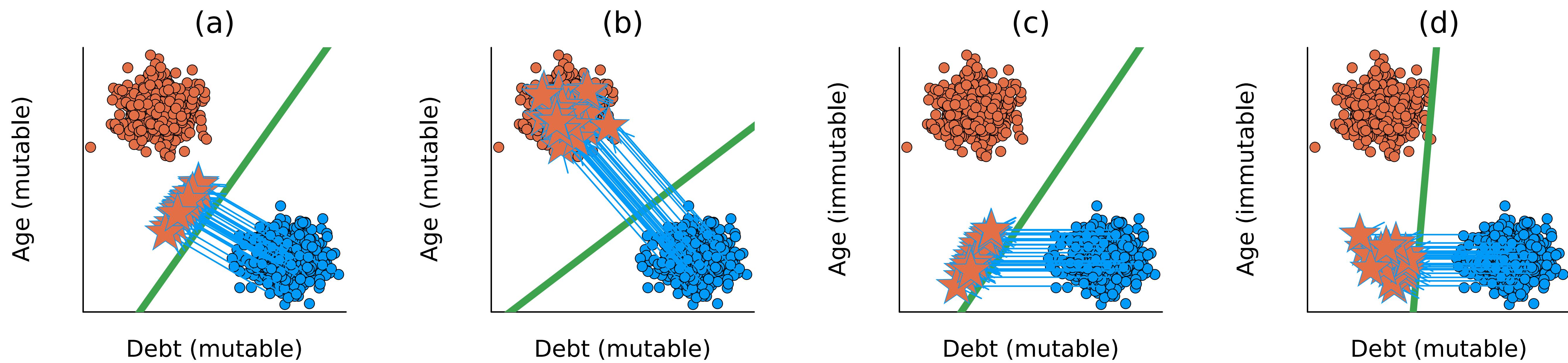


Counterfactual training leverages counterfactual explanations to make models more
 1) explainable, 2) sensitive to actionability constraints and 3) adversarially robust.



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)):

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

Pseudo-Code for counterfactual training:

```

Require: Training dataset  $\mathcal{D}$ , initialize model  $\mathbf{M}_\theta$ 
1: while not converged do
2:   Sample  $\mathbf{x}'_0 \sim \mathbf{X}$ ,  $y^+ \sim \mathcal{U}(\mathcal{Y})$  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  $(\mathbf{x}'_{\text{CE}i}, y^+_{i}, \mathbf{x}'_{\text{AE}i}, \mathbf{y}_{\text{AE}i}, \mathbf{x}^+_{i})_{i=1}^{n_{\text{CE}}}$ 
9:   for each batch do
10:    Backpropagate  $\nabla_\theta$  through equation (2)
11:   end for
12: end while
13: return  $\mathbf{M}_\theta$ 

```

SUMMARY

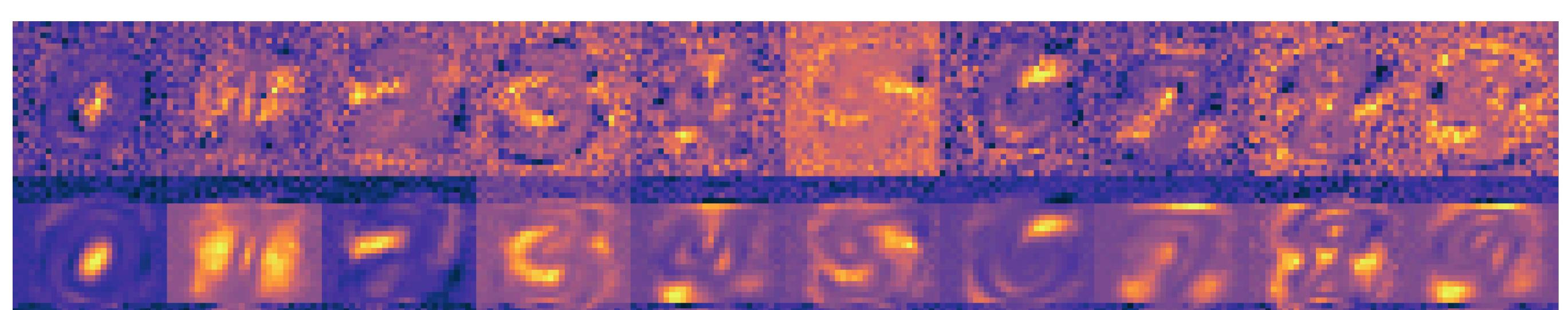
On-the-fly counterfactual explanations are used to induce explainability and actionability through contrastive divergence and recycled as adversarial examples to further improve robustness.

IMPROVED PLAUSIBILITY

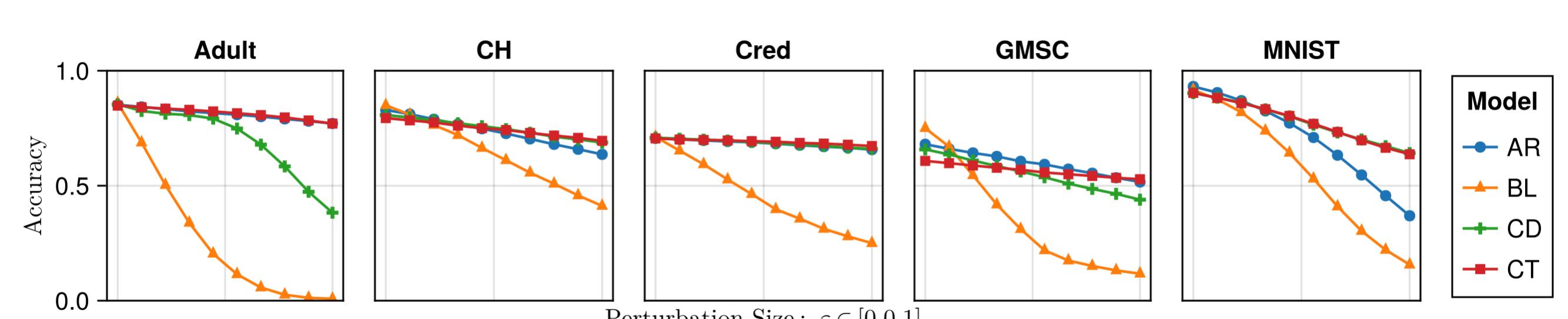
Data	IP (-%)	IP* (-%)	Cost (-%)	
LS	$26.26 \pm 0.67^*$	$51.28 \pm$	2.01^*	$16.41 \pm 0.57^*$
Circ	$58.88 \pm 0.37^*$	$93.84 \pm$	6.70^*	$42.99 \pm 0.85^*$
Moon	$19.59 \pm 0.73^*$	$8.00 \pm$	9.44	$5.16 \pm 1.00^*$
OL	-1.93 ± 1.12	$-27.70 \pm$	14.59	$40.86 \pm 2.30^*$
Adult	0.19 ± 1.05	$34.35 \pm$	5.61^*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



IMPROVED ACTIONABILITY



IMPROVED ROBUSTNESS



BIBLIOGRAPHY