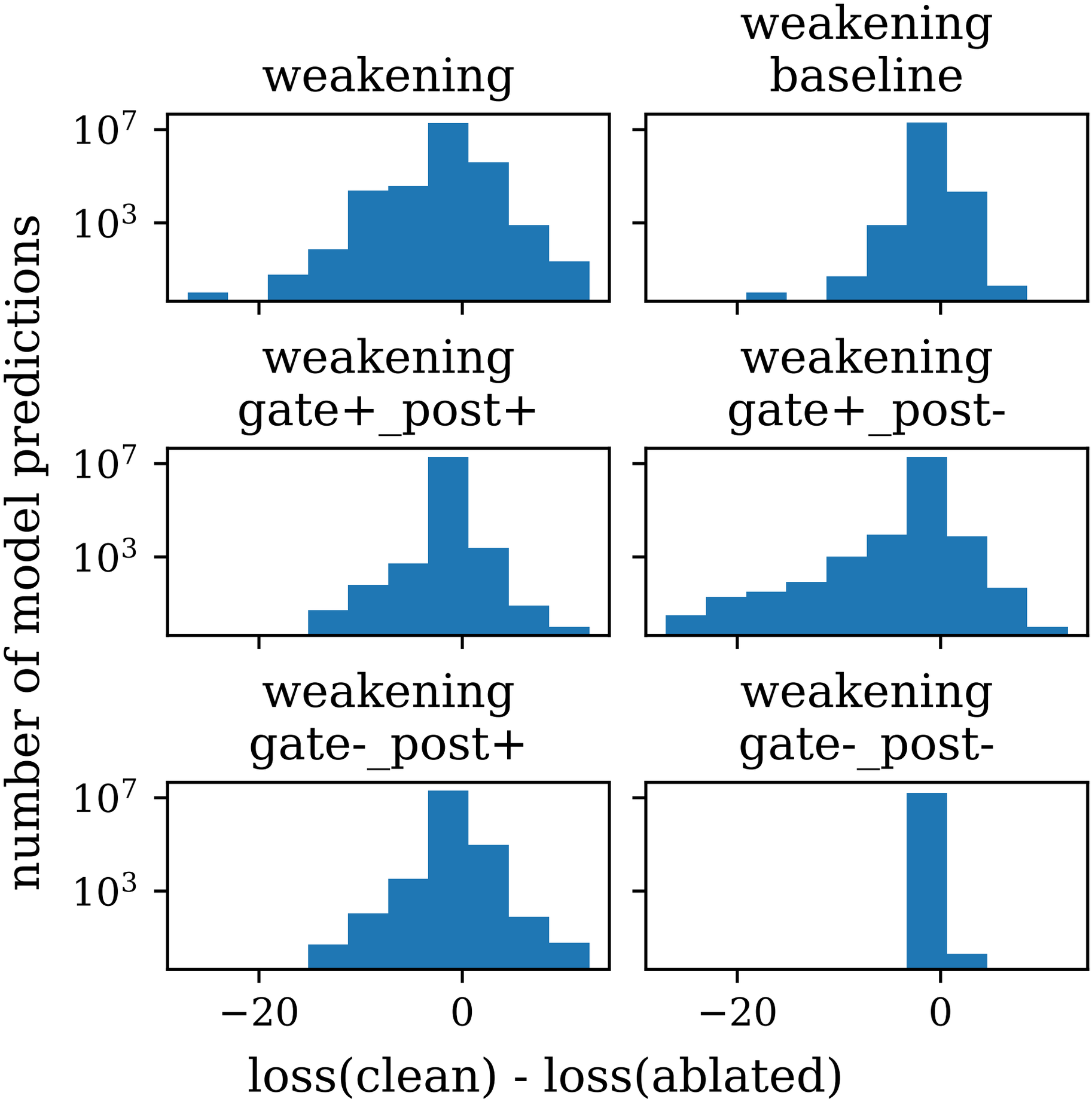
weakening
weakening baseline
weakening gate+_post+
weakening gate+_post-
weakening gate-_post+
weakening gate-_post-
number of model predictions
10^7
10^3
−20
0
loss(clean) - loss(ablated)