

 $\sigma = 2.0$ GT = 8, Pred = 2

(a) Explanations under Gaussian noise

 $\sigma = 0.0$ GT = 8, Pred = 8

Integrated Gradients

Gradient SHAP

Deep SHAP $\sigma = 0.5$ GT = 8. Pred = 8 $\sigma = 1.0$ GT = 8. Pred = 8

(b) Relative attribution change