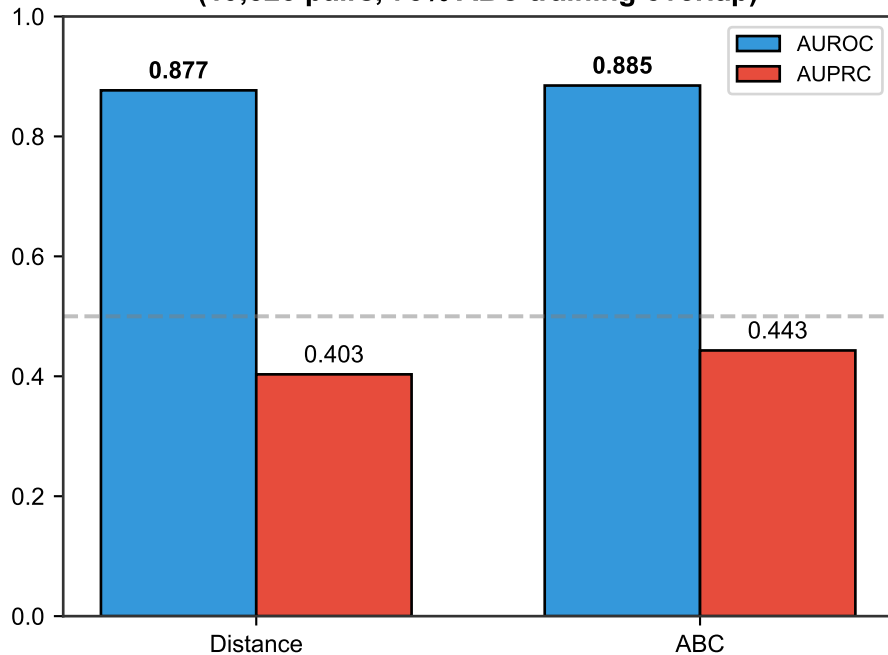**(A) Task B Full Dataset**
**(19,825 pairs, 78% ABC training overlap)**

**(B) Held-Out (Excluding ABC Training)**
**(4,378 pairs, leakage-free)**

Note: ABC cannot be evaluated on held-out data (would be 0% training overlap)