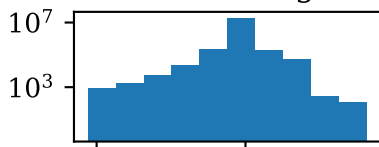
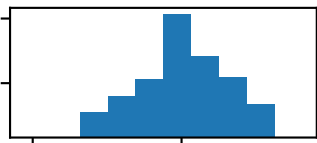


number of model predictions

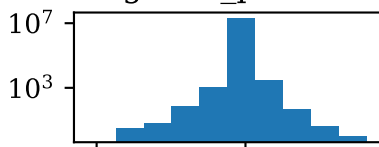
weakening



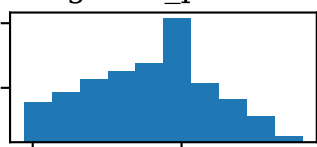
weakening  
baseline



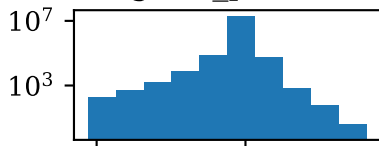
weakening  
gate+\_post+



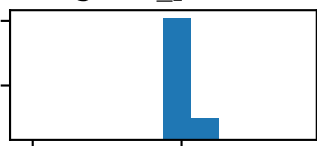
weakening  
gate+\_post-



weakening  
gate-\_post+



weakening  
gate-\_post-



entropy(clean) - entropy(ablated)