Workshop on Generative and Experimental Perspectives for Biomolecular Design @ ICLR 2024

Revealing data leakage in protein interaction benchmarks

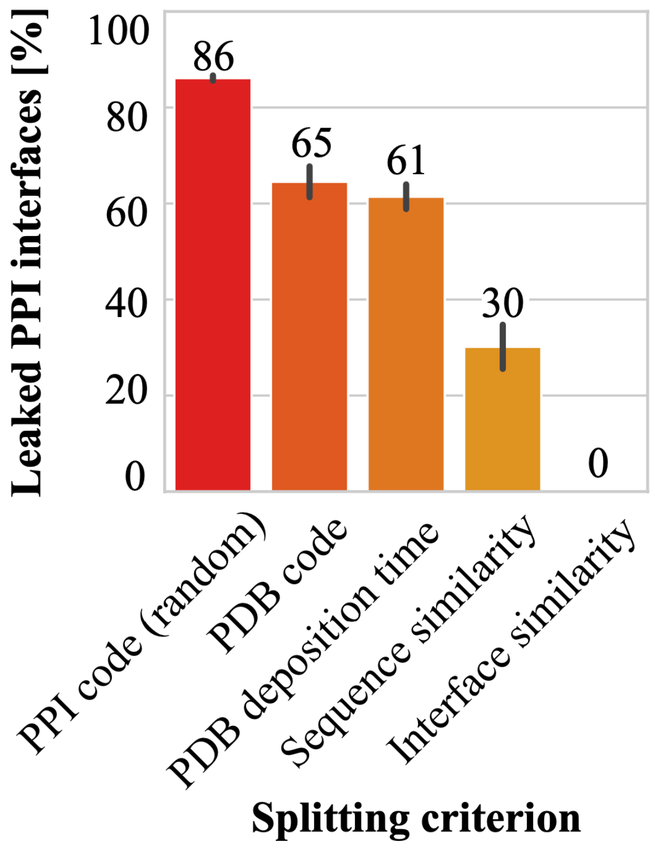
Anton Bushuiev[*1], **Roman Bushuiev**[*1,4], Jiri Sedlar[1], Tomas Pluskal[4], Jiri Damborsky[2,3], Stanislav Mazurenko[2,3], Josef Sivic[1]

1 CZECH INSTITUTE OF INFORMATICS ROBOTICS AND CYBERNETICS CTU IN PRAGUE
2 LOSCHMIDT LABORATORIES
3 FNUSA ICRC ST. ANNE'S UNIVERSITY HOSPITAL BRNO INTERNATIONAL CLINICAL RESEARCH CENTER
4 ÚOCHB AV ČR IOCB PRAGUE

Paper

iDist in the PPIRef package

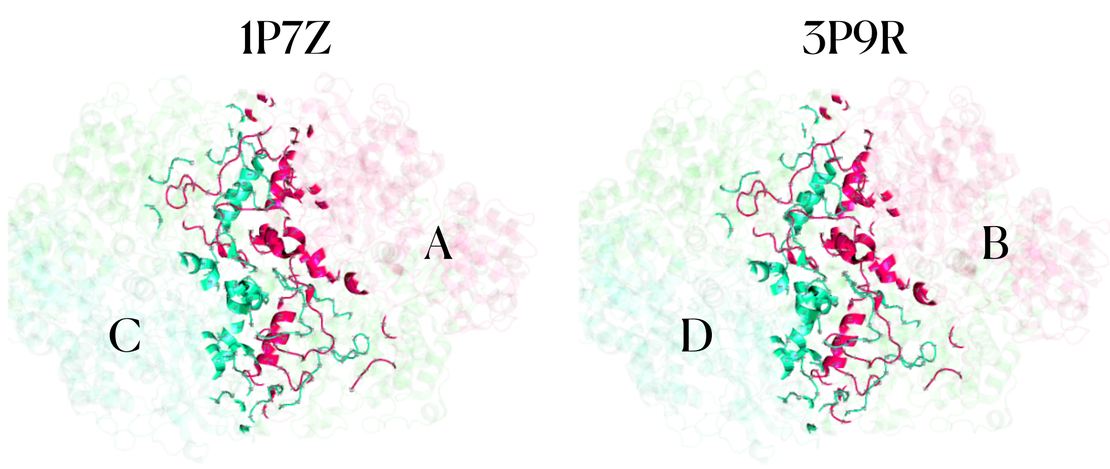
Data leakage problem



Percentage of test inerfaces that have a near duplicate in the training data (y-axis) using randomized 90%/10% splits of 50K PPIs from PDB (x-axis)

Method

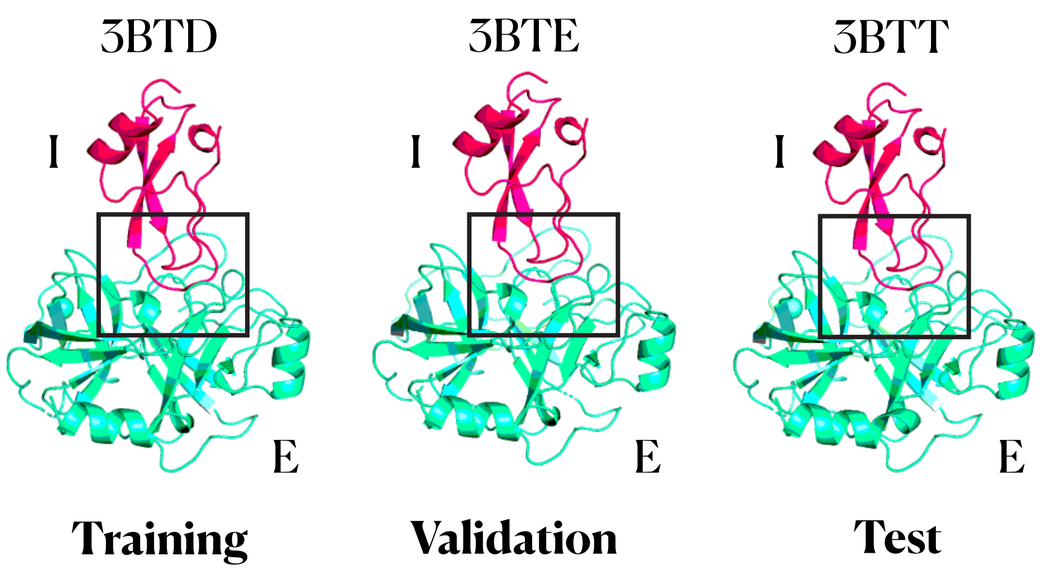
- We use the **iDist** algorithm [1] for large-scale search of test protein–protein interactions (PPIs) with near-duplicate interfaces in training data
- **iDist** is >**100x faster** than PPI alignment methods and finds same near duplicates with **99% precision** and **97% recall**



Example of a near duplicate detected with iDist

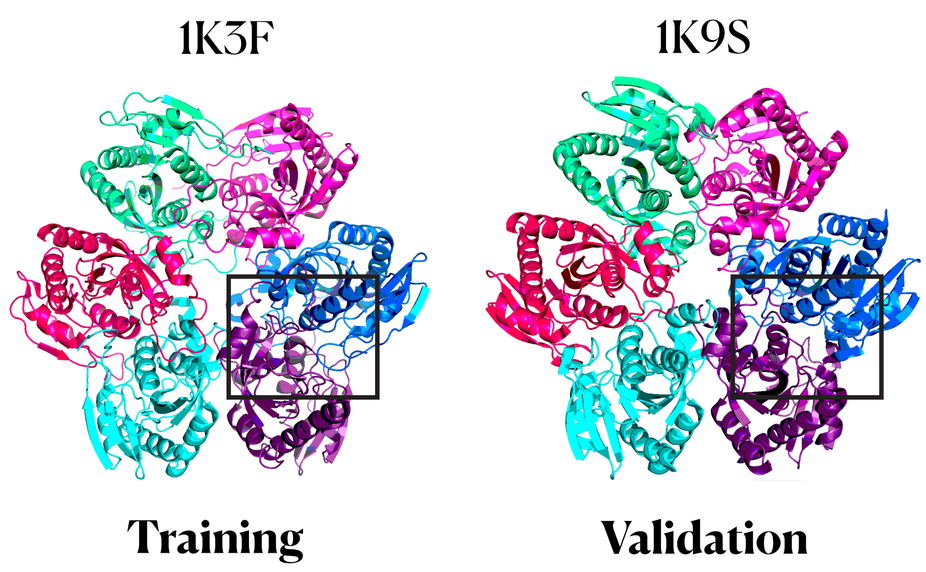
Leakage in existing datasets

- **61% data leakage in ICLR 2023 split of SKEMPI v2.0**, a standard dataset for protein–protein interaction (PPI) design
- → Splitting by PDB codes (and metadata in general) is not enough



Single-point mutants with near-duplicate interfaces in different PDB entries

- **53% data leakage in ICLR 2023 split of DIPS**, a standard dataset for PPI docking and protein interface prediction
- → Splitting by protein family (and sequence similarity in general) is not enough



Near-duplicate interfaces in the homooligomers where proteins have only 26.5% sequence similarity

Recommendations

1. Use 3D interface similarity as the standard criterion for splitting protein interactions
2. Thoroughly review the information provided by dataset authors
3. Quantify and report data leakage when there is no control over train-test splits

[1] Bushuiev, et al. "Learning to design protein–protein interactions with enhanced generalization". ICLR 2024