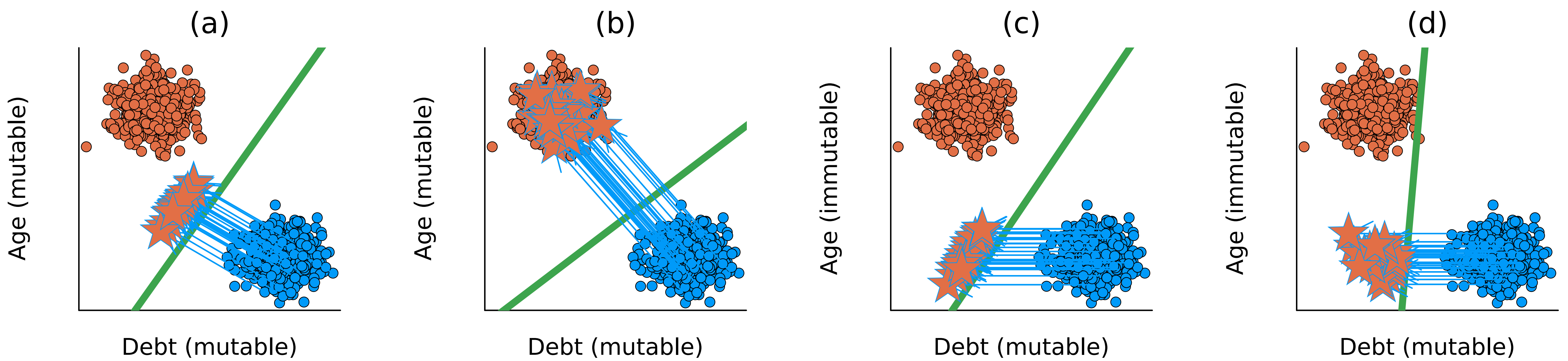


Counterfactual training leverages counterfactual explanations to make models more 1) explainable, 2) sensitive to actionability constraints and 3) adversarially robust.



METHODOLOGY

Counterfactual search objective (equation (1)):

$$\min_{\mathbf{x}' \in \mathcal{X}^D} \{y_{\text{loss}}(\mathbf{M}_{\theta}(\mathbf{x}'), \mathbf{y}^+) + \lambda_{\text{reg}}(\mathbf{x}')\}$$

Counterfactual training objective (equation (2)):

$$\min_{\theta} y_{\text{loss}}(\mathbf{M}_{\theta}(\mathbf{x}), \mathbf{y}) + \lambda_{\text{div}} \text{div}(\mathbf{x}^+, \mathbf{x}'_{\text{CE}}, \mathbf{y}^+; \theta) + \lambda_{\text{adv}} \text{advloss}(\mathbf{M}_{\theta}(\mathbf{x}'_{\text{AE}}), \mathbf{y}_{\text{AE}}) + \lambda_{\text{reg}} \text{ridge}(\mathbf{x}^+, \mathbf{x}'_{\text{CE}}, \mathbf{y}; \theta)$$

Pseudo-Code for counterfactual training:

Require: Training dataset \mathcal{D} , initialize model \mathbf{M}_{θ}

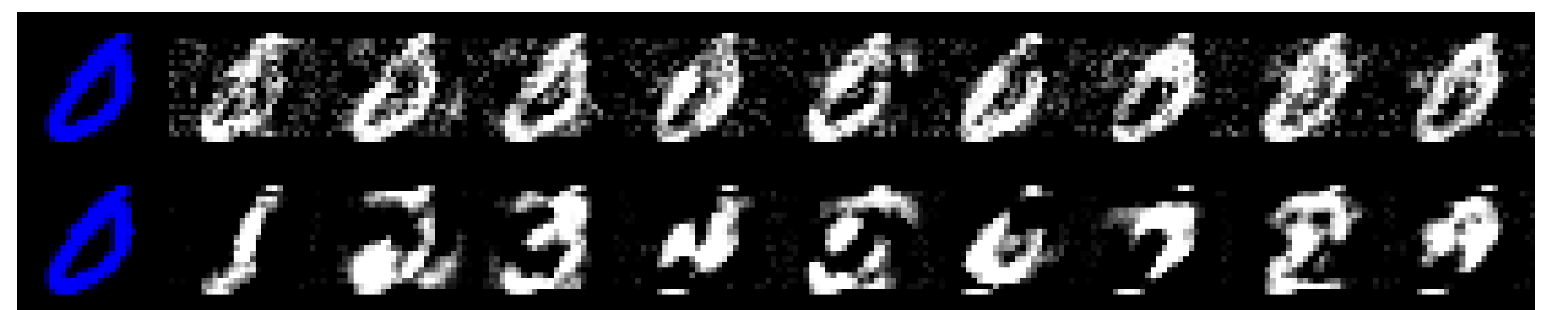
- 1: **while** not converged **do**
- 2: Sample $\mathbf{x}'_0 \sim \mathbf{X}$, $\mathbf{y}^+ \sim \mathcal{U}(\mathcal{Y})$ and $\mathbf{x}^+ \sim \mathbf{X}^+$
- 3: **for** $t = 1$ to T **do**
- 4: Backpropagate $\nabla_{\mathbf{x}'}$ through equation (1)
- 5: Store $\mathbf{x}'_{\text{CE}}, \mathbf{x}'_{\text{AE}}, \mathbf{y}_{\text{AE}}$
- 6: **end for**
- 7: Sample mini-batches $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^{n_b}$ from dataset \mathcal{D}
- 8: Distribute $(\mathbf{x}'_{\text{CE}_i}, \mathbf{y}^+_{i}, \mathbf{x}'_{\text{AE}_i}, \mathbf{y}_{\text{AE}_i}, \mathbf{x}^+_{i})_{i=1}^{n_{\text{CE}}}$
- 9: **for** each batch **do**
- 10: Backpropagate ∇_{θ} through equation (2)
- 11: **end for**
- 12: **end while**
- 13: **return** \mathbf{M}_{θ}

SUMMARY

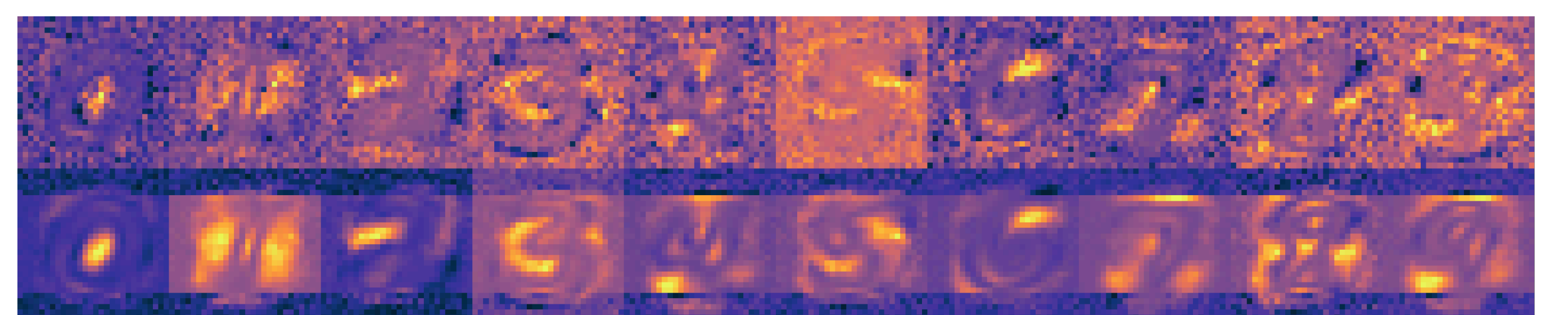
On-the-fly counterfactual explanations are used to induce explainability and actionability through contrastive divergence and recycled as adversarial examples to further improve robustness.

IMPROVED PLAUSIBILITY

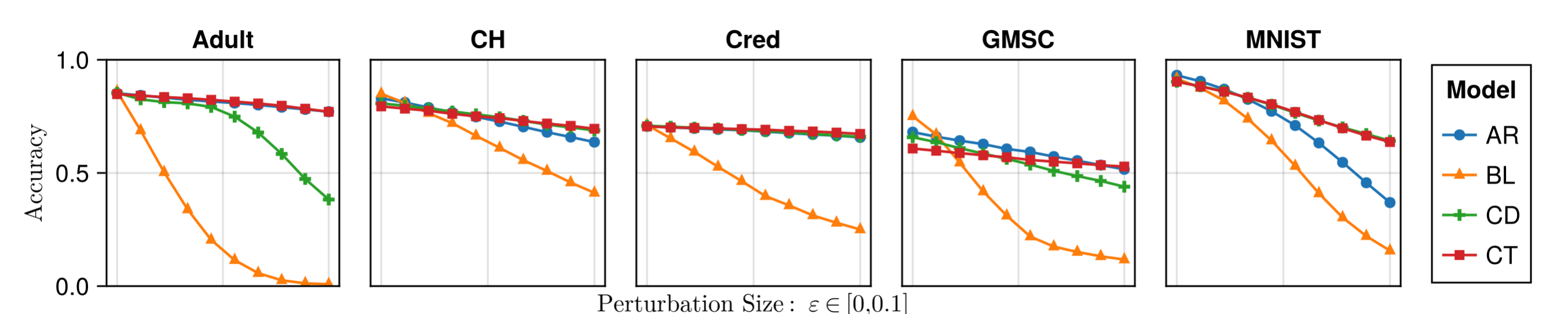
Data	IP (−%)	IP* (−%)	Cost (−%)
LS	26.26 ± 0.67*	51.28 ± 2.01*	16.41 ± 0.57*
Circ	58.88 ± 0.37*	93.84 ± 6.70*	42.99 ± 0.85*
Moon	19.59 ± 0.73*	8.00 ± 9.44	5.16 ± 1.00*
OL	−1.93 ± 1.12	−27.70 ± 14.59	40.86 ± 2.30*
Adult	0.19 ± 1.05	34.35 ± 5.61*	4.03 ± 4.03
CH	10.65 ± 1.47*	63.06 ± 4.25*	44.23 ± 1.43*
Cred	10.14 ± 1.59*	50.35 ± 12.26*	−18.17 ± 4.40*
GMSC	10.65 ± 2.28*	24.75 ± 4.84*	66.01 ± 1.41*
MNIST	6.36 ± 1.70*	−70.31 ± 217.60	−35.11 ± 6.96*
Avg.	15.64	25.29	18.49



IMPROVED ACTIONABILITY



IMPROVED ROBUSTNESS



BIBLIOGRAPHY