

RegulatoryBench: Task-Stratified Evaluation Framework



*Credible set to causal gene
(14,016 pairs, 560 loci)*

*CRISPRi-validated links
(19,825 pairs, 78% leakage)*

*Variant to regulatory element
(Future work)*

Key Insight: Task conflation inflates method performance.
Distance dominates Task A (AUROC=0.930); ABC wins Task B (AUROC=0.885).
Proper task separation reveals 78% leakage in current benchmarks.