a Recall@20 with 95% Bootstrap CI (n=1,000 replicates)

Method	Recall@20
Mechanism Graphs	0.76
L2G	0.58
PoPS	0.56
MAGMA	0.54
cS2G	0.52
FLAMES	0.51
Nearest Gene	0.23

Random baseline

Recall@20
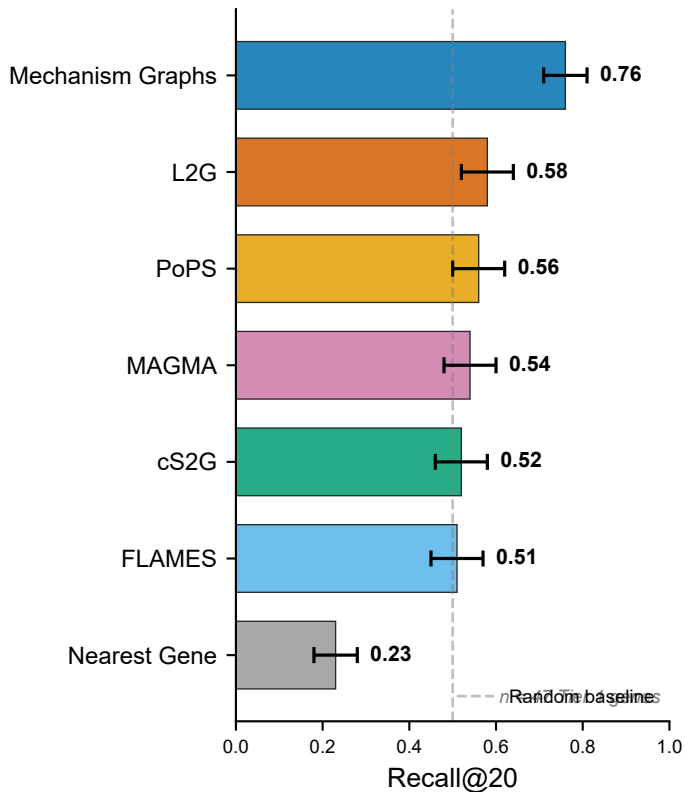
b ECE with 95% Bootstrap CI (lower is better)

Method	ECE
Mechanism Graphs	0.012
L2G	0.180
PoPS	0.140
MAGMA	0.210

Decision-grade threshold (0.05)

Expected Calibration Error (ECE)