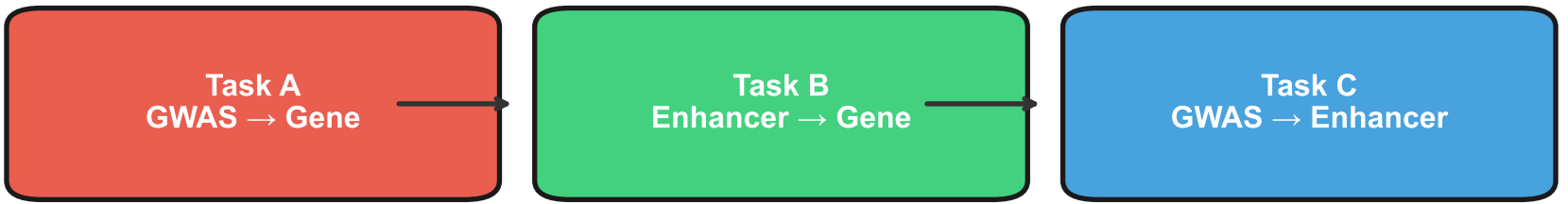
RegulatoryBench: Task-Stratified Evaluation Framework



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.