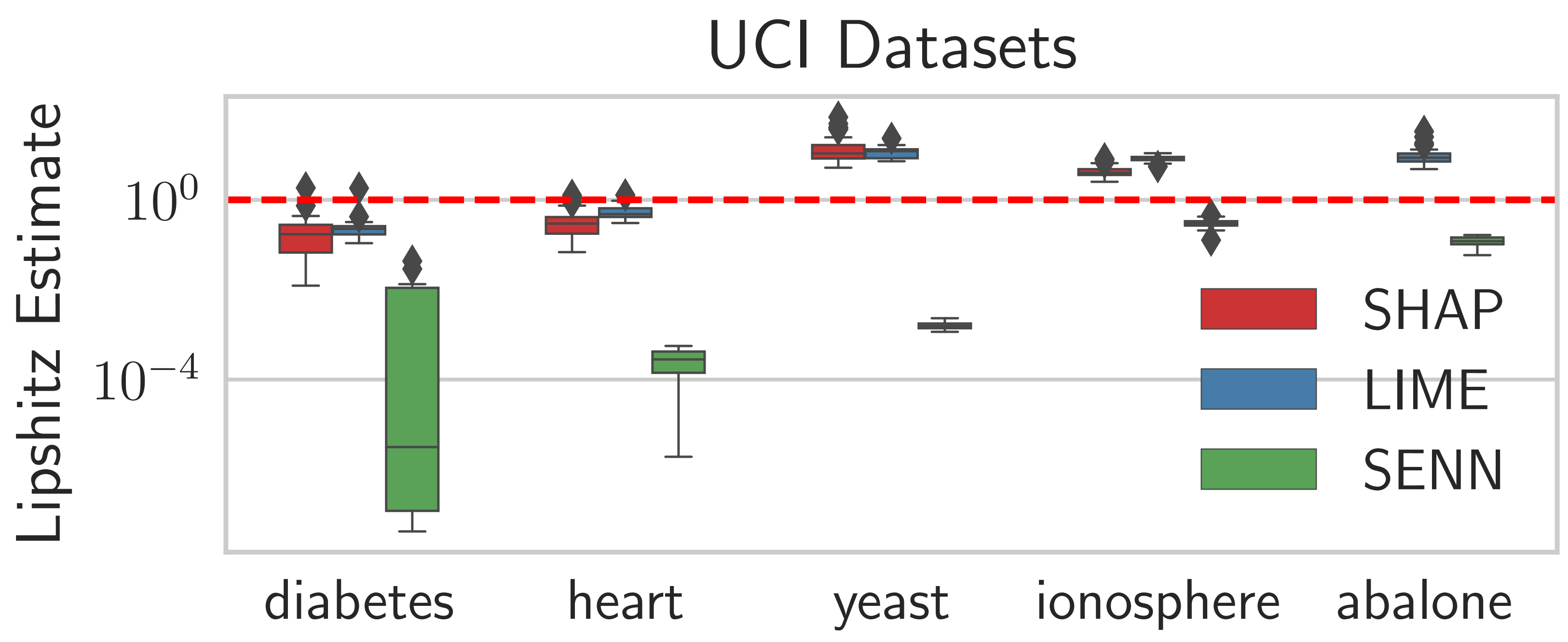
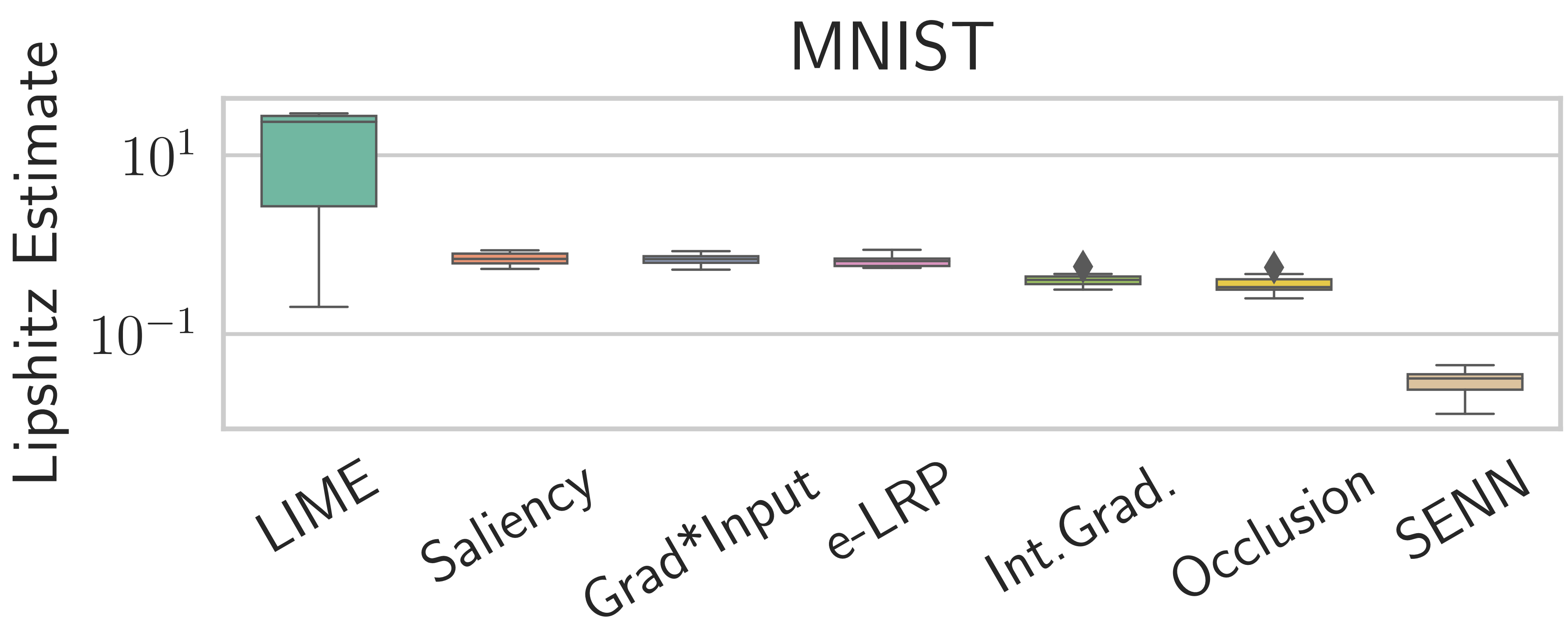
SENN
Robustness

Adversarial robustness estimation:

$$\hat{L}(x) = \arg \max_{\hat{x}\in B_\epsilon(x)} \|f_{expl}(\hat{x}) - f_{expl}(x)\|_2/\|h(\hat{x}) - h(x)\|$$

Results aggregated over full dataset:

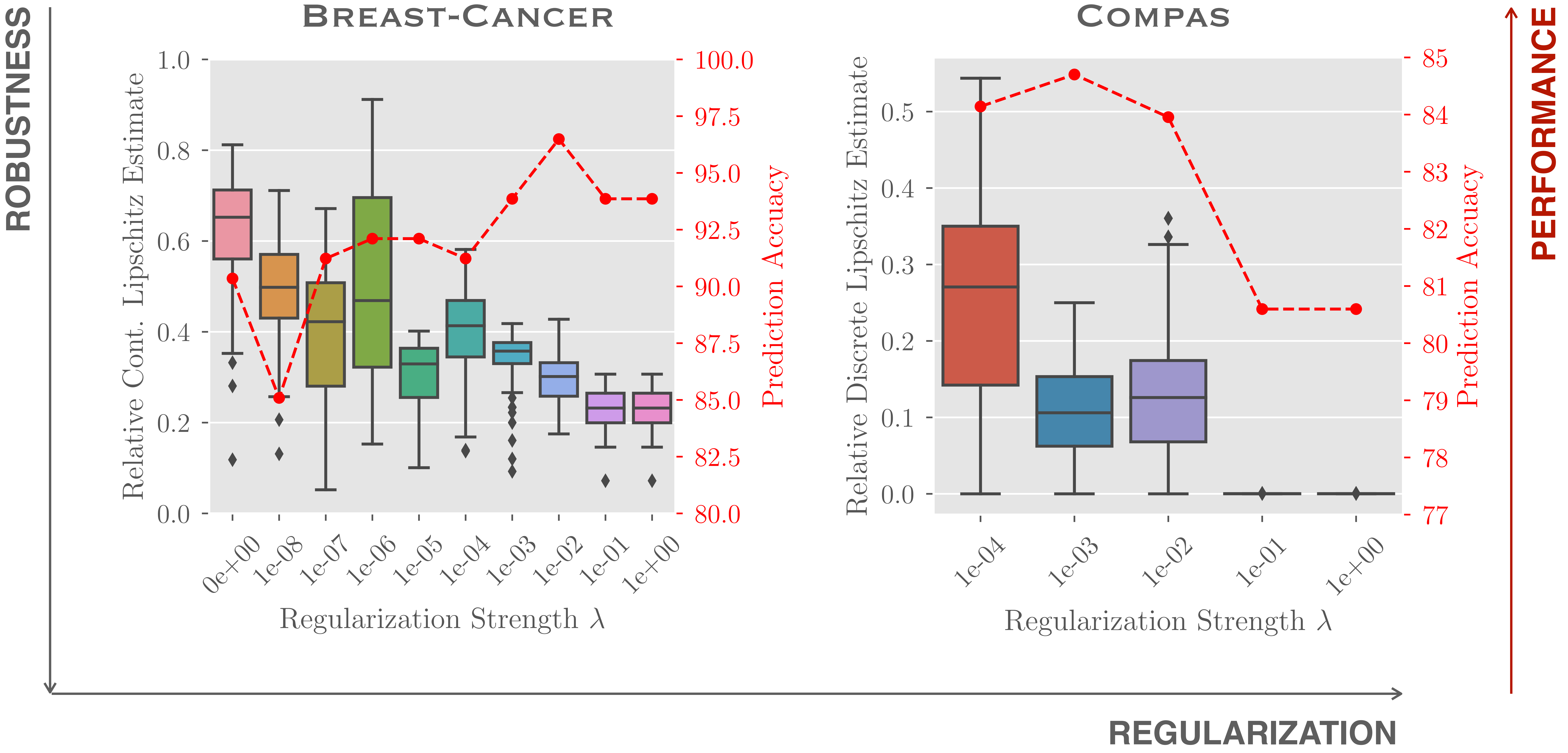




SENN

Robustness vs Performance

Effect of regularization on SENN's stability:


ROBUSTNESS
BREAST-CANCER
Relative Cont. Lipschitz Estimate
Prediction Accuacy
0e+00
1e-08
1e-07
1e-06
1e-05
1e-04
1e-03
1e-02
1e-01
1e+00
Regularization Strength λ
COMPAS
Relative Discrete Lipschitz Estimate
Prediction Accuacy
1e-04
1e-03
1e-02
1e-01
1e+00
Regularization Strength λ
PERFORMANCE
REGULARIZATION