} + \Delta : \mathbf{M}_\theta(\mathbf{x}') = \mathbf{y}^+ \neq \mathbf{y} = \mathbf{M}_\theta(\mathbf{x})$   
 150 where  $\arg \max_y \mathbf{y}^+ = y^+$  and  $y^+$  denotes the index of the target class.

151 We say that  $\mathbf{M}_\theta$  is **explainable** to the extent that faithfully generated counterfactuals are plausible (i.e. consistent with  
 152 the data) and actionable. Formally, we define these properties as follows:

- 153 1. (Plausibility)  $\int^A p(\mathbf{x}'|\mathbf{y}^+) d\mathbf{x} \rightarrow 1$  where  $A$  is some small region around  $\mathbf{x}'$ .  
 154 2. (Actionability) Permutations  $\Delta$  are subject to actionability constraints.

155 We consider counterfactuals as faithful to the extent that they are consistent with what the model has learned about the  
 156 input data. Let  $p_\theta(\mathbf{x}|\mathbf{y}^+)$  denote the conditional posterior over inputs, then formally:

- 157 3. (Faithfulness)  $\int^A p_\theta(\mathbf{x}'|\mathbf{y}^+) d\mathbf{x} \rightarrow 1$  where  $A$  is defined as above.

158 The definitions of faithfulness and plausibility in Definition 3.1 are the same as in Altmeyer et al. (2024), with adapted  
 159 notation. Actionability constraints in Definition 3.1 vary and depend on the context in which  $\mathbf{M}_\theta$  is deployed. In this  
 160 work, we focus on domain and mutability constraints for individual features  $x_d$  for  $d = 1, \dots, D$ . We limit ourselves  
 161 to classification tasks for reasons discussed in Section 5.

#### 162 3.1 Our Proposed Objective

163 Let  $\mathbf{x}'_t$  for  $t = 0, \dots, T$  denote a counterfactual explanation generated through gradient descent over  $T$  iterations  
 164 as initially proposed by Wachter, Mittelstadt, and Russell (2017). For our purposes, we let  $T$  vary and consider the  
 165 counterfactual search as converged as soon as the predicted probability for the target class has reached a pre-determined  
 166 threshold,  $\tau: \mathcal{S}(\mathbf{M}_\theta(\mathbf{x}'))[y^+] \geq \tau^2$ .

167 To train models with high explainability as defined in Definition 3.1, we propose to leverage counterfactuals in the  
 168 following objective,

$$\min_{\theta} \text{yloss}(\mathbf{M}_\theta(\mathbf{x}), \mathbf{y}) + \lambda_{\text{div}} \text{div}(\mathbf{x}, \mathbf{x}'_T, y; \theta) + \lambda_{\text{adv}} \text{advloss}(\mathbf{M}_\theta(\mathbf{x}'_{t \leq T}), \mathbf{y}) \quad (1)$$

169 where  $\text{yloss}(\cdot)$  is any conventional classification loss that induces discriminative performance (e.g. cross-entropy).  
 170 The two additional components in Equation 1 are explained in more detail below. For now, they can be sufficiently de-  
 171 scribed as inducing explainability directly and indirectly by penalizing: 1) the contrastive divergence,  $\text{div}(\cdot)$ , between  
 172 counterfactuals  $\mathbf{x}'_T$  and observed samples  $x$  and, 2) the adversarial loss,  $\text{advloss}(\cdot)$ , with respect to nascent counterfac-  
 173 tuals  $\mathbf{x}'_{t \leq T}$ . The tradeoff between the different components can be governed by adjusting the strengths of the penalties  
 174  $\lambda_{\text{div}}$  and  $\lambda_{\text{adv}}$ .

---

<sup>2</sup>For detailed background information on gradient-based counterfactual search and convergence see Section B.1.

175 **3.1.1 Directly Inducing Explainability through Contrastive Divergence**

176 Grathwohl et al. (2020) observe that any classifier can be re-interpreted as a joint energy-based model (JEM)  
 177 that learns to discriminate output classes conditional on inputs and generate inputs. They show that JEMs can be  
 178 trained to perform well at both tasks by directly maximizing the joint log-likelihood factorized as  $\log p_\theta(\mathbf{x}, \mathbf{y}) =$   
 179  $\log p_\theta(\mathbf{y}|\mathbf{x}) + \log p_\theta(\mathbf{x})$ . The first factor can be optimized using conventional cross-entropy as in Equation 1. To  
 180 optimize  $\log p_\theta(\mathbf{x})$  Grathwohl et al. (2020) minimize the contrastive divergence between samples drawn from  $p_\theta(\mathbf{x})$   
 181 and training observations, i.e. samples from  $p(\mathbf{x})$ .

182 A key empirical finding in Altmeyer et al. (2024) was that JEMs tend to do well with respect to the plausibility objec-  
 183 tive in Definition 3.1. If we consider samples drawn from  $p_\theta(\mathbf{x})$  as counterfactuals, this is an expected finding, because  
 184 the JEM objective effectively minimizes the divergence between the conditional posterior and  $p(\mathbf{x}|y^+)$ . To generate  
 185 samples, Grathwohl et al. (2020) rely on Stochastic Gradient Langevin Dynamics (SGLD) using an uninformative  
 186 prior for initialization. This is where we depart from their methodology: instead of generating samples through SGLD,  
 187 we propose using counterfactual explainers to generate counterfactuals for observed training samples. Specifically, we  
 188 have

$$\text{div}(\mathbf{x}, \mathbf{x}'_T, y; \theta) = \mathcal{E}_\theta(\mathbf{x}, y) - \mathcal{E}_\theta(\mathbf{x}'_T, y) \quad (2)$$

189 where  $\mathcal{E}_\theta(\cdot)$  denotes the energy function. In particular, we set  $\mathcal{E}_\theta(\mathbf{x}, y) = -\mathbf{M}_\theta(\mathbf{x})[y^+]$  where  $y^+$  denotes the index of  
 190 the target class. We generate samples  $\mathbf{x}'_T$  by first randomly sampling the target class  $y^+ \sim p(y)$  and then generating  
 191 a counterfactual explanation for that target over  $T$  iterations using a gradient-based counterfactual generator. This is  
 192 similar to how conditional sampling is used to draw from  $p_\theta(\mathbf{x})$  in Grathwohl et al. (2020).

193 Intuitively, the gradient of Equation 2 decreases the energy of observed training samples (positive samples) while at  
 194 same time increasing the energy of counterfactuals (negative samples) (Du and Mordatch 2020). As the generated  
 195 counterfactuals get more plausible (Definition 3.1) over the cause of training, these two opposing effects gradually  
 196 balance each out (Lippe 2024).

197 The departure from SGLD allows us to tap into the vast repertoire of explainers that have been proposed in the literature  
 198 to meet different desiderata. Typically, these methods facilitate the imposition of domain and mutability constraints,  
 199 for example. In principle, any existing approach for generating counterfactual explanations is viable, so long as it does  
 200 not violate the faithfulness condition. Like JEMs (Murphy 2022), counterfactual training can be considered as a form  
 201 of contrastive representation learning.

202 **3.1.2 Indirectly Inducing Explainability through Adversarial Robustness**

203 Based on our analysis in Section 2, counterfactuals  $\mathbf{x}'$  can be repurposed as additional training samples (Luu and Inoue  
 204 2023; Balashankar et al. 2023) or adversarial examples (Freiesleben 2022; Pawelczyk et al. 2022). This leaves some  
 205 flexibility with respect to the exact choice for  $\text{advloss}(\cdot)$  in Equation 1. An intuitive functional form to use, though  
 206 likely not the only reasonable choice, is inspired by adversarial training:

$$\begin{aligned} \text{advloss}(\mathbf{M}_\theta(\mathbf{x}'_{t \leq T}), \mathbf{y}; \varepsilon) &= \text{yloss}(\mathbf{M}_\theta(\mathbf{x}'_{t_\varepsilon}), \mathbf{y}) \\ t_\varepsilon &= \max_t \{t : \|\Delta_t\|_\infty < \varepsilon\} \end{aligned} \quad (3)$$

207 Under this choice, we consider nascent counterfactuals  $\mathbf{x}'_{t \leq T}$  as adversarial examples as long as the magnitude of the  
 208 perturbation to any individual feature is at most  $\varepsilon$ . This is closely aligned with Szegedy et al. (2013), who define an  
 209 adversarial attack as an “imperceptible non-random perturbation”. Thus, we choose to work with a different distinction  
 210 between CE and AE than Freiesleben (2022), who considers misclassification as the key distinguishing feature of AE.  
 211 One of the key observations in this work is that we can leverage counterfactual explanations during training and get  
 212 adversarial examples, essentially for free.

213 **3.2 Encoding Actionability Constraints**

214 Many existing counterfactual explainers support domain and mutability constraints out-of-the-box. In fact, both types  
 215 of constraints can be implemented for any counterfactual explainer that relies on gradient descent in the feature space  
 216 for optimization (Altmeyer, Deursen, et al. 2023). In this context, domain constraints can be imposed by simply  
 217 projecting counterfactuals back to the specified domain, if the previous gradient step resulted in updated feature values  
 218 that were out-of-domain. Mutability constraints can similarly be enforced by setting partial derivatives to zero to  
 219 ensure that features are only mutated in the allowed direction, if at all.

220 Since actionability constraints are binding at test time, we should also impose them when generating  $\mathbf{x}'$  during each  
 221 training iteration to align model representations with user requirements. Through their effect on  $\mathbf{x}'$ , both types of

222 constraints influence model outcomes through Equation 2. Here it is crucial that we avoid penalizing implausibility  
 223 that arises due to mutability constraints. For any mutability-constrained feature  $d$  this can be achieved by enforcing  
 224  $\mathbf{x}[d] - \mathbf{x}'[d] := 0$  whenever perturbing  $\mathbf{x}'[d]$  in the direction of  $\mathbf{x}[d]$  would violate mutability constraints. Specifically,  
 225 we set  $\mathbf{x}[d] := \mathbf{x}'[d]$  if

- 226 1. Feature  $d$  is strictly immutable in practice.
- 227 2. We have  $\mathbf{x}[d] > \mathbf{x}'[d]$  but feature  $d$  can only be decreased in practice.
- 228 3. We have  $\mathbf{x}[d] < \mathbf{x}'[d]$  but feature  $d$  can only be increased in practice.

229 From a Bayesian perspective, setting  $\mathbf{x}[d] := \mathbf{x}'[d]$  can be understood as assuming a point mass prior for  $p(\mathbf{x})$  with  
 230 respect to feature  $d$ . Intuitively, we think of this simply in terms ignoring implausibility costs with respect to immutable  
 231 features, which effectively forces the model to instead seek plausibility with respect to the remaining features. This  
 232 in turn results in lower overall sensitivity to immutable features, which we demonstrate empirically for different  
 233 classifiers in Section 4. Under certain conditions, this results holds theoretically[For the proof, see the supplementary  
 234 appendix.]:

235 **Proposition 3.1** (Protecting Immutable Features). *Let  $f_\theta(\mathbf{x}) = \mathcal{S}(\mathbf{M}_\theta(\mathbf{x})) = \mathcal{S}(\Theta\mathbf{x})$  denote a linear classifier with  
 236 softmax activation  $\mathcal{S}$  (i.e. multinomial logistic regression) where  $y \in \{1, \dots, K\} = \mathcal{K}$  and  $\mathbf{x} \in \mathbb{R}^D$ . If we assume  
 237 multivariate Gaussian class densities with common diagonal covariance matrix  $\Sigma_k = \Sigma$  for all  $k \in \mathcal{K}$ , then protecting  
 238 an immutable feature from the contrastive divergence penalty (Equation 2) will result in lower classifier sensitivity to  
 239 that feature relative to the remaining features, provided that at least one of those is mutable and discriminative.*

240 It is worth highlighting that Proposition 3.1 assumes independence of features. This raises a valid concern about the  
 241 effect of protecting immutable features in the presence of proxy features that remain unprotected. We discuss this  
 242 limitation in Section 5.

### 243 3.3 Illustration

244 To better convey the intuition underlying our proposed method, we illustrate different model outcomes in Example 3.1.

245 **Example 3.1** (Prediction of Consumer Credit Default). Suppose we are interested in predicting the likelihood that  
 246 loan applicants default on their credit. We have access to historical data on previous loan takers comprised of a binary  
 247 outcome variable ( $y \in \{1 = \text{default}, 2 = \text{no default}\}$ ) two input features: 1) the subjects' *age*, which we define as  
 248 immutable, and 2) the subjects' existing level of *debt*, which we define as mutable.

249 We have simulated this scenario using synthetic data with independent features and Gaussian class-conditional densities  
 250 in Figure 1. The four panels in Figure 1 show the outcomes for different training procedures using the same model  
 251 architecture each time (a linear classifier). In each case, we show the linear decision boundary (green) and the training  
 252 data colored according to their ground-truth label: orange points belong to the target class,  $y^+ = 2$ , blue points belong  
 253 to the non-target class,  $y^- = 1$ . Stars indicate counterfactuals in the target class generated at test time using generic  
 254 gradient descent until convergence.

255 In panel (a), we have trained our model conventionally, and we do not impose mutability constraints at test time. The  
 256 generated counterfactuals are all valid, but not plausible: they are clearly distinguishable from the ground-truth data. In  
 257 panel (b), we have trained our model with counterfactual training, once again not imposing mutability constraints at test  
 258 time. We observe that the counterfactuals are clearly plausible, therefore meeting the first objective of Definition 3.1.

259 In panel (c), we have used conventional training again, this time imposing the mutability constraint on *age* at test  
 260 time. Counterfactuals are valid but involve some substantial reductions in *debt* for some individuals, in particular  
 261 very young applicants. By comparison, counterfactual paths are shorter on average in panel (d), where we have used  
 262 counterfactual training and protected immutable features as described in Section 3.2. In particular, we observe that due  
 263 to the classifier's lower sensitivity to *age*, recourse recommendations with respect to *debt* are much more homogenous,  
 264 in that they do not disproportionately punish younger individuals. The counterfactuals are also plausible with respect  
 265 to the mutable feature. Thus, we consider the model in panel (d) as the most explainable according to Definition 3.1.

## 266 4 Experiments

267 In this section, we present experiments that we have conducted in order to answer the following research questions:

268 **Research Question 4.1** (Plausibility). *Does our proposed counterfactual training objective (Equation 1) induce mod-  
 269 els to learn plausible explanations?*

270 **Research Question 4.2** (Actionability). *Does our proposed counterfactual training objective (Equation 1) yield more  
 271 favorable algorithmic recourse outcomes in the presence of actionability constraints?*

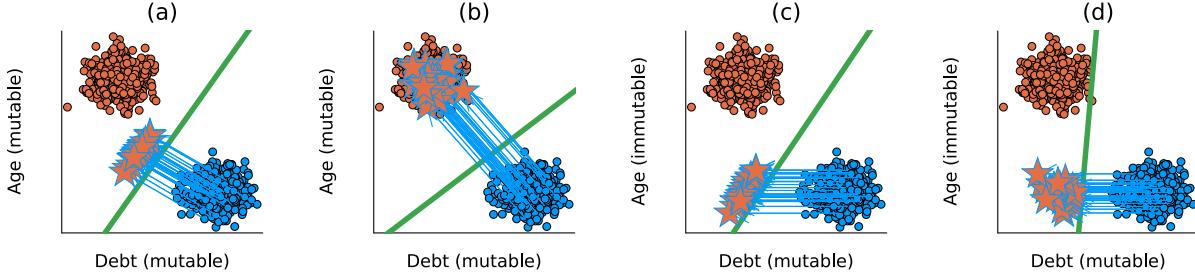


Figure 1: Visual illustration of how counterfactual training improves explainability. See Example 3.1 for details.

272 Beyond this, we are also interested in understanding how robust our answers to RQ 4.1 and RQ 4.2 are:

273 **Research Question 4.3** (Hyperparameters). *What are the effects of different hyperparameter choices with respect to*  
274 *Equation 1?*

#### 275 4.1 Experimental Setup

#### 276 4.2 Experimental Results

### 277 5 Discussion

- 278 1. Limited to classification models.  
279 2. Proxy attributes of immutable features.  
280 3. Increased training time.

### 281 6 Conclusion

## 282 References

- 283 Abbasnejad, Ehsan, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. 2020. “Counterfactual  
284 Vision and Language Learning.” In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition  
285 (CVPR)*, 10041–51. <https://doi.org/10.1109/CVPR42600.2020.01006>.
- 286 Altmeyer, Patrick, Arie van Deursen, et al. 2023. “Explaining Black-Box Models Through Counterfactuals.” In  
287 *Proceedings of the JuliaCon Conferences*, 1:130. 1.
- 288 Altmeyer, Patrick, Mojtaba Farmanbar, Arie van Deursen, and Cynthia CS Liem. 2024. “Faithful Model Explanations  
289 Through Energy-Constrained Conformal Counterfactuals.” In *Proceedings of the AAAI Conference on Artificial  
290 Intelligence*, 38:10829–37. 10.
- 291 Augustin, Maximilian, Alexander Meinke, and Matthias Hein. 2020. “Adversarial Robustness on in-and Out-  
292 Distribution Improves Explainability.” In *European Conference on Computer Vision*, 228–45. Springer.
- 293 Balashankar, Ananth, Xuezhi Wang, Yao Qin, Ben Packer, Nithum Thain, Ed Chi, Jilin Chen, and Alex Beutel. 2023.  
294 “Improving Classifier Robustness Through Active Generative Counterfactual Data Augmentation.” In *Findings of  
295 the Association for Computational Linguistics: EMNLP 2023*, 127–39.
- 296 Du, Yilun, and Igor Mordatch. 2020. “Implicit Generation and Generalization in Energy-Based Models.” <https://arxiv.org/abs/1903.08689>.
- 297 Freiesleben, Timo. 2022. “The Intriguing Relation Between Counterfactual Explanations and Adversarial Examples.”  
298 *Minds and Machines* 32 (1): 77–109.
- 299 Goodfellow, Ian J, Jonathon Shlens, and Christian Szegedy. 2014. “Explaining and Harnessing Adversarial Examples.”  
300 <https://arxiv.org/abs/1412.6572>.
- 301 Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.
- 302 Grathwohl, Will, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swer-  
303 sky. 2020. “Your Classifier Is Secretly an Energy Based Model and You Should Treat It Like One.” In *International  
304 Conference on Learning Representations*.
- 305 Guidotti, Riccardo. 2022. “Counterfactual Explanations and How to Find Them: Literature Review and Benchmark-  
306 ing.” *Data Mining and Knowledge Discovery*, 1–55.
- 307 Guo, Hangzhi, Thanh H. Nguyen, and Amulya Yadav. 2023. “CounterNet: End-to-End Training of Prediction Aware  
308 Counterfactual Explanations.” In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery  
309 and Data Mining*, 577–89. KDD ’23. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/3580305.3599290>.

- 312 Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Springer New  
 313 York. <https://doi.org/10.1007/978-0-387-84858-7>.
- 314 Kolter, Zico. 2023.“Keynote Addresses: SaTML 2023 .” In *2023 IEEE Conference on Secure and Trustworthy  
 315 Machine Learning (SaTML)*, xvi–. Los Alamitos, CA, USA: IEEE Computer Society. [https://doi.org/10.1109/Sa  
 TML54575.2023.00009](https://doi.org/10.1109/Sa<br/>
  316 TML54575.2023.00009).
- 317 Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. 2017. “Simple and Scalable Predictive Uncer-  
 318 tainty Estimation Using Deep Ensembles.” *Advances in Neural Information Processing Systems* 30.
- 319 Lippe, Phillip. 2024. “UvA Deep Learning Tutorials.” <https://uvadlc-notebooks.readthedocs.io/en/latest/>.
- 320 Luu, Hoai Linh, and Naoya Inoue. 2023. “Counterfactual Adversarial Training for Improving Robustness of Pre-  
 321 Trained Language Models.” In *Proceedings of the 37th Pacific Asia Conference on Language, Information and  
 322 Computation*, 881–88.
- 323 Murphy, Kevin P. 2022. *Probabilistic Machine Learning: An Introduction*. MIT Press.
- 324 O’Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*.  
 325 Crown.
- 326 Pawelczyk, Martin, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. 2022. “Exploring  
 327 Counterfactual Explanations Through the Lens of Adversarial Examples: A Theoretical and Empirical Analysis.”  
 328 In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, edited by Gustau  
 329 Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, 151:4574–94. Proceedings of Machine Learning Research.  
 330 PMLR. <https://proceedings.mlr.press/v151/pawelczyk22a.html>.
- 331 Poyiadzi, Rafael, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. 2020. “FACE: Feasible and  
 332 Actionable Counterfactual Explanations.” In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*,  
 333 344–50.
- 334 Ross, Alexis, Himabindu Lakkaraju, and Osbert Bastani. 2024. “Learning Models for Actionable Recourse.” In  
 335 *Proceedings of the 35th International Conference on Neural Information Processing Systems*. NIPS ’21. Red  
 336 Hook, NY, USA: Curran Associates Inc.
- 337 Sauer, Axel, and Andreas Geiger. 2021. “Counterfactual Generative Networks.” <https://arxiv.org/abs/2101.06046>.
- 338 Schut, Lisa, Oscar Key, Rory Mc Grath, Luca Costabello, Bogdan Sacaleanu, Yarin Gal, et al. 2021. “Generating  
 339 Interpretable Counterfactual Explanations By Implicit Minimisation of Epistemic and Aleatoric Uncertainties.” In  
 340 *International Conference on Artificial Intelligence and Statistics*, 1756–64. PMLR.
- 341 Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus.  
 342 2013. “Intriguing Properties of Neural Networks.” <https://arxiv.org/abs/1312.6199>.
- 343 Teney, Damien, Ehsan Abbasnedjad, and Anton van den Hengel. 2020. “Learning What Makes a Difference from  
 344 Counterfactual Examples and Gradient Supervision.” In *Computer Vision–ECCV 2020: 16th European Confer-  
 345 ence, Glasgow, UK, August 23–28, 2020, Proceedings, Part x 16*, 580–99. Springer.
- 346 Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2017. “Counterfactual Explanations Without Opening the Black  
 347 Box: Automated Decisions and the GDPR.” *Harv. JL & Tech.* 31: 841. <https://doi.org/10.2139/ssrn.3063289>.
- 348 Wilson, Andrew Gordon. 2020. “The Case for Bayesian Deep Learning.” <https://arxiv.org/abs/2001.10995>.
- 349 Wu, Tongshuang, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. “Polyjuice: Generating Counterfactuals  
 350 for Explaining, Evaluating, and Improving Models.” In *Proceedings of the 59th Annual Meeting of the Associa-  
 351 tion for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing  
 352 (Volume 1: Long Papers)*, edited by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, 6707–23. Online:  
 353 Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-long.523>.

354 **A Notation**

- 355 •  $y^+$ : The target class and also the index of the target class.
- 356 •  $y^-$ : The non-target class and also the index of non-the target class.
- 357 •  $\mathbf{y}^+$ : The one-hot encoded output vector for the target class.
- 358 •  $\theta$ : Model parameters (unspecified).
- 359 •  $\Theta$ : Matrix of parameters.

360 **B Technical Details of Our Approach**

361 **B.1 Generating Counterfactuals through Gradient Descent**

362 In this section, we provide some background on gradient-based counterfactual generators (Section B.1.1) and discuss  
363 how we define convergence in this context (Section B.1.2).

364 **B.1.1 Background**

365 **B.1.2 Convergence**

366 An important consideration when generating counterfactual explanations using gradient-based methods is how to  
367 define convergence. Two common choices are to 1) perform gradient descent over a fixed number of iterations  $T$ , or  
368 2) conclude the search as soon as the predicted probability for the target class has reached a pre-determined threshold,  
369  $\tau: \mathcal{S}(\mathbf{M}_\theta(\mathbf{x}'))[y^+] \geq \tau$ . We prefer the latter for our purposes, because it explicitly defines convergence in terms of the  
370 black-box model,  $\mathbf{M}(\mathbf{x})$ .

371 Defining convergence in this way allows for a more intuitive interpretation of the resulting counterfactual outcomes  
372 than with fixed  $T$ . Specifically, it allows us to think of counterfactuals as explaining ‘high-confidence’ predictions by  
373 the model for the target class  $y^+$ . Depending on the context and application, different choices of  $\tau$  can be considered  
374 as representing ‘high-confidence’ predictions.

375 **B.2 Protecting Mutability Constraints with Linear Classifiers**

376 In Section 3.2 we explain that to avoid penalizing implausibility that arises due to mutability constraints, we impose a  
377 point mass prior on  $p(\mathbf{x})$  for the corresponding feature. We argue in Section 3.2 that this approach induces models to  
378 be less sensitive to immutable features and demonstrate this empirically in Section 4. Below we derive the analytical  
379 results in Proposition 3.1.

380 *Proof.* Let  $d_{\text{mtbl}}$  and  $d_{\text{immmtbl}}$  denote some mutable and immutable feature, respectively. Suppose that  $\mu_{y^-, d_{\text{immmtbl}}} <$   
381  $\mu_{y^+, d_{\text{immmtbl}}}$  and  $\mu_{y^-, d_{\text{mtbl}}} > \mu_{y^+, d_{\text{mtbl}}}$ , where  $\mu_{k,d}$  denotes the conditional sample mean of feature  $d$  in class  $k$ . In words,  
382 we assume that the immutable feature tends to take lower values for samples in the non-target class  $y^-$  than in the  
383 target class  $y^+$ . We assume the opposite to hold for the mutable feature.

384 Assuming multivariate Gaussian class densities with common diagonal covariance matrix  $\Sigma_k = \Sigma$  for all  $k \in \mathcal{K}$ , we  
385 have for the log likelihood ratio between any two classes  $k, m \in \mathcal{K}$  (Hastie, Tibshirani, and Friedman 2009):

$$\log \frac{p(k|\mathbf{x})}{p(m|\mathbf{x})} = \mathbf{x}^\top \Sigma^{-1} (\mu_k - \mu_m) + \text{const} \quad (4)$$

386 By independence of  $x_1, \dots, x_D$ , the full log-likelihood ratio decomposes into:

$$\log \frac{p(k|\mathbf{x})}{p(m|\mathbf{x})} = \sum_{d=1}^D \frac{\mu_{k,d} - \mu_{m,d}}{\sigma_d^2} x_d + \text{const} \quad (5)$$

387 By the properties of our classifier (*multinomial logistic regression*), we have:

$$\log \frac{p(k|\mathbf{x})}{p(m|\mathbf{x})} = \sum_{d=1}^D (\theta_{k,d} - \theta_{m,d}) x_d + \text{const} \quad (6)$$

388 where  $\theta_{k,d} = \Theta[k, d]$  denotes the coefficient on feature  $d$  for class  $k$ .

389 Based on Equation 5 and Equation 6 we can identify that  $(\mu_{k,d} - \mu_{m,d}) \propto (\theta_{k,d} - \theta_{m,d})$  under the assumptions we  
390 made above. Hence, we have that  $(\theta_{y^-, d_{\text{immmtbl}}} - \theta_{y^+, d_{\text{immmtbl}}}) < 0$  and  $(\theta_{y^-, d_{\text{mtbl}}} - \theta_{y^+, d_{\text{mtbl}}}) > 0$

391 Let  $\mathbf{x}'$  denote some randomly chosen individual from class  $y^-$  and let  $y^+ \sim p(y)$  denote the randomly chosen target  
 392 class. Then the partial derivative of the contrastive divergence penalty Equation 2 with respect to coefficient  $\theta_{y^+, d}$  is  
 393 equal to

$$\frac{\partial}{\partial \theta_{y^+, d}} (\text{div}(\mathbf{x}, \mathbf{x}', \mathbf{y}; \theta)) = \frac{\partial}{\partial \theta_{y^+, d}} ((-\mathbf{M}_\theta(\mathbf{x})[y^+]) - (-\mathbf{M}_\theta(\mathbf{x}')[y^+])) = x'_d - x_d \quad (7)$$

394 and equal to zero everywhere else.

395 Since  $(\mu_{y^-, d_{\text{immtbl}}} < \mu_{y^+, d_{\text{immtbl}}})$  we are more likely to have  $(x'_{d_{\text{immtbl}}} - x_{d_{\text{immtbl}}}) < 0$  than vice versa at initialization.  
 396 Similarly, we are more likely to have  $(x'_{d_{\text{mtbl}}} - x_{d_{\text{mtbl}}}) > 0$  since  $(\mu_{y^-, d_{\text{mtbl}}} > \mu_{y^+, d_{\text{mtbl}}})$ .

397 This implies that if we do not protect feature  $d_{\text{immtbl}}$ , the contrastive divergence penalty will decrease  $\theta_{y^-, d_{\text{immtbl}}}$  thereby  
 398 exacerbating the existing effect  $(\theta_{y^-, d_{\text{immtbl}}} - \theta_{y^+, d_{\text{immtbl}}}) < 0$ . In words, not protecting the immutable feature would have  
 399 the undesirable effect of making the classifier more sensitive to this feature, in that it would be more likely to predict  
 400 class  $y^-$  as opposed to  $y^+$  for lower values of  $d_{\text{immtbl}}$ .

401 By the same rationale, the contrastive divergence penalty can generally be expected to increase  $\theta_{y^-, d_{\text{mtbl}}}$  exacerbating  
 402  $(\theta_{y^-, d_{\text{mtbl}}} - \theta_{y^+, d_{\text{mtbl}}}) > 0$ . In words, this has the effect of making the classifier more sensitive to the mutable feature, in  
 403 that it would be more likely to predict class  $y^-$  as opposed to  $y^+$  for higher values of  $d_{\text{mtbl}}$ .

404 Thus, our proposed approach of protecting feature  $d_{\text{immtbl}}$  has the net affect of decreasing the classifier's sensitivity  
 405 to the immutable feature relative to the mutable feature (i.e. no change in sensitivity for  $d_{\text{immtbl}}$  relative to increased  
 406 sensitivity for  $d_{\text{mtbl}}$ ).  $\square$

### ⚠ Warning

@Cynthia, @Arie, I have tentatively phrased the above in terms of a theorem and proof. This is something I've so far shied away from because I feel a bit out of my depth when it comes to mathematical proofs. The above makes intuitive sense to me, but I don't know for sure if it's correct.

407

## 408 B.3 Domain Constraints

409 We apply domain constraints on counterfactuals during training and evaluation. There are at least two good reasons for  
 410 doing so. Firstly, within the context of explainability and algorithmic recourse, real-world attributes are often domain  
 411 constrained: the *age* feature, for example, is lower bounded by zero and upper bounded by the maximum human  
 412 lifespan. Secondly, domain constraints help mitigate training instabilities commonly associated with energy-based  
 413 modelling (Grathwohl et al. 2020; Altmeyer et al. 2024).

414 For our image datasets, features are pixel values and hence the domain is constrained by the lower and upper bound  
 415 of values that pixels can take depending on how they are scaled (in our case  $[-1, 1]$ ). For all other features  $d$  in our  
 416 synthetic and tabular datasets, we automatically infer domain constraints  $[x_d^{\text{LB}}, x_d^{\text{UB}}]$  as follows,

$$\begin{aligned} x_d^{\text{LB}} &= \arg \min_{x_d} \{\mu_d - n_{\sigma_d} \sigma_d, \arg \min_{x_d} x_d\} \\ x_d^{\text{UB}} &= \arg \max_{x_d} \{\mu_d + n_{\sigma_d} \sigma_d, \arg \max_{x_d} x_d\} \end{aligned} \quad (8)$$

417 where  $\mu_d$  and  $\sigma_d$  denote the sample mean and standard deviation of feature  $d$ . We set  $n_{\sigma_d} = 3$  across the board but  
 418 higher values and hence wider bounds may be appropriate depending on the application.

## 419 C Detailed Results

### ⚠ Warning

@Cynthia, @Arie, I'm including some preliminary results here but will rerun experiments in the coming days. You'll notice some odd outlier results (huge values) for the initial grid search (Section C.2). My belief is that this is once again due to *ECCo* overshooting for some hyperparameter choices, that we have discussed before. A simple solution to this story is to actually impose domain constraints. Let's see what the final results show us

420

(I also still plan to make slight changes to the implementation), but in any case I think it might even be worth to report results for the initial grid search in the final appendix (they do include good results for CT for certain hyperparameter ranges and highlight limitations inherited from energy-based modelling).

421

## 422 C.1 Qualitative Findings for Image Data

**i** Note

@Cynthia, @Arie, Figure A1 shows much more plausible (faithful) counterfactuals for a model with CT than the model with conventional training (Figure A2). In fact, this is not even using *ECCo+* and still showing better results than the best results we achieved in our AAAI paper for JEM ensembles.

423

## 424 C.2 Initial Grid Search

425 For the initial round of experiments we

### 426 C.2.1 Generator Parameters

427 The hyperparameter grids for the first investigation of the effect of generator parameters are shown in Parameters C.1  
428 and Parameters C.2.

#### 429 Parameters C.1 (Training Phase).

- 430 • Generator Parameters:
  - 431 –  $\lambda_{\text{cost}}$ : 0.0, 0.001, 0.1
  - 432 –  $\lambda_{\text{div}}$ : 0.01, 0.05, 0.1, 0.5, 1.0, 5.0, 10.0, 15.0
  - 433 – Learning Rate: 1.0
  - 434 – Maximum Iterations: 20, 50, 100
  - 435 – Optimizerimizer: sgd
- 436 • Generator: ecco, generic, omni, revise
- 437 • Training Parameters:
  - 438 – Objective: full, vanilla

#### 439 Parameters C.2 (Evaluation Phase).

- 440 • Counterfactual Parameters:
  - 441 – Convergence: max\_iter
  - 442 – Maximum Iterations: 100
  - 443 – No. Individuals: 100
  - 444 – No. Runs: 5
- 445 • Generator Parameters:
  - 446 –  $\lambda_{\text{cost}}$ : 0.0
  - 447 –  $\lambda_{\text{div}}$ : 0.1, 0.5, 1.0, 5.0, 10.0, 20.0
  - 448 – Learning Rate: 1.0
  - 449 – Maximum Iterations: 50
  - 450 – Optimizerimizer: sgd

#### 451 C.2.1.1 Linearly Separable

- 452 • **Energy Penalty** (Table A1): *ECCo* generally does yield better results than *Vanilla* for higher choices of the  
453 energy penalty (10,15) during training. *Generic* performs poorly accross the board. *Omni* seems to have an  
454 anchoring effect, in that it never performs terribly but also never as good as the best *ECCo* results. *REVISE*  
455 performs poorly across the board.
- 456 • **Cost** (Table A2): Results for all generators (except *Omni*) are quite bad, which can likely be attributed to  
457 extremely bad results for some choices of the **Energy Penalty** (results here are averaged). For *ECCo* and  
458 *Generic*, higher cost values generally lead to worse results.
- 459 • **Maximum Iterations**: No clear patterns recognizable, so it seems that smaller choices are ok.
- 460 • **Validity**: *ECCo* almost always valid except for very low values during training and high values at evaluation  
461 time. *Generic* often has poor validity.

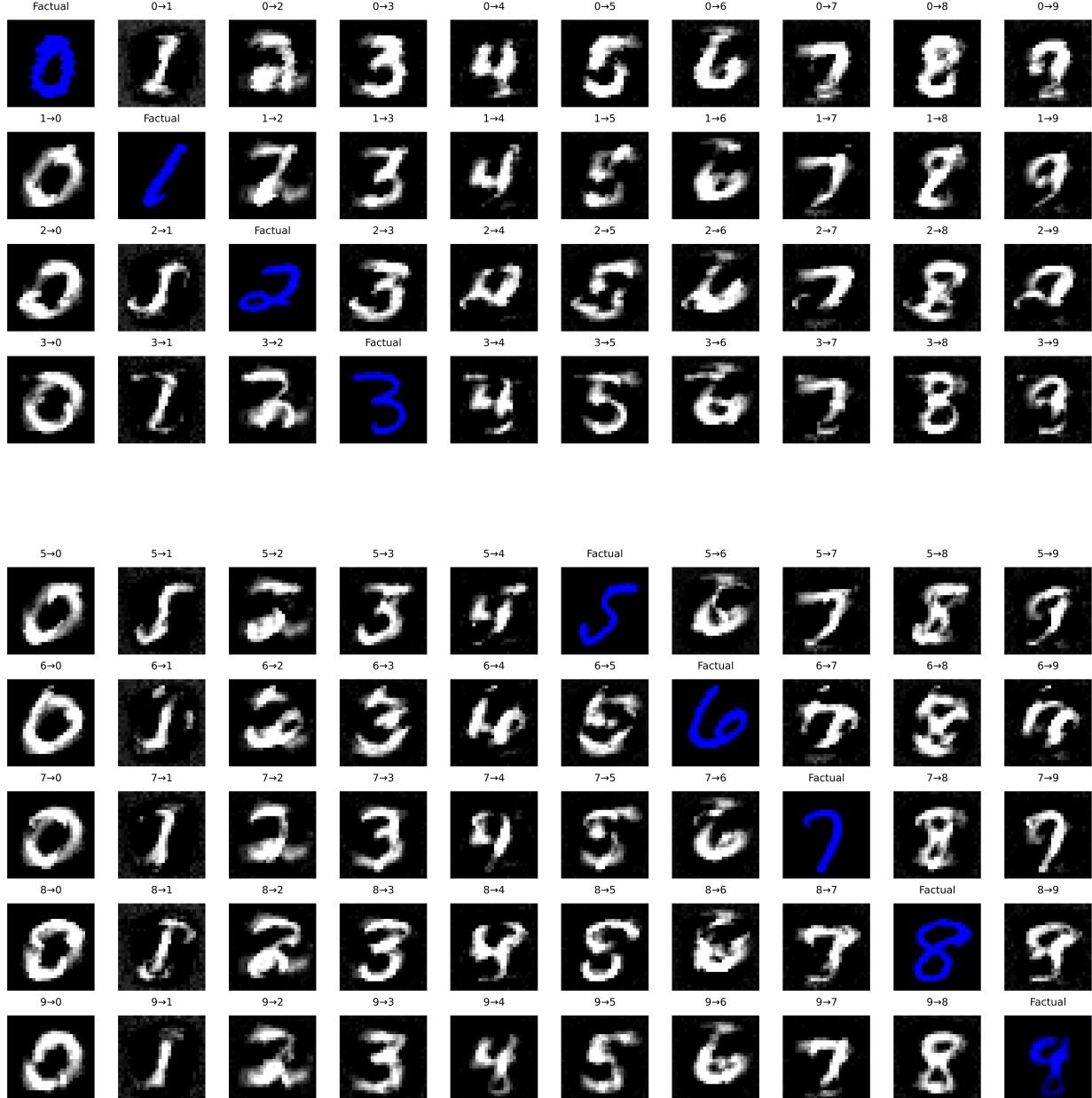


Figure A1: Counterfactual images for *MLP* with counterfactual training. The underlying generator, *ECCo*, aims to generate counterfactuals that are faithful to the model (Altmeyer et al. 2024).

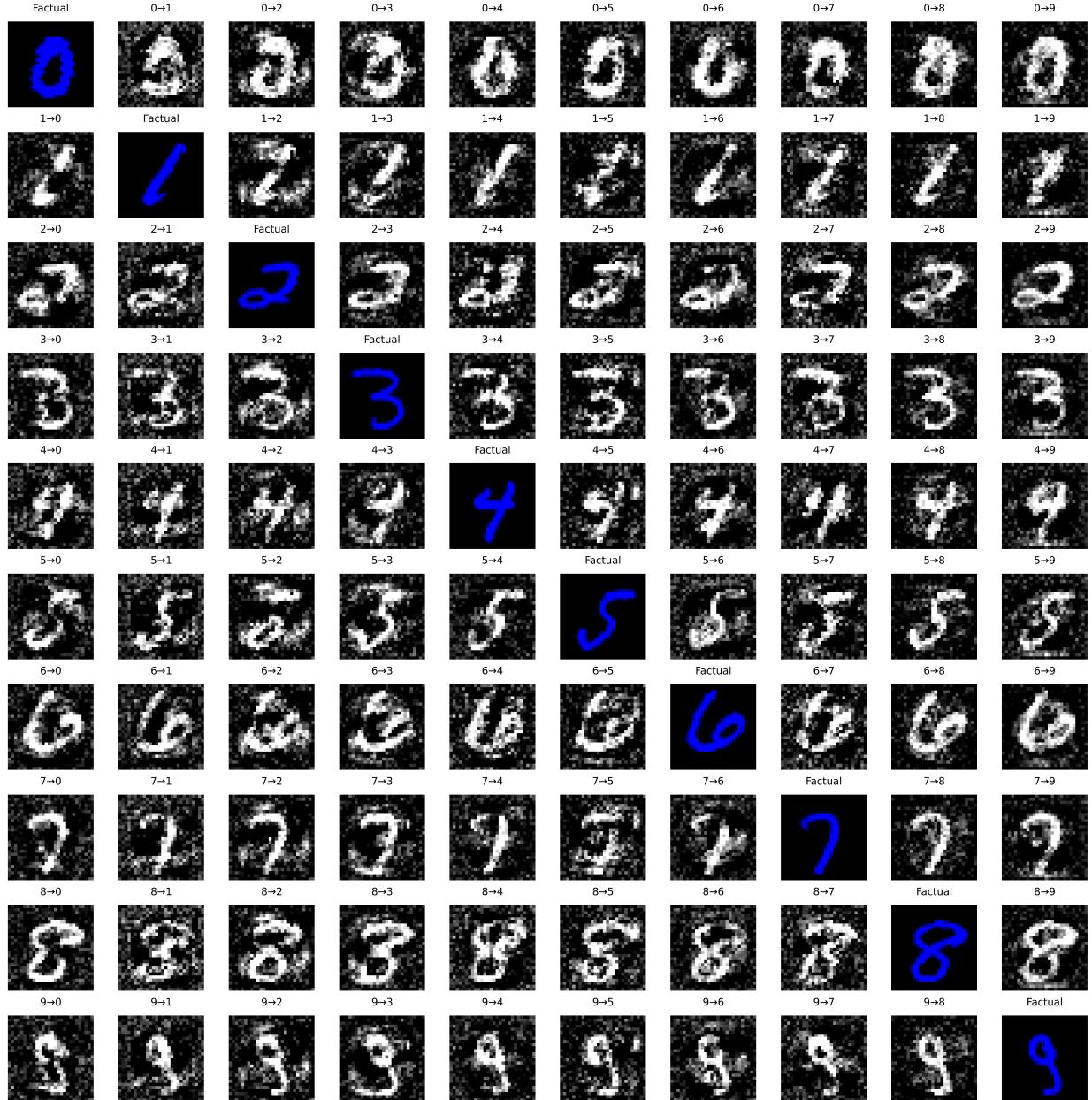


Figure A2: Counterfactual images for *MLP* with conventional training. The underlying generator, *ECCo*, aims to generate counterfactuals that are faithful to the model (Altmeyer et al. 2024).

462

- **Accuracy:** Seems largely unaffected.

Table A1: Results for Linearly Separable data by energy penalty.

Objective	$\lambda_{\text{div}}(\text{train})$	Generator	Value	Std
full	0.01	<i>ECCo</i>	$-9.91 \cdot 10^{11}$	$2.25 \cdot 10^{12}$
full	0.01	<i>Generic</i>	$-5.71 \cdot 10^{17}$	$1.3 \cdot 10^{18}$
<b>full</b>	<b>0.01</b>	<b>Omniscient</b>	<b>-2.54</b>	<b>0.116</b>
full	0.01	<i>REVISE</i>	-15.6	13.2
vanilla	0.01	<i>ECCo</i>	-4.28	3.52
vanilla	0.01	<i>Generic</i>	-4.45	3.47
vanilla	0.01	<i>Omniscient</i>	-5.12	4.46
vanilla	0.01	<i>REVISE</i>	-4.91	4.24
full	0.05	<i>ECCo</i>	$-5.63 \cdot 10^5$	$1.28 \cdot 10^6$
full	0.05	<i>Generic</i>	$-8.35 \cdot 10^{17}$	$1.9 \cdot 10^{18}$
<b>full</b>	<b>0.05</b>	<b>Omniscient</b>	<b>-2.53</b>	<b>0.114</b>
full	0.05	<i>REVISE</i>	-15	12.6
vanilla	0.05	<i>ECCo</i>	-4.4	3.66
vanilla	0.05	<i>Generic</i>	-4.38	3.48
vanilla	0.05	<i>Omniscient</i>	-5.25	4.62
vanilla	0.05	<i>REVISE</i>	-4.94	4.22
full	0.1	<i>ECCo</i>	$-6.74 \cdot 10^5$	$1.53 \cdot 10^6$
full	0.1	<i>Generic</i>	$-1.71 \cdot 10^{11}$	$3.9 \cdot 10^{11}$
<b>full</b>	<b>0.1</b>	<b>Omniscient</b>	<b>-2.56</b>	<b>0.124</b>
full	0.1	<i>REVISE</i>	-15.6	13.2
vanilla	0.1	<i>ECCo</i>	-4.28	3.52
vanilla	0.1	<i>Generic</i>	-4.45	3.48
vanilla	0.1	<i>Omniscient</i>	-5.12	4.46
vanilla	0.1	<i>REVISE</i>	-4.91	4.25
full	0.5	<i>ECCo</i>	-11.8	9.83
full	0.5	<i>Generic</i>	$-1.06 \cdot 10^{18}$	$2.42 \cdot 10^{18}$
<b>full</b>	<b>0.5</b>	<b>Omniscient</b>	<b>-2.54</b>	<b>0.123</b>
full	0.5	<i>REVISE</i>	-15	12.6
vanilla	0.5	<i>ECCo</i>	-4.4	3.65
vanilla	0.5	<i>Generic</i>	-4.38	3.48
vanilla	0.5	<i>Omniscient</i>	-5.25	4.61
vanilla	0.5	<i>REVISE</i>	-4.95	4.22
full	1	<i>ECCo</i>	-11.5	11.1
full	1	<i>Generic</i>	$-1.71 \cdot 10^{11}$	$3.88 \cdot 10^{11}$
<b>full</b>	<b>1</b>	<b>Omniscient</b>	<b>-2.59</b>	<b>0.117</b>
full	1	<i>REVISE</i>	-15.7	13.3
vanilla	1	<i>ECCo</i>	-4.28	3.51
vanilla	1	<i>Generic</i>	-4.44	3.47
vanilla	1	<i>Omniscient</i>	-5.11	4.46
vanilla	1	<i>REVISE</i>	-4.91	4.25
full	5	<i>ECCo</i>	-3.99	3.12
full	5	<i>Generic</i>	$-4.88 \cdot 10^{17}$	$1.11 \cdot 10^{18}$
<b>full</b>	<b>5</b>	<b>Omniscient</b>	<b>-2.53</b>	<b>0.117</b>
full	5	<i>REVISE</i>	-14.6	12.1
vanilla	5	<i>ECCo</i>	-4.4	3.65
vanilla	5	<i>Generic</i>	-4.38	3.48
vanilla	5	<i>Omniscient</i>	-5.25	4.61
vanilla	5	<i>REVISE</i>	-4.95	4.22
<b>full</b>	<b>10</b>	<b>ECCo</b>	<b>-2.31</b>	<b>0.735</b>
full	10	<i>Generic</i>	$-1.7 \cdot 10^{11}$	$3.86 \cdot 10^{11}$
full	10	<i>Omniscient</i>	-2.53	0.117

Continuing table below.

Objective	$\lambda_{\text{div}}(\text{train})$	Generator	Value	Std
full	10	<i>REVISE</i>	-15.5	13
vanilla	10	<i>ECCo</i>	-4.28	3.51
vanilla	10	<i>Generic</i>	-4.44	3.47
vanilla	10	<i>Omniscient</i>	-5.12	4.46
vanilla	10	<i>REVISE</i>	-4.91	4.24
<b>full</b>	<b>15</b>	<b><i>ECCo</i></b>	<b>-2.01</b>	<b>0.488</b>
full	15	<i>Generic</i>	$-4.91 \cdot 10^{17}$	$1.12 \cdot 10^{18}$
full	15	<i>Omniscient</i>	-2.53	0.116
full	15	<i>REVISE</i>	-14.4	11.7
vanilla	15	<i>ECCo</i>	-4.4	3.65
vanilla	15	<i>Generic</i>	-4.38	3.48
vanilla	15	<i>Omniscient</i>	-5.25	4.6
vanilla	15	<i>REVISE</i>	-4.95	4.23

Table A2: Results for Linearly Separable data by cost penalty.

Objective	$\lambda_{\text{cost}}(\text{train})$	Generator	Value	Std
full	0	<i>ECCo</i>	$-5.32 \cdot 10^3$	$1.21 \cdot 10^4$
full	0	<i>Generic</i>	$-1.03 \cdot 10^{18}$	$2.34 \cdot 10^{18}$
<b>full</b>	<b>0</b>	<b><i>Omniscient</i></b>	<b>-2.64</b>	<b>0.125</b>
full	0	<i>REVISE</i>	-15.4	12.9
vanilla	0	<i>ECCo</i>	-4.34	3.58
vanilla	0	<i>Generic</i>	-4.41	3.48
vanilla	0	<i>Omniscient</i>	-5.18	4.54
vanilla	0	<i>REVISE</i>	-4.93	4.23
full	0.001	<i>ECCo</i>	-362	811
full	0.001	<i>Generic</i>	$-2.65 \cdot 10^{17}$	$6.04 \cdot 10^{17}$
<b>full</b>	<b>0.001</b>	<b><i>Omniscient</i></b>	<b>-2.49</b>	<b>0.115</b>
full	0.001	<i>REVISE</i>	-15.5	13
vanilla	0.001	<i>ECCo</i>	-4.34	3.58
vanilla	0.001	<i>Generic</i>	-4.41	3.48
vanilla	0.001	<i>Omniscient</i>	-5.18	4.53
vanilla	0.001	<i>REVISE</i>	-4.93	4.23
full	0.1	<i>ECCo</i>	$-3.72 \cdot 10^{11}$	$8.46 \cdot 10^{11}$
full	0.1	<i>Generic</i>	$-4.48 \cdot 10^{14}$	$1.02 \cdot 10^{15}$
<b>full</b>	<b>0.1</b>	<b><i>Omniscient</i></b>	<b>-2.5</b>	<b>0.112</b>
full	0.1	<i>REVISE</i>	-14.6	12.2
vanilla	0.1	<i>ECCo</i>	-4.34	3.58
vanilla	0.1	<i>Generic</i>	-4.41	3.48
vanilla	0.1	<i>Omniscient</i>	-5.18	4.54
vanilla	0.1	<i>REVISE</i>	-4.93	4.24

## 463 C.2.1.2 Moons

- 464 • **Energy Penalty** (Table A3): *ECCo* consistently yields better results than *Vanilla*, except for very low choices  
465 of the energy penalty during training for which it performs abysmal. *Generic* performs quite badly across  
466 the board for high enough choices of the energy penalty at evaluation time. *Omni* has small positive effect.  
467 *REVISE* performs poorly across the board.
- 468 • **Cost (distance penalty)**: *Generic* generally does better for higher values, while *ECCo* does better for lower  
469 values.
- 470 • **Maximum Iterations**: No clear patterns recognizable, so it seems that smaller choices are ok.
- 471 • **Validity**: *ECCo* generally achieves full validity except for very low choices the energy penalty during training  
472 and high choices at evaluation time. *Generic* performs poorly for high choices of the energy penalty during  
473 evaluation.

474  
475

- **Accuracy:** Largely unaffected although *ECCo* suffers a bit for very low choices the energy penalty during training. *REVISE* suffers a lot in general (around 10 percentage points).

Table A3: Results for Moons data by energy penalty.

Objective	$\lambda_{\text{div}}(\text{train})$	Generator	Value	Std
full	0.01	<i>ECCo</i>	$-2.8 \cdot 10^{22}$	$6.39 \cdot 10^{22}$
full	0.01	<i>Generic</i>	$-4.89 \cdot 10^{30}$	$1.11 \cdot 10^{31}$
<b>full</b>	<b>0.01</b>	<b>Omniscient</b>	<b>-4.74</b>	<b>5.08</b>
full	0.01	<i>REVISE</i>	-572	$1.25 \cdot 10^3$
vanilla	0.01	<i>ECCo</i>	-15.5	17.3
vanilla	0.01	<i>Generic</i>	-10.9	11.9
vanilla	0.01	<i>Omniscient</i>	-12.7	14.4
vanilla	0.01	<i>REVISE</i>	-11.2	13
full	0.05	<i>ECCo</i>	$-1.55 \cdot 10^{16}$	$3.52 \cdot 10^{16}$
full	0.05	<i>Generic</i>	$-2.22 \cdot 10^{20}$	$5 \cdot 10^{20}$
<b>full</b>	<b>0.05</b>	<b>Omniscient</b>	<b>-4.41</b>	<b>4.48</b>
full	0.05	<i>REVISE</i>	$-1.04 \cdot 10^3$	$2.3 \cdot 10^3$
vanilla	0.05	<i>ECCo</i>	-15.5	17.2
vanilla	0.05	<i>Generic</i>	-11.7	12.8
vanilla	0.05	<i>Omniscient</i>	-12.4	14.1
vanilla	0.05	<i>REVISE</i>	-11.3	13.1
full	0.1	<i>ECCo</i>	$-3.41 \cdot 10^3$	$7.73 \cdot 10^3$
full	0.1	<i>Generic</i>	$-5.22 \cdot 10^{30}$	$1.19 \cdot 10^{31}$
<b>full</b>	<b>0.1</b>	<b>Omniscient</b>	<b>-4.78</b>	<b>5.12</b>
full	0.1	<i>REVISE</i>	-288	594
vanilla	0.1	<i>ECCo</i>	-15.5	17.2
vanilla	0.1	<i>Generic</i>	-10.9	11.9
vanilla	0.1	<i>Omniscient</i>	-12.7	14.4
vanilla	0.1	<i>REVISE</i>	-11.3	13.1
full	0.5	<i>ECCo</i>	-7.09	7.51
full	0.5	<i>Generic</i>	$-1.11 \cdot 10^{31}$	$2.53 \cdot 10^{31}$
<b>full</b>	<b>0.5</b>	<b>Omniscient</b>	<b>-4.58</b>	<b>4.83</b>
full	0.5	<i>REVISE</i>	$-1.19 \cdot 10^3$	$2.64 \cdot 10^3$
vanilla	0.5	<i>ECCo</i>	-15.5	17.2
vanilla	0.5	<i>Generic</i>	-11.7	12.8
vanilla	0.5	<i>Omniscient</i>	-12.4	14.1
vanilla	0.5	<i>REVISE</i>	-11.3	13.1
full	1	<i>ECCo</i>	-6.06	6.33
full	1	<i>Generic</i>	$-1.58 \cdot 10^{33}$	$3.59 \cdot 10^{33}$
<b>full</b>	<b>1</b>	<b>Omniscient</b>	<b>-4.66</b>	<b>4.89</b>
full	1	<i>REVISE</i>	$-1.16 \cdot 10^3$	$2.59 \cdot 10^3$
vanilla	1	<i>ECCo</i>	-15.5	17.3
vanilla	1	<i>Generic</i>	-10.9	11.9
vanilla	1	<i>Omniscient</i>	-12.7	14.4
vanilla	1	<i>REVISE</i>	-11.3	13.1
<b>full</b>	<b>5</b>	<b>ECCo</b>	<b>-2.57</b>	<b>2.07</b>
full	5	<i>Generic</i>	$-1.17 \cdot 10^{28}$	$2.66 \cdot 10^{28}$
full	5	<i>Omniscient</i>	-4.29	4.31
full	5	<i>REVISE</i>	-530	$1.16 \cdot 10^3$
vanilla	5	<i>ECCo</i>	-15.5	17.2
vanilla	5	<i>Generic</i>	-11.7	12.7
vanilla	5	<i>Omniscient</i>	-12.4	14.1
vanilla	5	<i>REVISE</i>	-11.3	13.1
<b>full</b>	<b>10</b>	<b>ECCo</b>	<b>-1.76</b>	<b>0.974</b>
full	10	<i>Generic</i>	$-1.54 \cdot 10^{33}$	$3.51 \cdot 10^{33}$

Continuing table below.

Objective	$\lambda_{\text{div}}(\text{train})$	Generator	Value	Std
full	10	<i>Omniscient</i>	-4.44	4.56
full	10	<i>REVISE</i>	$-1.52 \cdot 10^3$	$3.4 \cdot 10^3$
vanilla	10	<i>ECCo</i>	-15.5	17.3
vanilla	10	<i>Generic</i>	-10.9	11.9
vanilla	10	<i>Omniscient</i>	-12.7	14.4
vanilla	10	<i>REVISE</i>	-11.3	13.1
<b>full</b>	<b>15</b>	<b>ECCo</b>	<b>-1.37</b>	<b>0.365</b>
full	15	<i>Generic</i>	$-5.32 \cdot 10^{28}$	$1.21 \cdot 10^{29}$
full	15	<i>Omniscient</i>	-4.34	4.38
full	15	<i>REVISE</i>	-473	$1.03 \cdot 10^3$
vanilla	15	<i>ECCo</i>	-15.5	17.2
vanilla	15	<i>Generic</i>	-11.7	12.8
vanilla	15	<i>Omniscient</i>	-12.4	14.1
vanilla	15	<i>REVISE</i>	-11.3	13.1

## 476 C.2.1.3 Circles

- 477 • **Energy Penalty** (Table A4): *ECCo* consistently yields better results than *Vanilla*, though primarily for low to  
 478 medium choices of the energy penalty ( $<=5$ ) during training. The same goes for *Generic*, which sometimes  
 479 outperforms *ECCo* (for small energy penalty at evaluation time). *Omni* does alright for lower energy penalty  
 480 at evaluation time, but loses out for higher choices. *REVISE* performs poorly across the board (except very  
 481 low choices at evaluation time).
- 482 • **Cost (distance penalty)**: *ECCo* and *Generic* generally achieve the best results when no cost penalty is used  
 483 during training. Both *Omni* and *REVISE* are largely unaffected.
- 484 • **Maximum Iterations**: *ECCo* consistently yields better results for higher numbers of iterations. *Generic*  
 485 generally does best for a medium number (50). *Omni* is sometimes invalid (???).
- 486 • **Validity**: *ECCo* tends to outperform its *Vanilla* counterpart, though primarily for low to medium choices of  
 487 the energy penalty ( $<=5$ ) during training and evaluation. *Vanilla* typically worse across the board.
- 488 • **Accuracy**: Mostly unaffected, but *REVISE* again consistently some deterioration and *ECCo* deteriorates for  
 489 high choices of energy penalty during training, reflecting other outcomes above.

Table A4: Results for Circles data by energy penalty.

Objective	$\lambda_{\text{div}}(\text{train})$	Generator	Value	Std
<b>full</b>	<b>0.01</b>	<b>ECCo</b>	<b>-1.26</b>	<b>0.423</b>
full	0.01	<i>Generic</i>	-1.49	0.71
full	0.01	<i>Omniscient</i>	-5.21	5.25
full	0.01	<i>REVISE</i>	$-2.71 \cdot 10^{26}$	$6.37 \cdot 10^{26}$
vanilla	0.01	<i>ECCo</i>	-9.33	7.34
vanilla	0.01	<i>Generic</i>	-8.89	6.88
vanilla	0.01	<i>Omniscient</i>	-8.67	6.87
vanilla	0.01	<i>REVISE</i>	-8.65	6.8
full	0.05	<i>ECCo</i>	-1.29	0.397
<b>full</b>	<b>0.05</b>	<b>Generic</b>	<b>-1.21</b>	<b>0.356</b>
full	0.05	<i>Omniscient</i>	-5.08	5.09
full	0.05	<i>REVISE</i>	$-5.91 \cdot 10^{27}$	$1.36 \cdot 10^{28}$
vanilla	0.05	<i>ECCo</i>	-9.35	7.32
vanilla	0.05	<i>Generic</i>	-8.85	6.87
vanilla	0.05	<i>Omniscient</i>	-8.7	6.96
vanilla	0.05	<i>REVISE</i>	-8.52	6.76
<b>full</b>	<b>0.1</b>	<b>ECCo</b>	<b>-1.2</b>	<b>0.383</b>
full	0.1	<i>Generic</i>	-1.5	0.735
full	0.1	<i>Omniscient</i>	-5.17	5.23
full	0.1	<i>REVISE</i>	$-3.06 \cdot 10^{26}$	$7.7 \cdot 10^{26}$

Continuing table below.

<b>Objective</b>	$\lambda_{\text{div}}(\text{train})$	<b>Generator</b>	<b>Value</b>	<b>Std</b>
vanilla	0.1	<i>ECCo</i>	-9.33	7.32
vanilla	0.1	<i>Generic</i>	-8.88	6.86
vanilla	0.1	<i>Omniscient</i>	-8.69	6.9
vanilla	0.1	<i>REVISE</i>	-8.68	6.81
<b>full</b>	<b>0.5</b>	<b>ECCo</b>	<b>-1.12</b>	<b>0.217</b>
full	0.5	<i>Generic</i>	-1.21	0.352
full	0.5	<i>Omniscient</i>	-5.09	5.12
full	0.5	<i>REVISE</i>	$-5.97 \cdot 10^{27}$	$1.37 \cdot 10^{28}$
vanilla	0.5	<i>ECCo</i>	-9.35	7.3
vanilla	0.5	<i>Generic</i>	-8.89	6.92
vanilla	0.5	<i>Omniscient</i>	-8.68	6.93
vanilla	0.5	<i>REVISE</i>	-8.53	6.75
<b>full</b>	<b>1</b>	<b>ECCo</b>	<b>-1.1</b>	<b>0.163</b>
full	1	<i>Generic</i>	-1.49	0.726
full	1	<i>Omniscient</i>	-5.16	5.2
full	1	<i>REVISE</i>	$-3.09 \cdot 10^{26}$	$7.22 \cdot 10^{26}$
vanilla	1	<i>ECCo</i>	-9.34	7.36
vanilla	1	<i>Generic</i>	-8.86	6.85
vanilla	1	<i>Omniscient</i>	-8.7	6.9
vanilla	1	<i>REVISE</i>	-8.69	6.85
full	5	<i>ECCo</i>	-1.75	0.154
<b>full</b>	<b>5</b>	<b>Generic</b>	<b>-1.21</b>	<b>0.363</b>
full	5	<i>Omniscient</i>	-5.14	5.16
full	5	<i>REVISE</i>	$-1.1 \cdot 10^{28}$	$2.5 \cdot 10^{28}$
vanilla	5	<i>ECCo</i>	-9.36	7.32
vanilla	5	<i>Generic</i>	-8.88	6.91
vanilla	5	<i>Omniscient</i>	-8.7	6.93
vanilla	5	<i>REVISE</i>	-8.52	6.73
full	10	<i>ECCo</i>	$-1.02 \cdot 10^6$	$2.32 \cdot 10^6$
<b>full</b>	<b>10</b>	<b>Generic</b>	<b>-1.49</b>	<b>0.702</b>
full	10	<i>Omniscient</i>	-5.13	5.16
full	10	<i>REVISE</i>	$-3.74 \cdot 10^{26}$	$9.09 \cdot 10^{26}$
vanilla	10	<i>ECCo</i>	-9.31	7.33
vanilla	10	<i>Generic</i>	-8.87	6.86
vanilla	10	<i>Omniscient</i>	-8.7	6.89
vanilla	10	<i>REVISE</i>	-8.69	6.83
full	15	<i>ECCo</i>	$-3.31 \cdot 10^{13}$	$7.54 \cdot 10^{13}$
<b>full</b>	<b>15</b>	<b>Generic</b>	<b>-1.22</b>	<b>0.37</b>
full	15	<i>Omniscient</i>	-5.2	5.23
full	15	<i>REVISE</i>	$-9.01 \cdot 10^{27}$	$2.06 \cdot 10^{28}$
vanilla	15	<i>ECCo</i>	-9.38	7.34
vanilla	15	<i>Generic</i>	-8.86	6.87
vanilla	15	<i>Omniscient</i>	-8.69	6.96
vanilla	15	<i>REVISE</i>	-8.51	6.73