



Figure B3: Distribution of Oracle Penalized LogP Scores We plot the distribution of oracle scores for the top 128 surrogate model-ranked designs in black, and the distribution for all 2,048 generated designs in light gray for each of the offline model-based optimization methods assessed in our work across 10 random seeds. While GABO and BO-qEI have similar distributions, GABO is able to more reliably rank top-performing designs higher, such that these designs can be identified even under limited oracle query budgets.