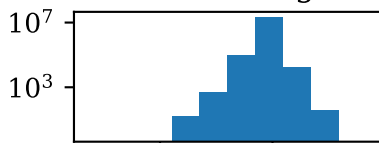
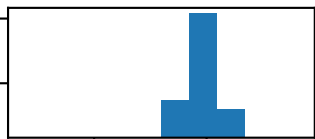


number of model predictions

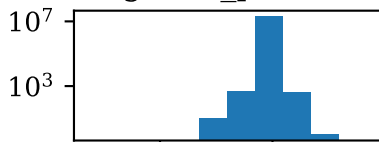
weakening



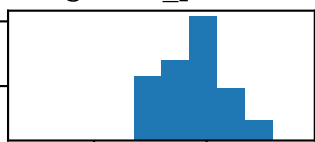
weakening
baseline



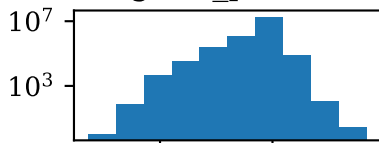
weakening
gate+_post+



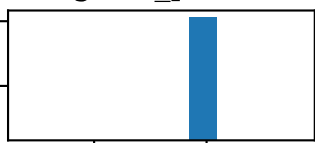
weakening
gate+_post-



weakening
gate-_post+



weakening
gate-_post-



loss(clean) - loss(ablated)