Counterfactual Concept Bottleneck Models

Gabriele Dominici, Pietro Barbiero, Mateo Espinosa Zarlenga, Alberto Termine, Martin Gjoreski, Giuseppe Marra, Marc Langheinrich*

Counterfactual Concept Bottleneck Models ***predict*** *concept & task labels (the* ***"What?"****),* ***simulate*** *concept interventions (the* ***"How?"****) and* ***imagine*** *interpretable counterfactuals (the* ***"Why not?"****)*

Università della Svizzera italiana
IBM
Scuola Normale Superiore
KU LEUVEN

Enhance human-machine interactions

What?

Predict class labels for new inputs → Prediction
e.g., “What is the diagnosis for the X-ray image?”

How?

Simulate changes to evaluate how they impacts class predictions → Intervention
e.g., “How would the absence of collapsed lung impact the presence of pneumothorax?”

Why not?

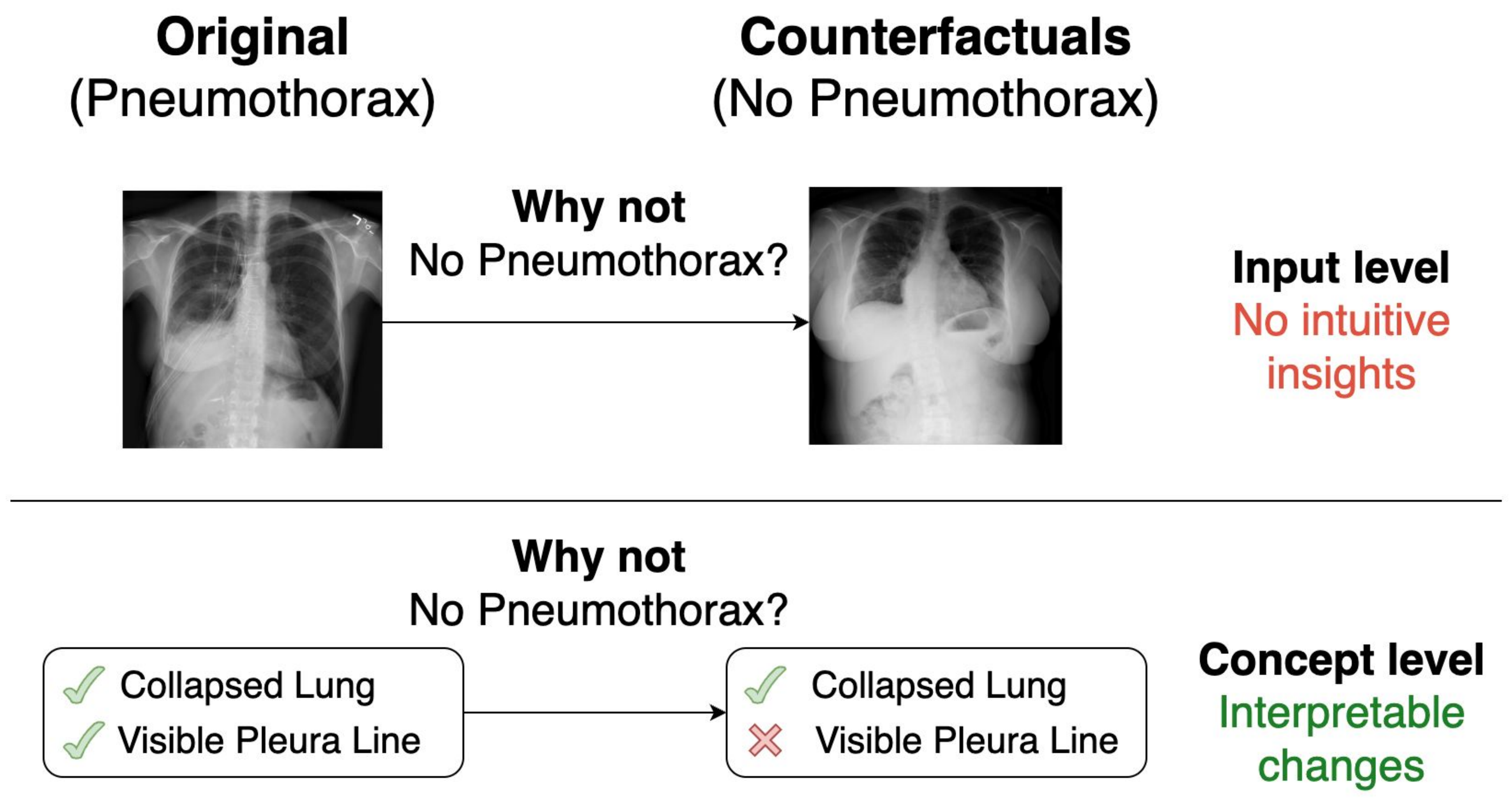
Imagine alternative scenarios that result in different class predictions → Counterfactuals
e.g., “Why is the patient not having pneumothorax?”

No model is trained to answer all these three question at once

Black Box: ✅ What? ❌ How? ❌ Why not?

Concept Bottleneck Models: ✅ What? ✅ How? ❌ Why not?

Post-hoc Counterfactuals Generators: ✅ What? ❌ How? ✅ Why not?
⚠️ Not Interpretable Counterfactuals



Counterfactuals
Concept Bottleneck Models

Results - What?

Task and Concept AUC

	dsprites		MNIST add		CUB		CIFAR10		SIIM Pneumothorax	
	Task (%) (↑)	Concept (%) (↑)	Task (%) (↑)	Concept (%) (↑)	Task (%) (↑)	Concept (%) (↑)	Task (%) (↑)	Concept (%) (↑)	Task (%) (↑)	Concept (%) (↑)
Black Box	99.9 ± 0.0	-	98.4 ± 0.0	-	94.4 ± 0.0	-	99.7 ± 0.0	-	95.1 ± 0.2	
CBM	99.6 ± 0.1	98.8 ± 0.1	98.4 ± 0.0	99.7 ± 0.0	93.0 ± 0.0	85.3 ± 0.1	99.3 ± 0.0	99.1 ± 0.0	89.4 ± 0.6	99.6 ± 0.1
CEM	99.8 ± 0.1	98.6 ± 0.1	98.4 ± 0.0	99.5 ± 0.0	93.3 ± 0.1	85.7 ± 0.1	99.8 ± 0.0	99.2 ± 0.0	94.2 ± 0.1	99.5 ± 0.0
BayCon	99.6 ± 0.0	98.7 ± 0.1	98.4 ± 0.0	99.7 ± 0.0	93.1 ± 0.1	85.3 ± 0.1	99.4 ± 0.0	99.1 ± 0.0	89.3 ± 0.8	99.6 ± 0.1
CCHVAE	99.5 ± 0.1	98.9 ± 0.1	98.4 ± 0.0	99.7 ± 0.0	93.2 ± 0.1	85.4 ± 0.1	99.3 ± 0.0	99.2 ± 0.0	89.6 ± 0.5	99.3 ± 0.1
VAECF	99.5 ± 0.0	98.8 ± 0.1	98.4 ± 0.0	99.7 ± 0.0	92.9 ± 0.2	85.4 ± 0.1	99.3 ± 0.0	99.2 ± 0.0	89.5 ± 0.8	99.6 ± 0.1
VCNET	99.7 ± 0.0	98.6 ± 0.1	98.2 ± 0.0	99.6 ± 0.0	89.6 ± 0.5	80.7 ± 0.1	99.1 ± 0.1	97.4 ± 0.2	91.4 ± 0.7	99.2 ± 0.0
CF-CBM	99.7 ± 0.0	98.7 ± 0.1	97.0 ± 0.1	99.3 ± 0.0	91.4 ± 0.2	84.8 ± 0.2	99.4 ± 0.0	96.6 ± 0.0	92.0 ± 0.5	98.2 ± 0.0

Results - How?

Concept Importance on dSprites

	Post-hoc	Joint (ours)
Shape 1	0.16 ± 0.03	0.30 ± 0.15
Shape 2	0.13 ± 0.03	0.18 ± 0.08
Shape 3	0.16 ± 0.03	0.32 ± 0.09
Two obj	0.04 ± 0.01	0.16 ± 0.07
Colours (confounder)	0.17 ± 0.03	0.02 ± 0.01

Concept Importance distributions on MNIST and CUB

Dataset	Percentile	Mean Post-hoc	Mean Joint (ours)	p-value
MNIST	Top 10% (↑)	0.1589 ± 0.0203	$\mathbf{0.1827} \pm 0.0676$	**0.0001**
	Bottom 50% (↓)	0.0308 ± 0.0175	$\mathbf{0.0257} \pm 0.0143$	0.2437
CUB	Top 10% (↑)	0.0093 ± 0.0029	$\mathbf{0.0131} \pm 0.0044$	**0.0**
	Bottom 50% (↓)	0.0016 ± 0.0006	$\mathbf{0.0013} \pm 0.0008$	**0.0**

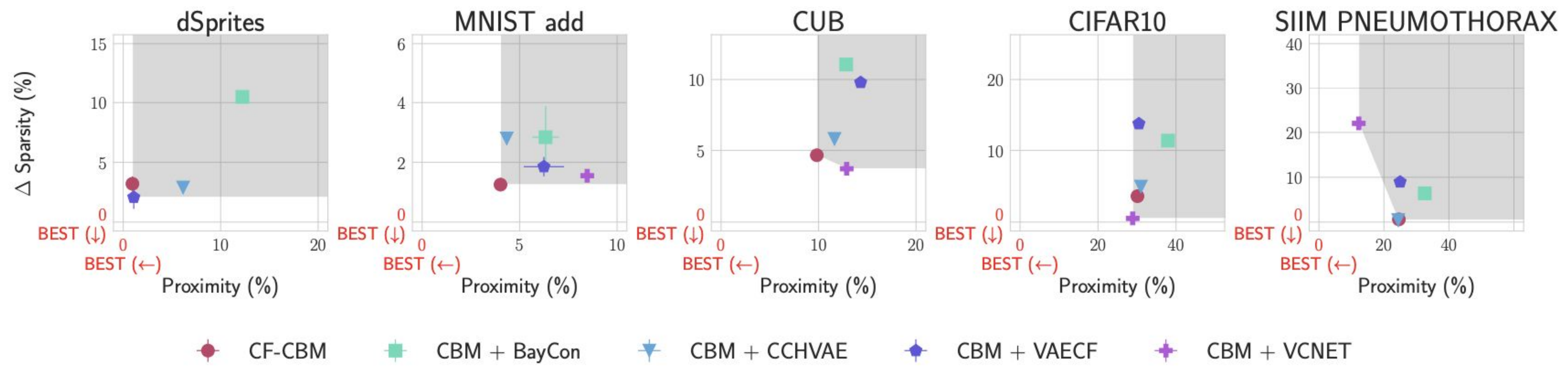
Results - Why not?

Validity

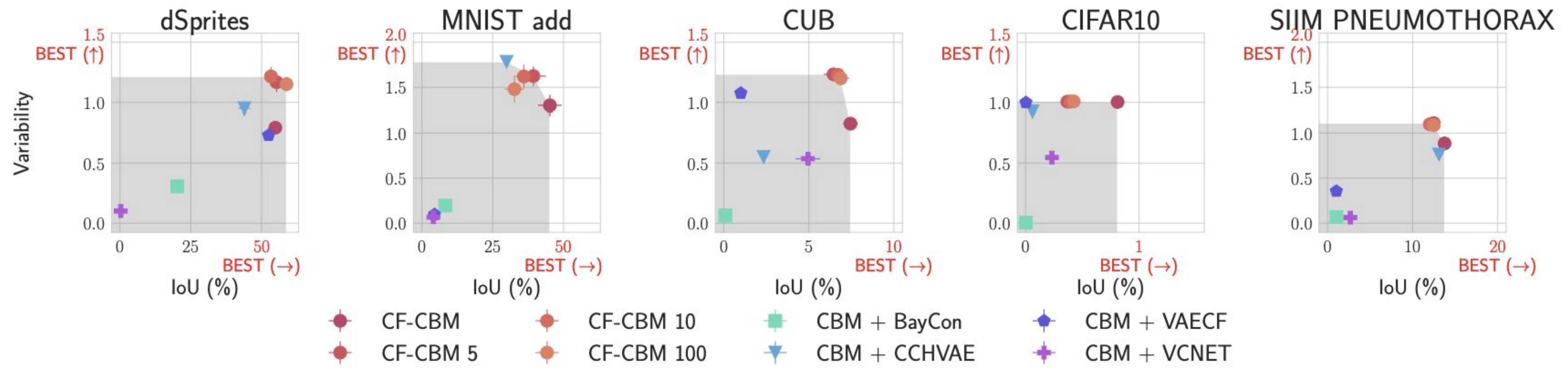
	dSprites (↑)	MNIST add (↑)	CUB (↑)	CIFAR10 (↑)	SIIM Pneumothorax (↑)	avg. (↑)
CBM+BayCon	$\mathbf{100.0 \pm 0.0}$	94.6 ± 0.5	66.2 ± 0.5	91.0 ± 3.3	$\mathbf{100.0 \pm 0.0}$	90.4 ± 0.7
CBM+CCHVAE	98.4 ± 0.6	94.7 ± 0.2	78.7 ± 0.4	93.6 ± 0.1	93.0 ± 0.6	91.7 ± 0.2
CBM+VAECF	$\mathbf{100.0 \pm 0.0}$	82.0 ± 5.9	95.2 ± 0.2	91.0 ± 0.8	$\mathbf{100.0 \pm 0.0}$	93.4 ± 1.2
CBM+VCNET	$\mathbf{100.0 \pm 0.0}$	89.9 ± 0.9	93.1 ± 0.5	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$	96.6 ± 0.2
CF-CBM (ours)	$\mathbf{100.0 \pm 0.0}$	$\mathbf{96.4 \pm 0.7}$	$\mathbf{99.0 \pm 0.1}$	$\mathbf{100.0 \pm 0.0}$	$\mathbf{99.9 \pm 0.1}$	$\mathbf{99.06 \pm 0.1}$

Results - Why not?

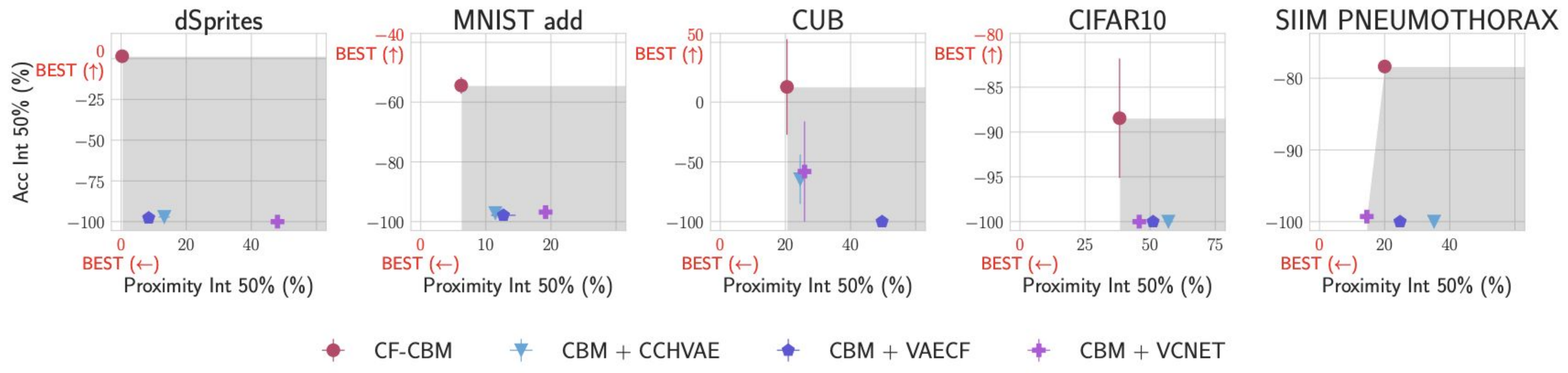
Sparsity / Proximity



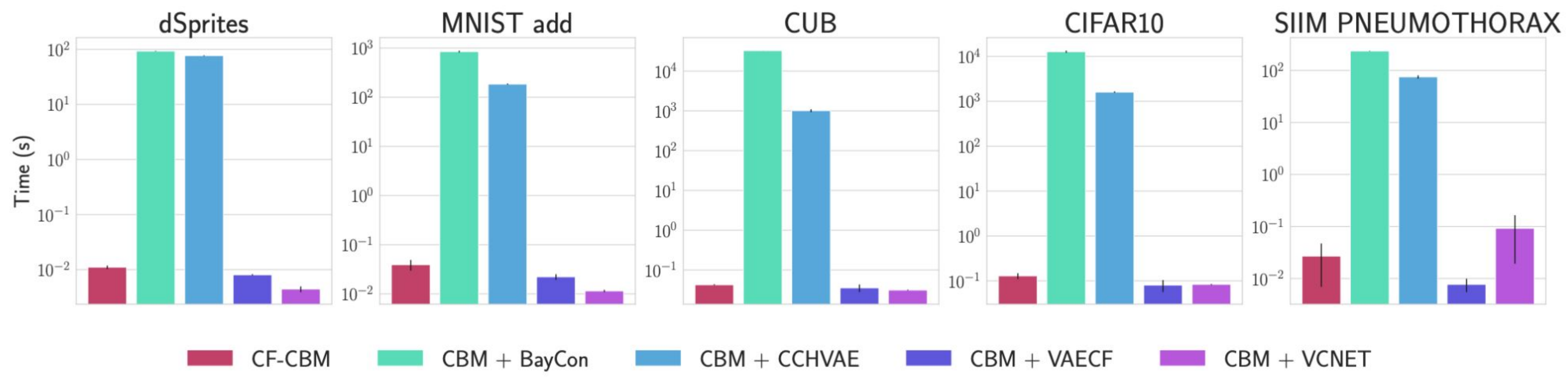
Variability / IoU



Accuracy / Proximity after proposed Interventions



Computing Time



Counterfactual Concept Bottleneck Models

Gabriele Dominici, Pietro Barbiero, Mateo Espinosa Zarlenga, Alberto Termine, Martin Gjoreski, Giuseppe Marra, Marc Langheinrich*