Moving Beyond Memorisation

Resources for Generalisable Protein Complex Prediction

Connected concepts in datasets, benchmarking, and training

plinder

The Protein-Ligand Interaction Dataset and Evaluation Resource

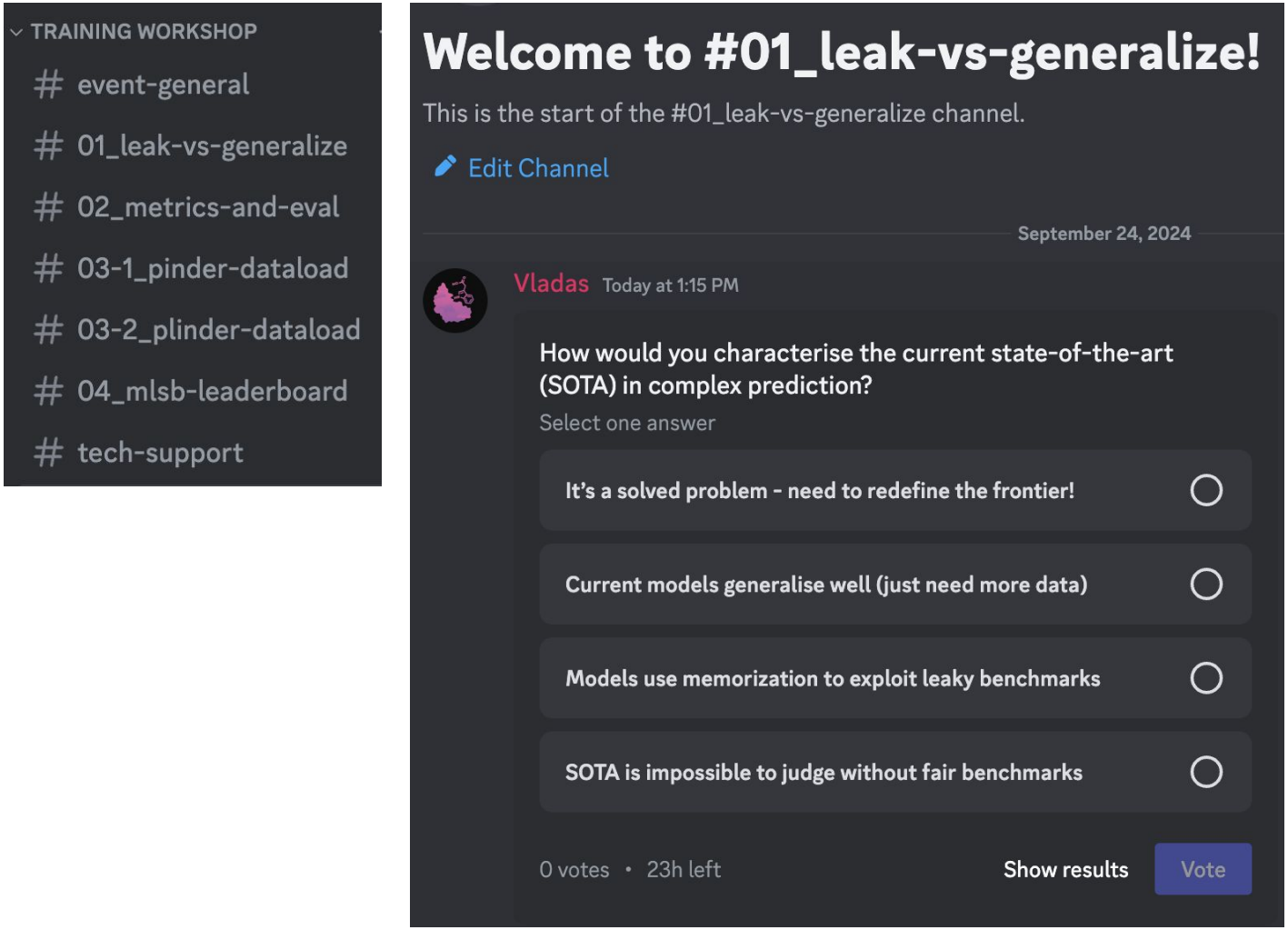
more data
more diversity
more flexible

pinder

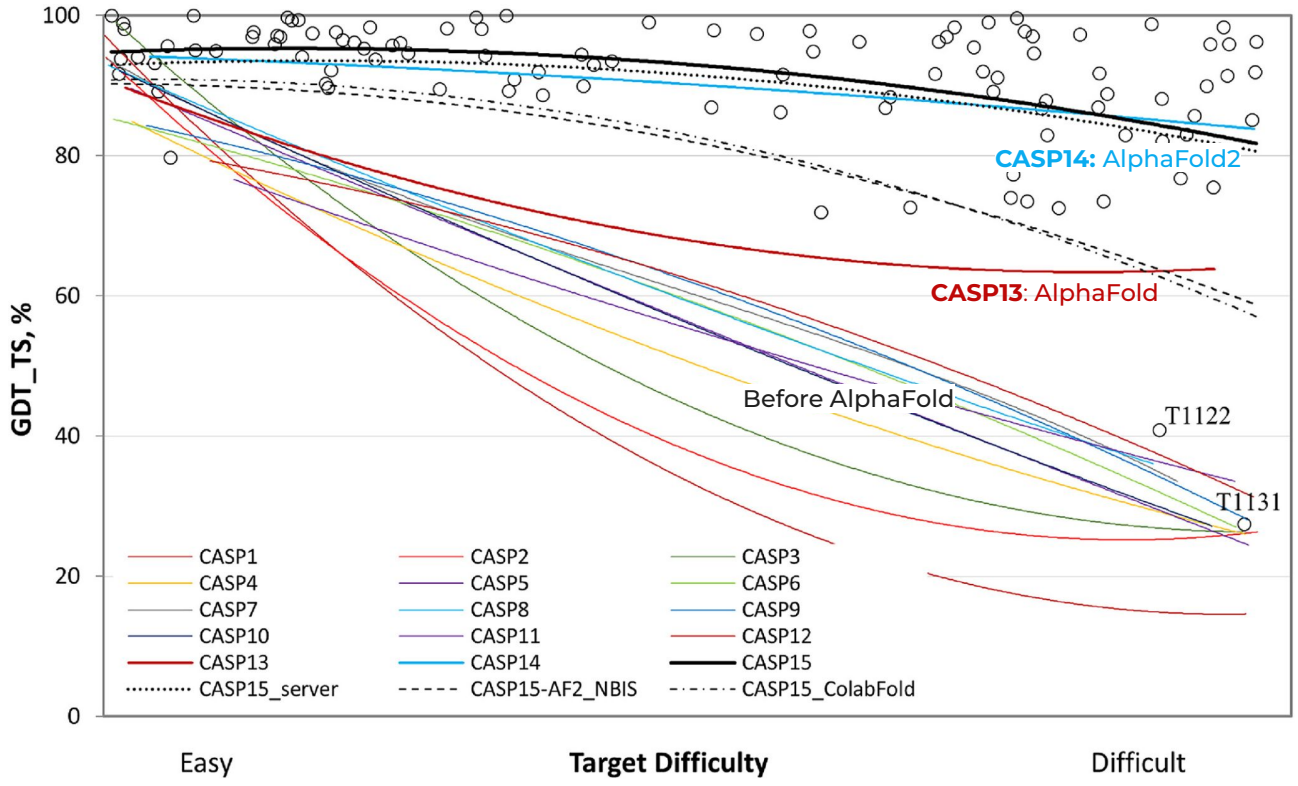
The Protein Interaction Dataset and Evaluation Resource

higher quality
less leakage
better evaluation

2024-09-24


TRAINING WORKSHOP
event-general
01_leak-vs-generalize
02_metrics-and-eval
03-1_pinder-dataload
03-2_plinder-dataload
04_mlsb-leaderboard
tech-support
Welcome to #01_leak-vs-generalize!
This is the start of the #01_leak-vs-generalize channel.
Edit Channel
September 24, 2024
Vladas Today at 1:15 PM
How would you characterise the current state-of-the-art (SOTA) in complex prediction?
Select one answer
It's a solved problem - need to redefine the frontier!
Current models generalise well (just need more data)
Models use memorization to exploit leaky benchmarks
SOTA is impossible to judge without fair benchmarks
0 votes • 23h left
Show results
Vote

Breakthroughs require generalisation

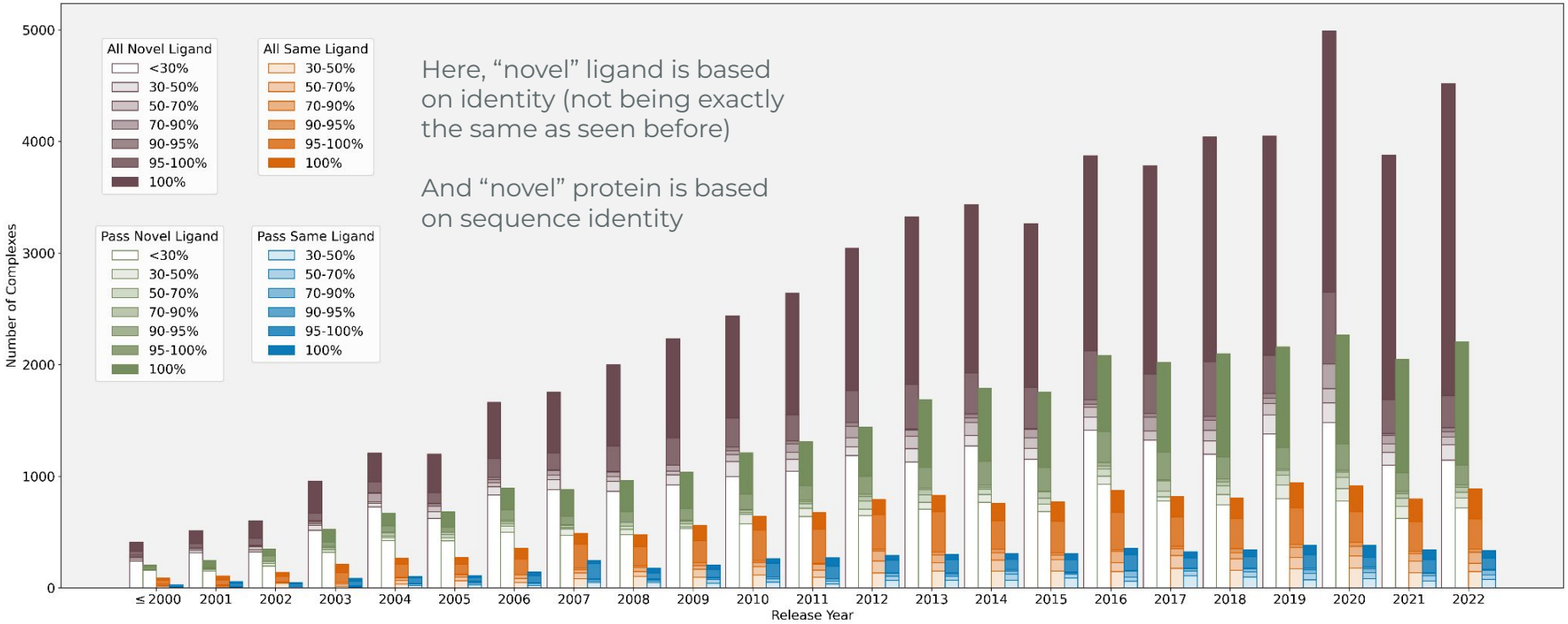


AlphaFold2 was a breakthrough due to "**Difficult**" not "Easy"

(in <u>monomer</u> prediction)

Kryshtafovych, Andriy, et al. *Proteins: Structure, Function, and Bioinformatics* 91.12 (2023): 1539-1549.

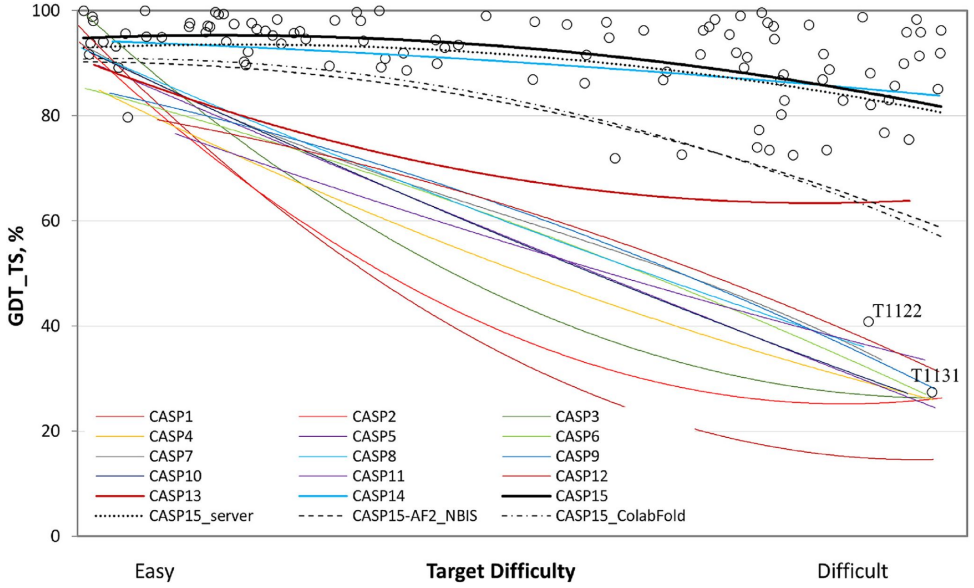
“New” PDBs != Novel complexes



Protein-ligand complexes released per year compared to everything before

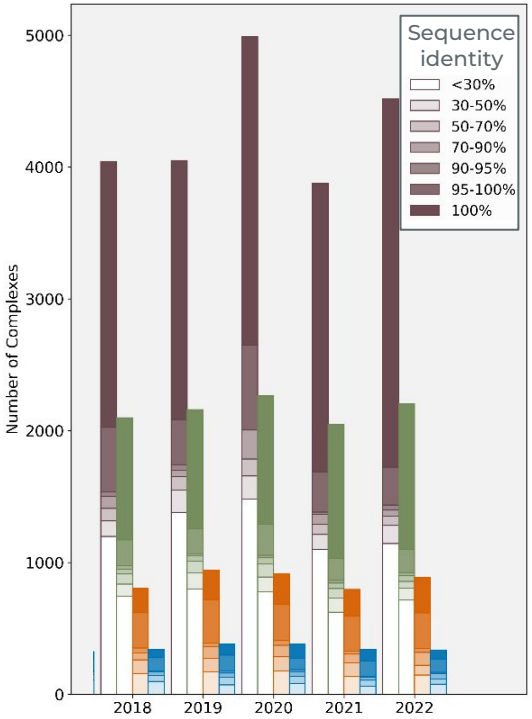
Leemann, Michèle, et al. *Proteins: Structure, Function, and Bioinformatics* 91.12 (2023): 1912-1924.

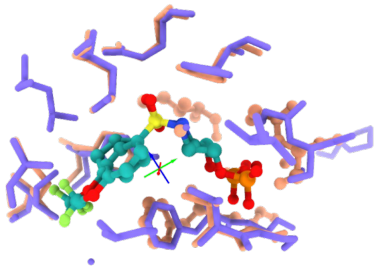
Time splits overestimate performance



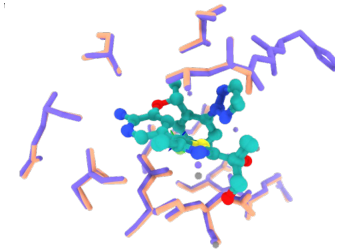
AF2 was a breakthrough
due to “Difficult” not “Easy”
(in monomer prediction)

“New” PDBs != Novel complexes





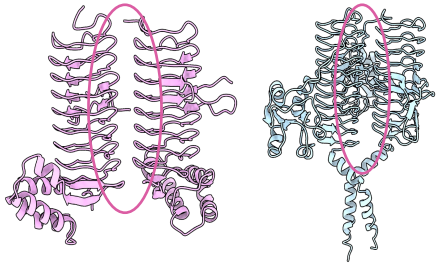
7L03 | PoseBusters
2CLI | PDBBind



7R7R | PoseBusters
4CD0 | PDBBind

“Novelty” definition depends on the task

- Generalisation difficulty
  = Interaction dissimilarity

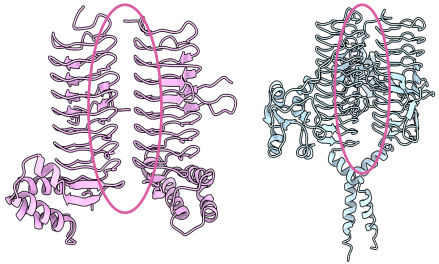


7OJQ
ProteinFlow
Train Split

2IUA
ProteinFlow
Test Split

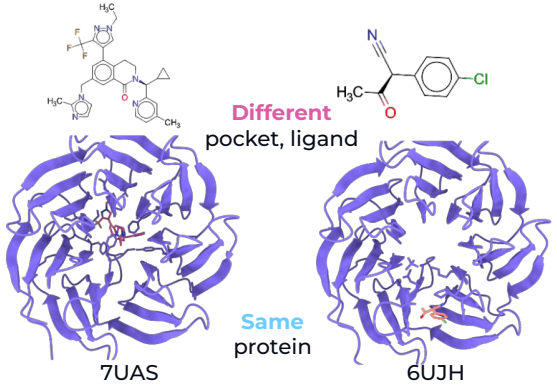
“Novelty” definition depends on the task

- Generalisation difficulty
  = Interaction dissimilarity



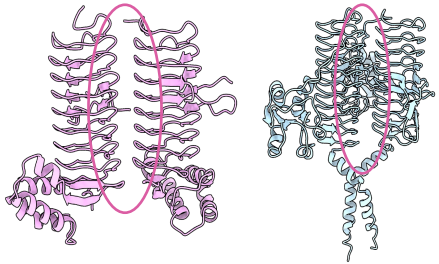
7OJQ
ProteinFlow
Train Split

2IUA
ProteinFlow
Test Split



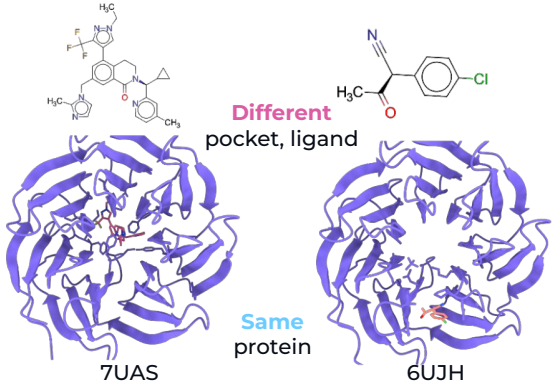
“Novelty” definition depends on the task

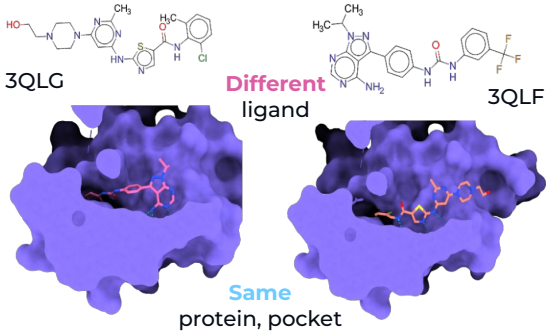
- Generalisation difficulty
  = Interaction dissimilarity

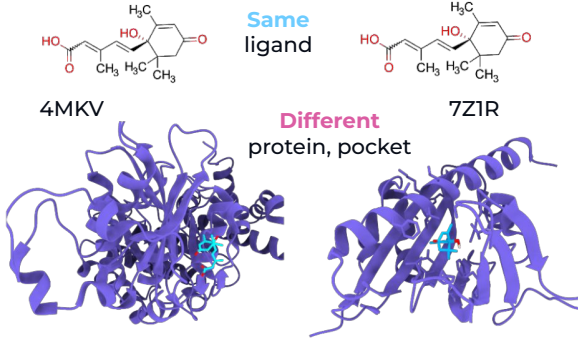


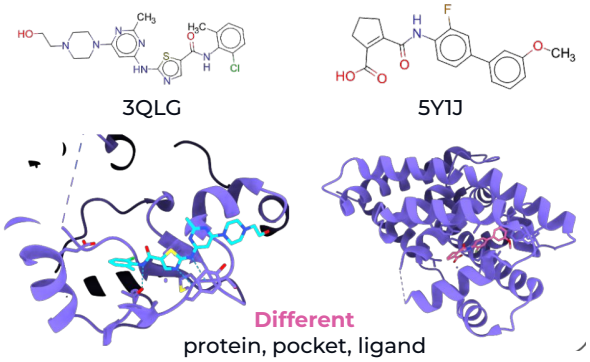
7OJQ
ProteinFlow
Train Split

2IUA
ProteinFlow
Test Split



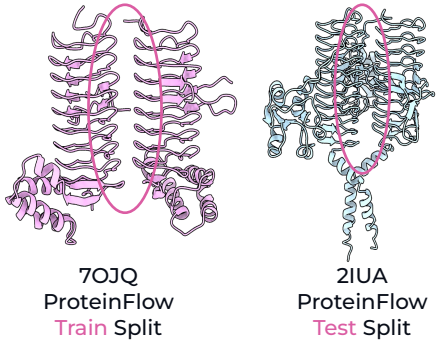


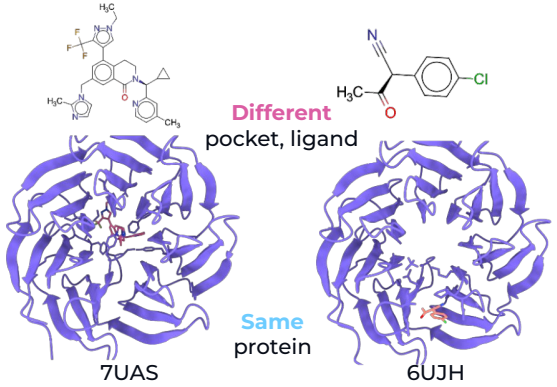


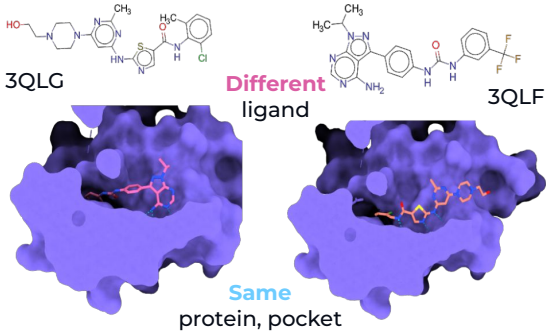


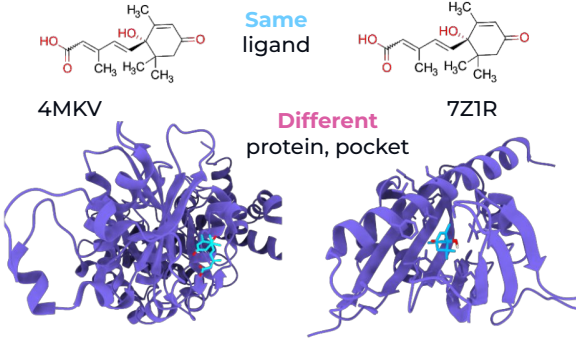
“Novelty” definition depends on the task

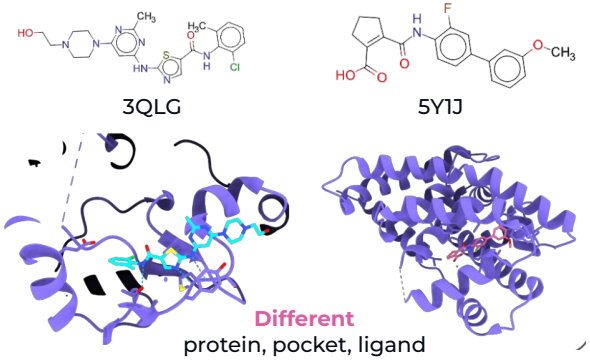
- Generalisation difficulty
  = Interaction dissimilarity
- P(L)INDER *measures, clusters, and splits* on interaction similarities





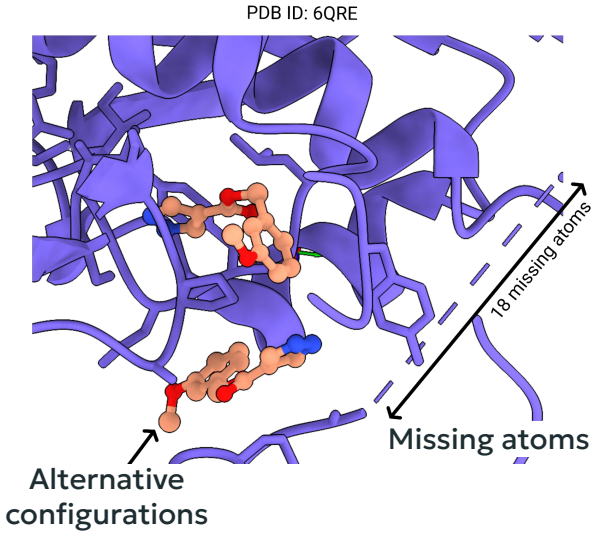






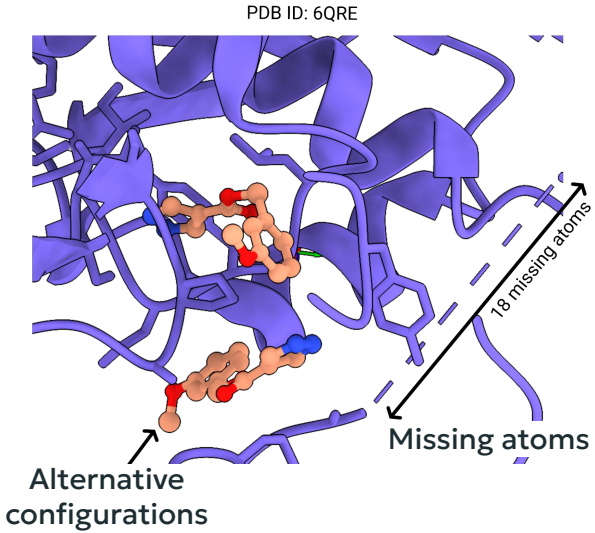
Why quality matters

- Experimental structures vary in quality

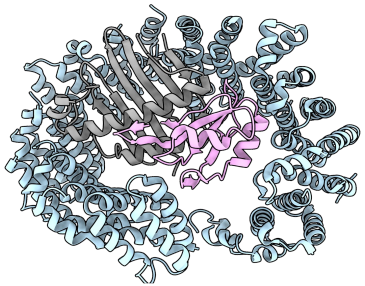


Why quality matters

- Experimental structures vary in quality

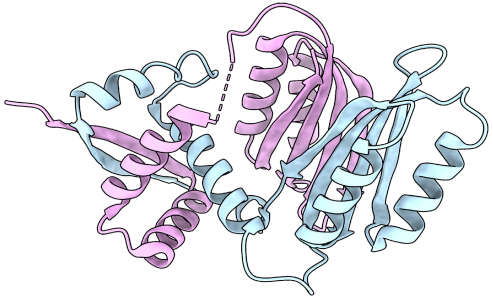


Spurious interactions



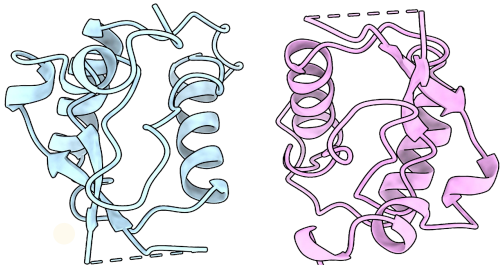
2X1G | DIPS-Plus
Interaction probability = 0.09

Detached parts



2WVB | ProteinFlow

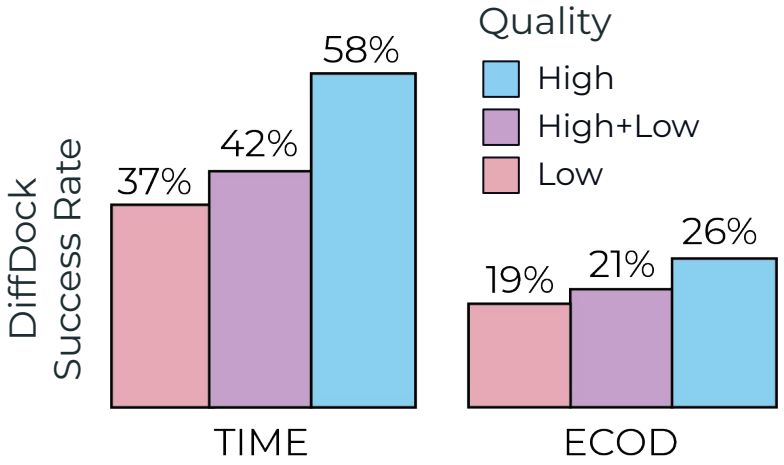
Interfacial gaps

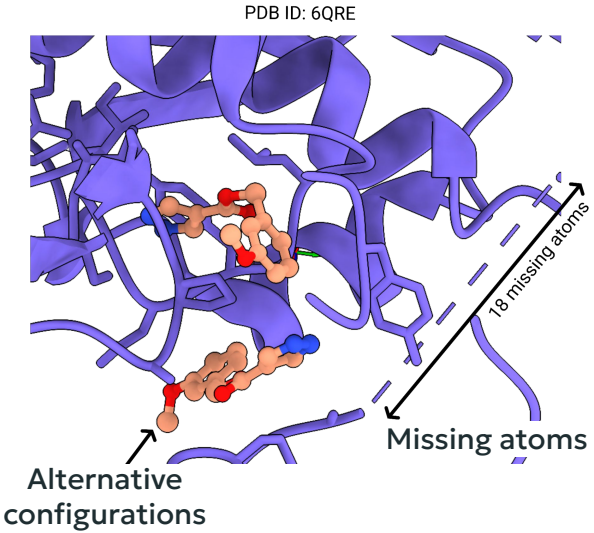


1EFN | DB5

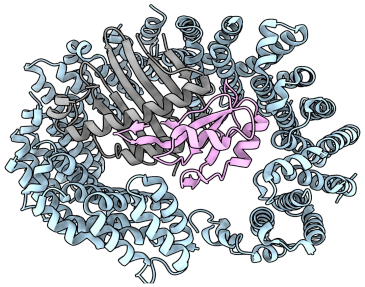
Why quality matters

- Experimental structures vary in quality
- Need **reliable** ground truth for evaluation



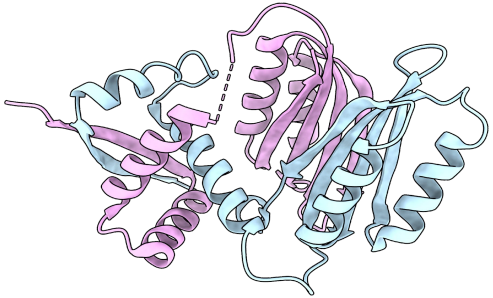


Spurious interactions



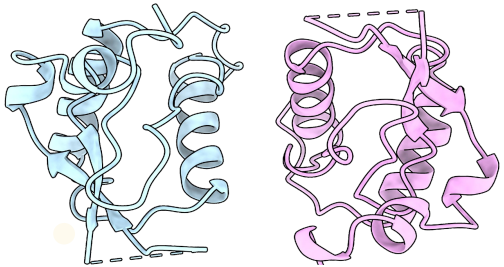
2X1G | DIPS-Plus
Interaction probability = 0.09

Detached parts



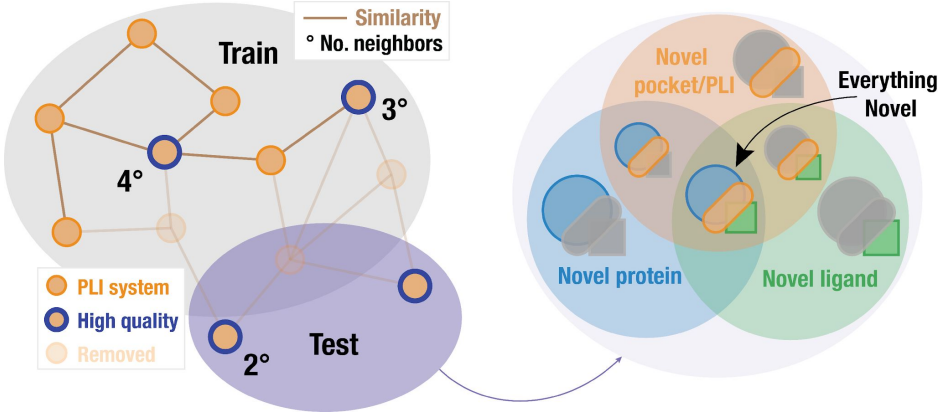
2WVB | ProteinFlow

Interfacial gaps



1EFN | DB5

Split into relevant train/val/test sets



Optimize for

- Test set **quality**
- Test set **diversity**
- Minimal information **leakage**
- Training set **size** and **diversity**

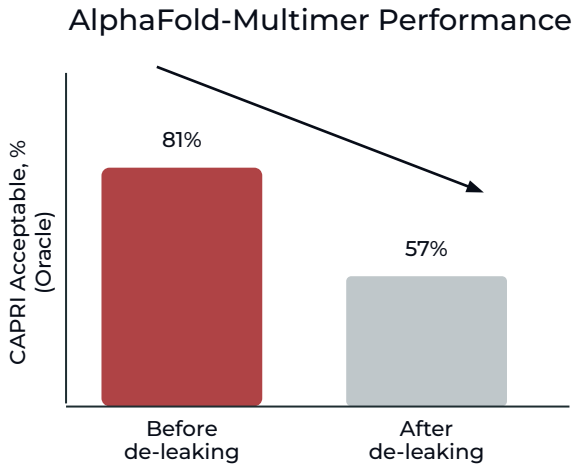
Stratify into

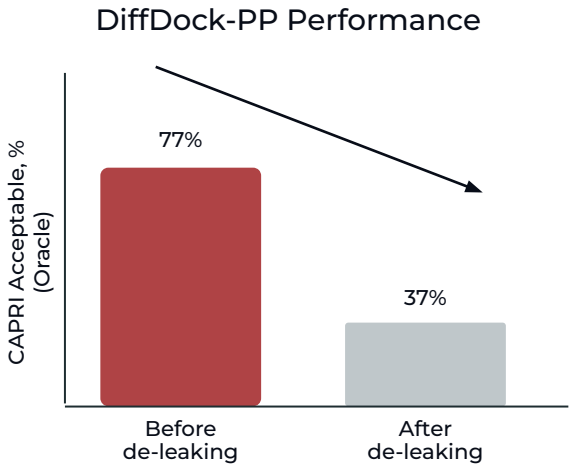
- Novel **protein**
- Novel **ligand**
- Novel **pocket**/**Xns**
- **Everything** novel

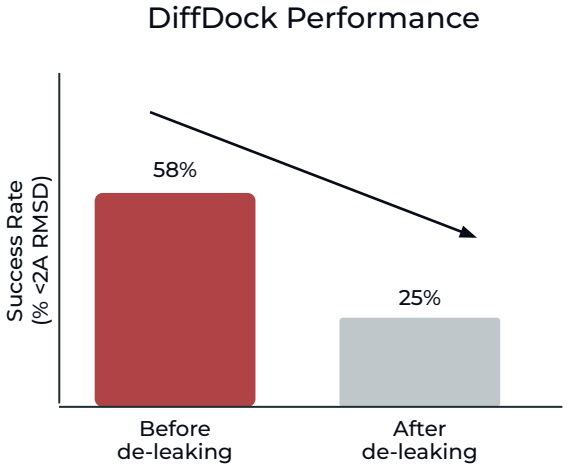
```
function SPLITTING(S, C, G, m, M)
    proto_test ← ∅
    for s ∈ S do
        if pass_quality (s) then
            N_s ← ∅
            for g ∈ [1, |G⟧ do
                leaked ← neighbors (s, G_g)
                N_s ← N_s ∪ leaked
            if m < |N_s| < M then
                proto_test.insert (s)
    ▷ Sort s ∈ proto_test by |N_s|
    test ← ∅
    for c ∈ C do
        test ← test ∪ upto 1 from proto_test for c
    train ← S
    for s ∈ test do
        train ← train \ N_s
```

Systems
Clusters
Graphs
min & **M**ax. removal

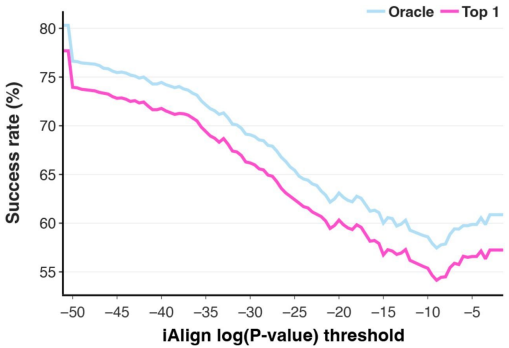
Lack of generalisation ~= memorisation

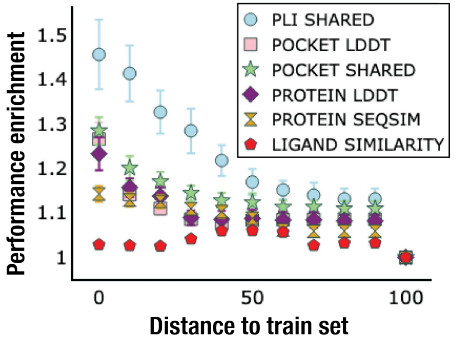




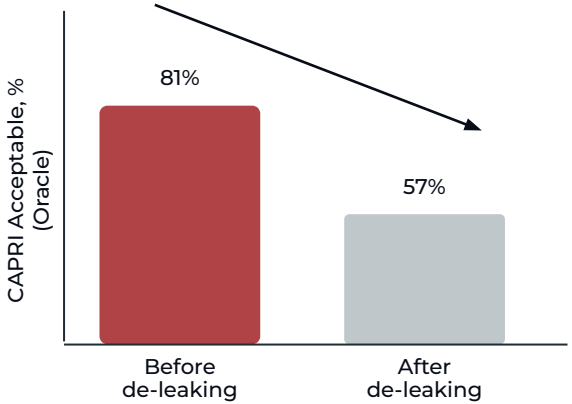


Lack of generalisation ~= memorisation

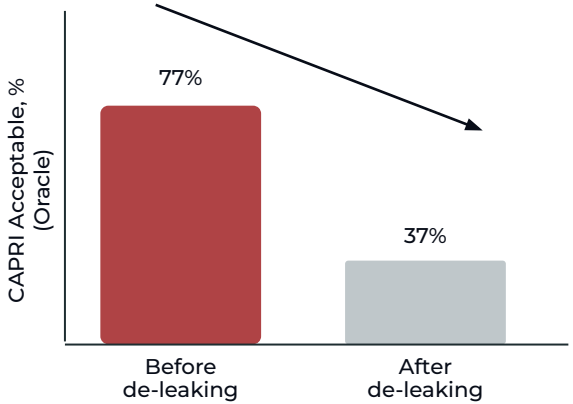




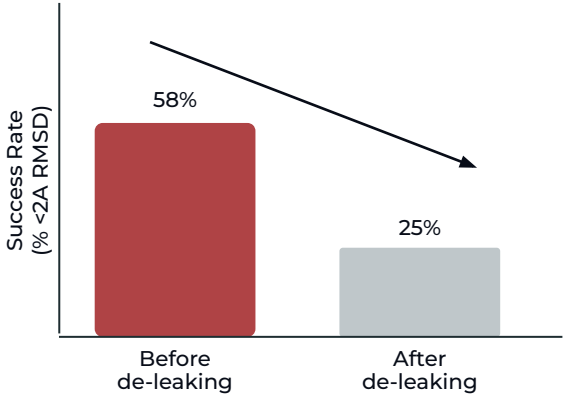
AlphaFold-Multimer Performance



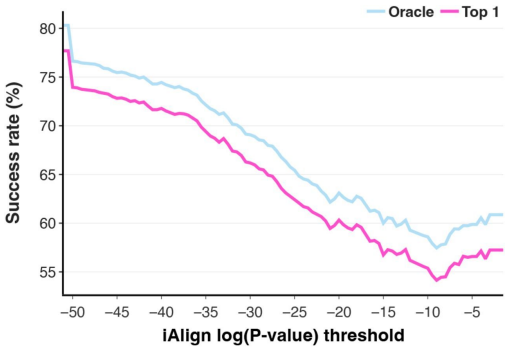
DiffDock-PP Performance



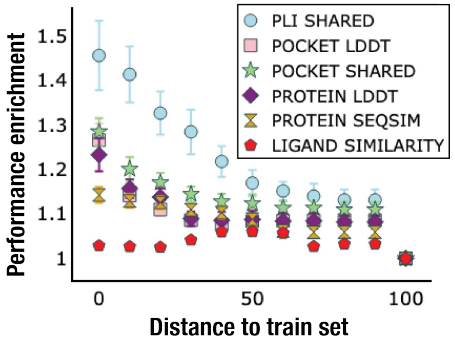
DiffDock Performance



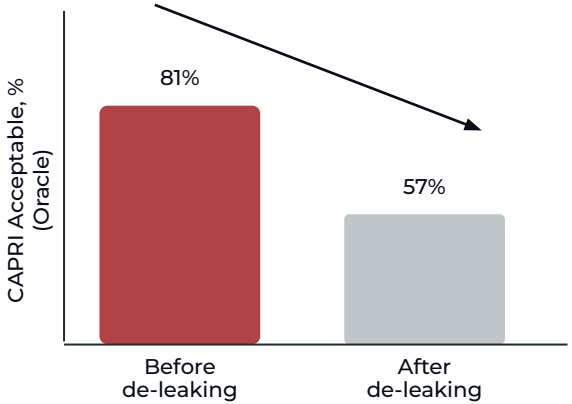
Lack of generalisation ~= memorisation



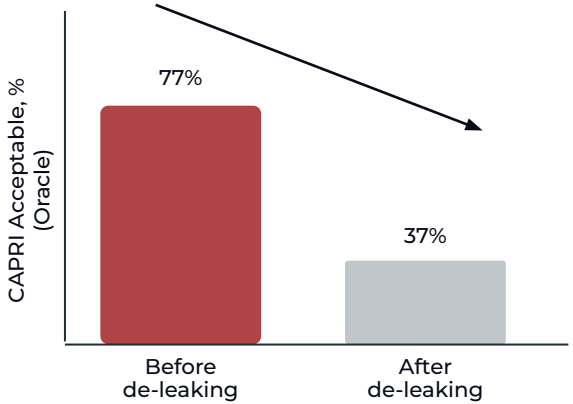
Complex prediction still needs a breakthrough



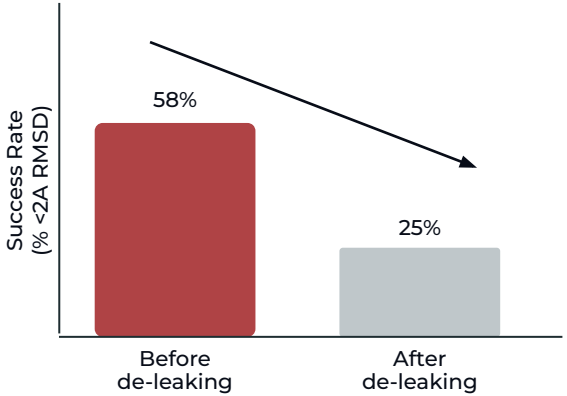
AlphaFold-Multimer Performance



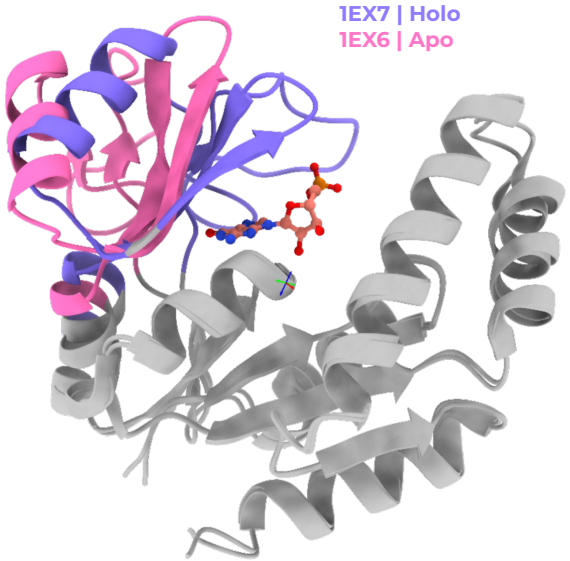
DiffDock-PP Performance

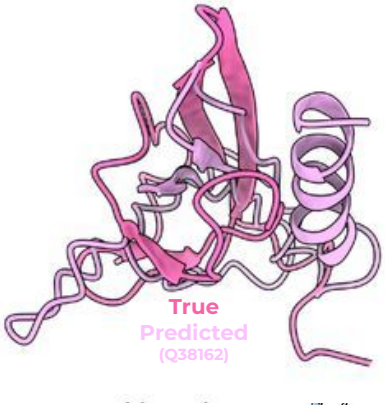


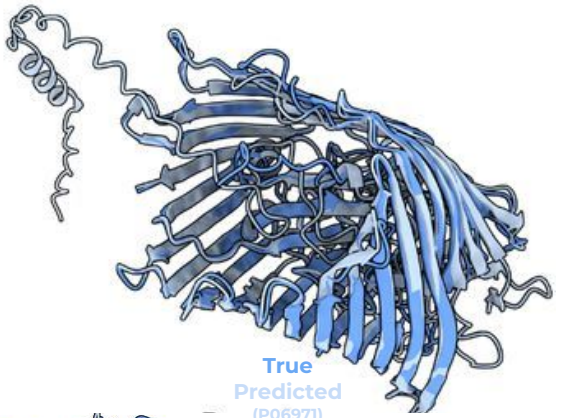
DiffDock Performance

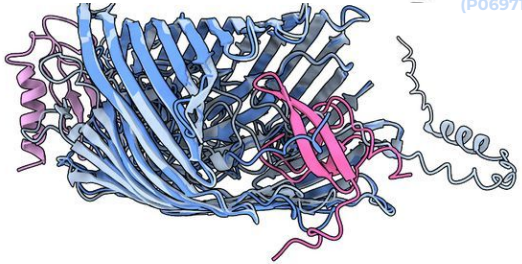


Realistic inference != Rigid re-docking

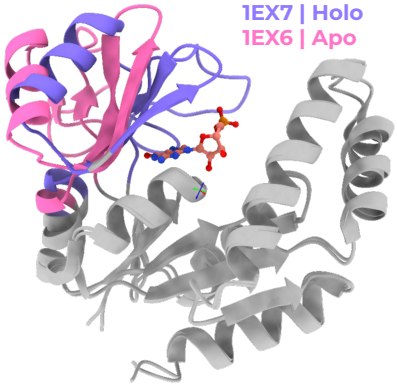


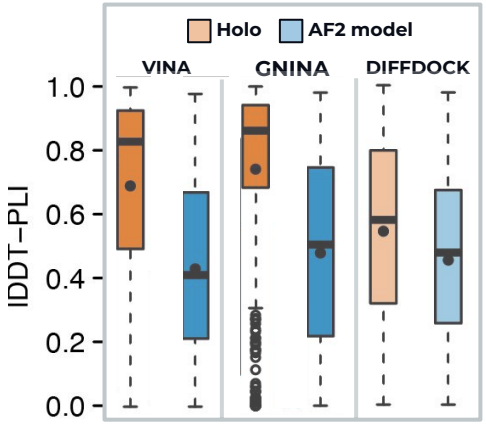


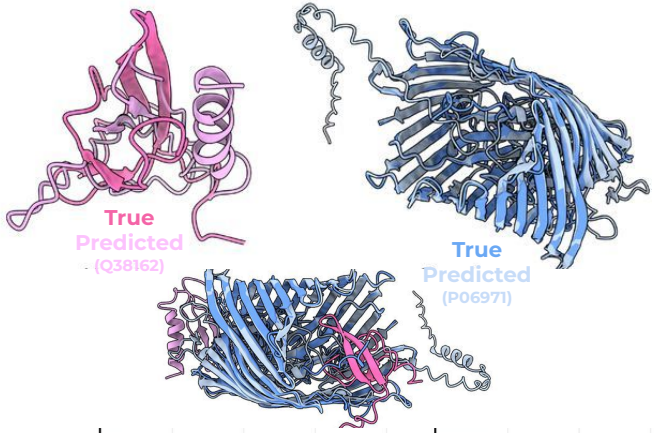


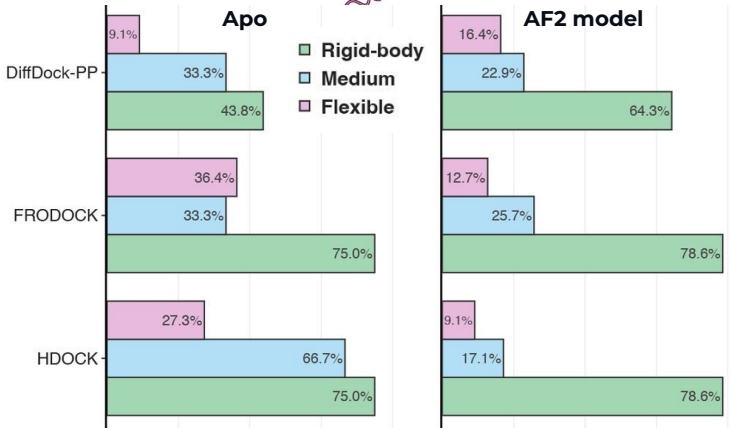


Realistic inference != Rigid re-docking

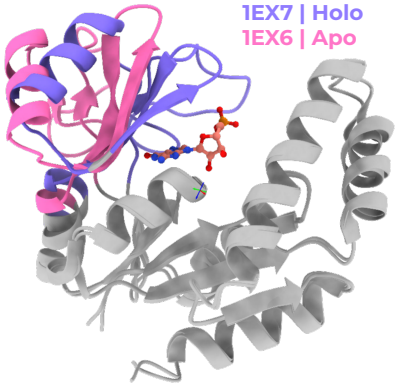






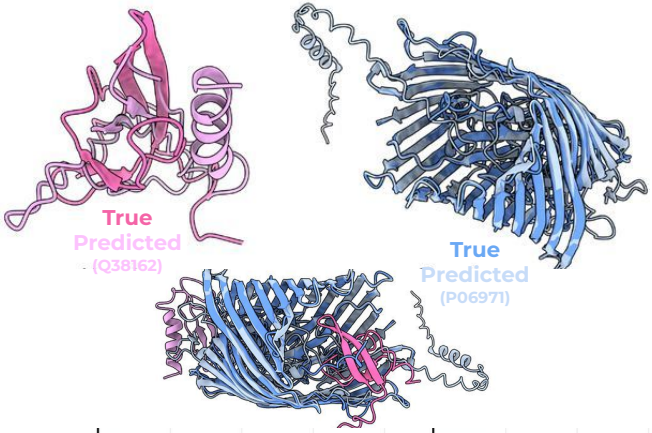


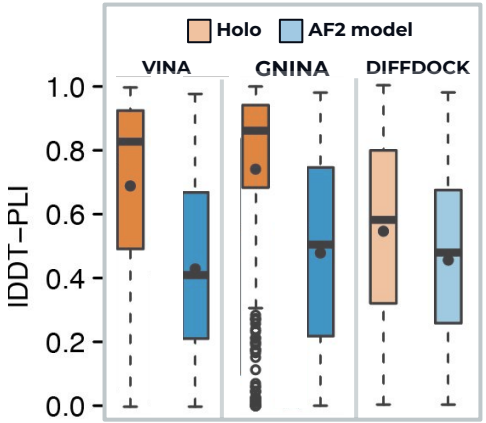
Realistic inference != Rigid re-docking

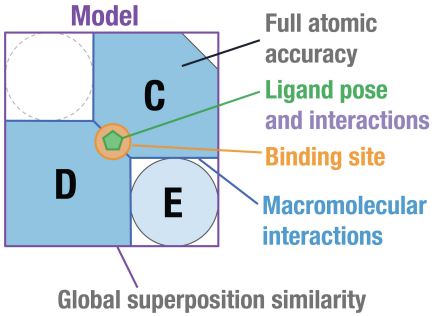


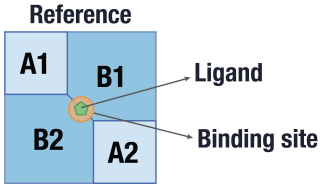
Need diverse and flexible metrics

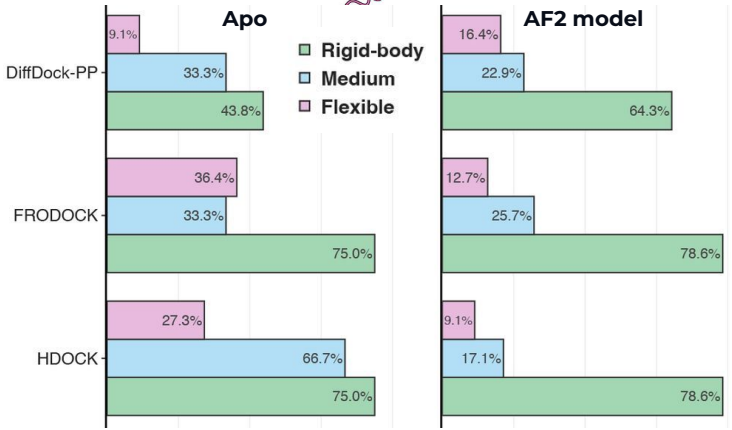
- Beyond **rigid placement** RMSD
- Evaluate **ligand** poses, **pocket** residues and **interactions**











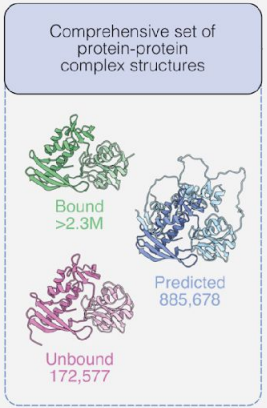
Automated and reproducible end-to-end workflow

pinder

Ingest

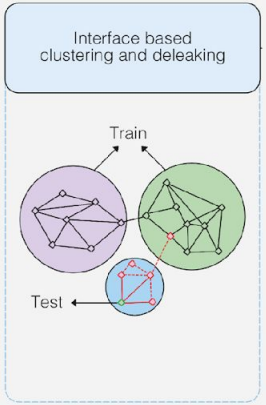
- Fully automated and reproducible pipeline
- 2,319,564 systems with 9,430 unique ECOD domain pairs across 6,529 families
- 100+ annotations
- Interface quality assessment with 10+ metrics

Expand



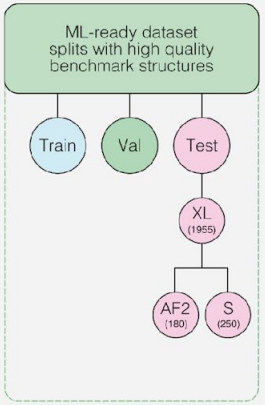
- Paired unbound and AlphaFold2 predicted monomers
- PPI interfaces stratified by flexibility

Split



- FoldSeek and MMSeqs based interface similarity comparison followed by transitive graph-clustering and additional deleaking via iAlign
- Extensive orthogonal leakage validation via ECOD-, PFAM-overlap & other metrics
- Maximum val and test set quality with minimal leakage

Evaluate

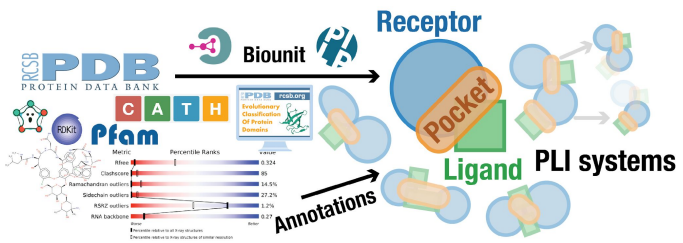


- Large split (XL) with smaller subset (S)
- AF2-training cutoff & interface structural deleaked XL subset (AF2)
- Automated evaluation harness with 38 CASP-CAPRI compatible metrics and leaderboard included
- Holo/Apo/Predicted input leaderboards across protein flexibility levels (easy, medium, hard)

Automated and reproducible end-to-end workflow

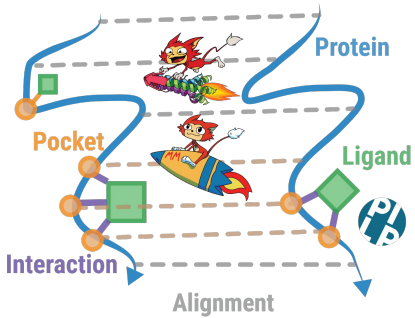
plinder

Ingest



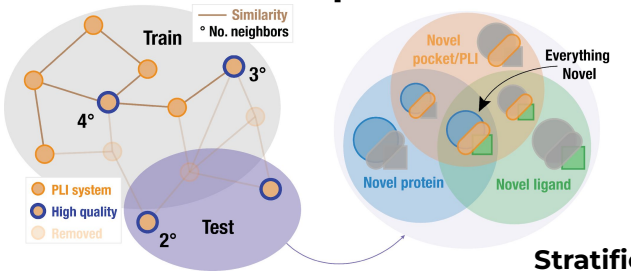
- Collect **all** PLIs in the PDB
- **Pair** *apo* and AF2 structures
- Annotate ***everything***
- Keep up with PDB **updates**

Compare



Sequence & ***structure*** similarity at **protein**, **pocket**, **ligand**, and **interaction** levels

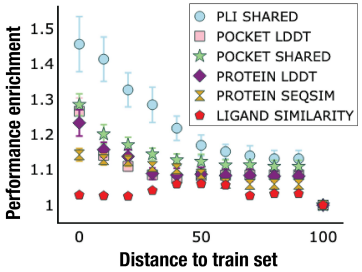
Split



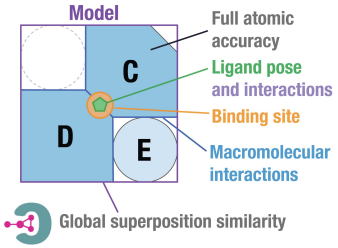
Maximizing test set **quality and diversity** while *minimizing* information **leakage**

Stratification for different tasks

Evaluate



Resources to **inspect biases** in your models



Comprehensive **evaluation** harness

Better datasets & splits = Better models

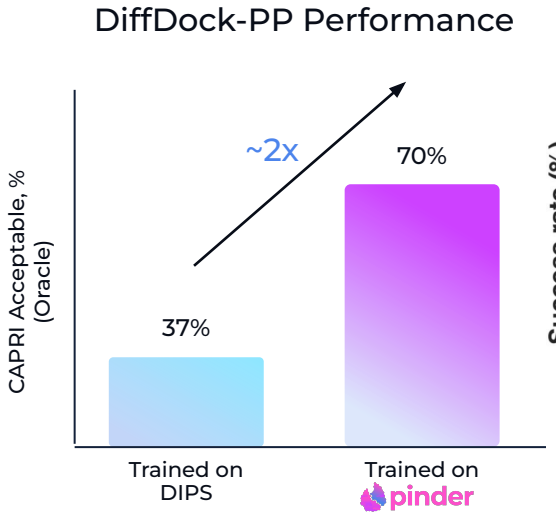
- **Bigger** more **diverse** training sets
- Assess **val** to stop training at the **right** time
- **Augment** training with **conformational change**
- Higher **quality** and **diverse** test sets
- **Realistic** inference with meaningful **evaluation**

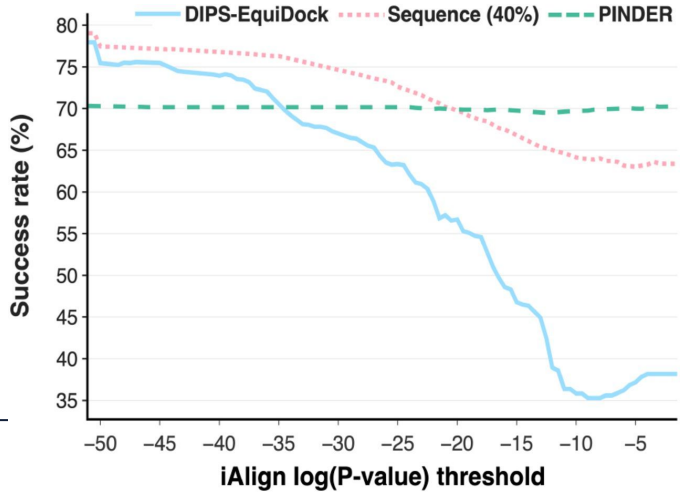
plinder
The Protein-Ligand Interaction Dataset and Evaluation Resource

more data
more diversity
more flexible

pinder
The Protein Interaction Dataset and Evaluation Resource

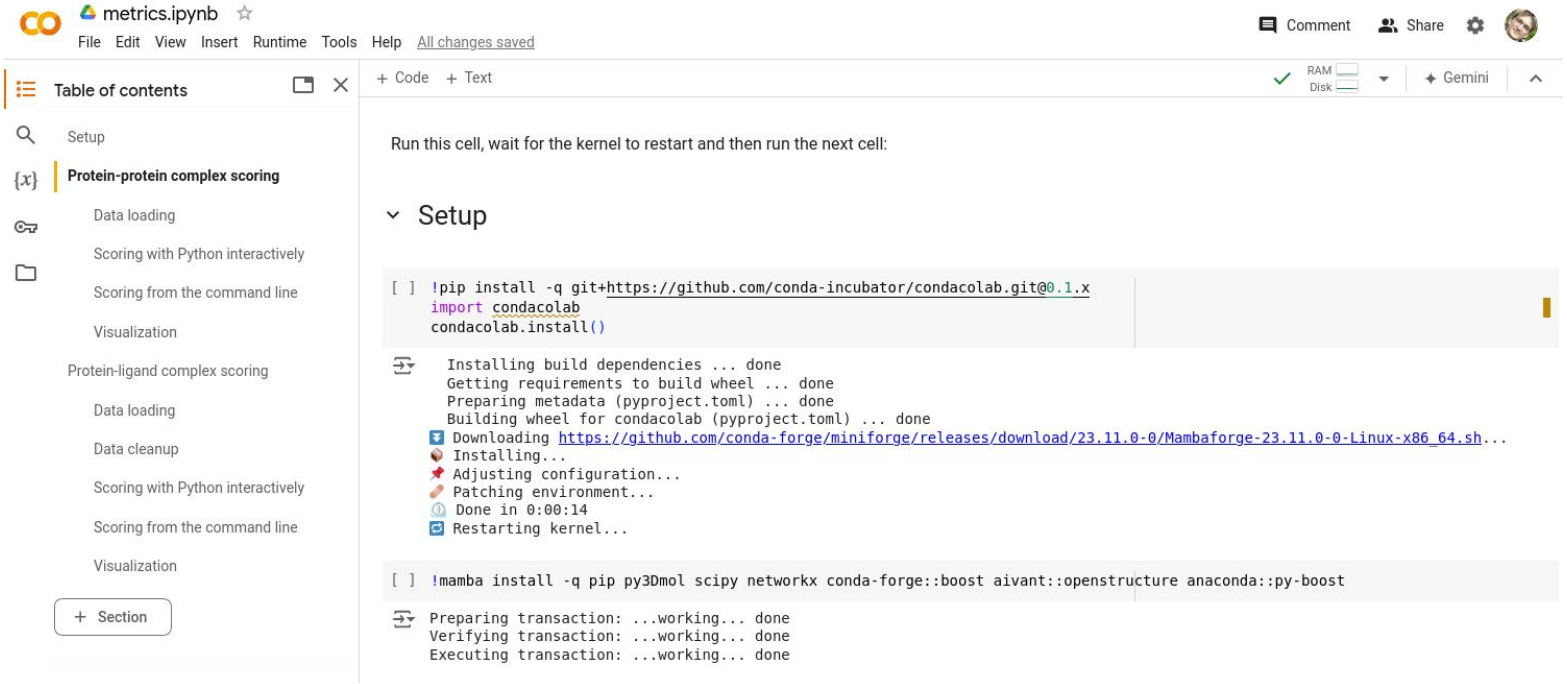
higher quality
less leakage
better evaluation





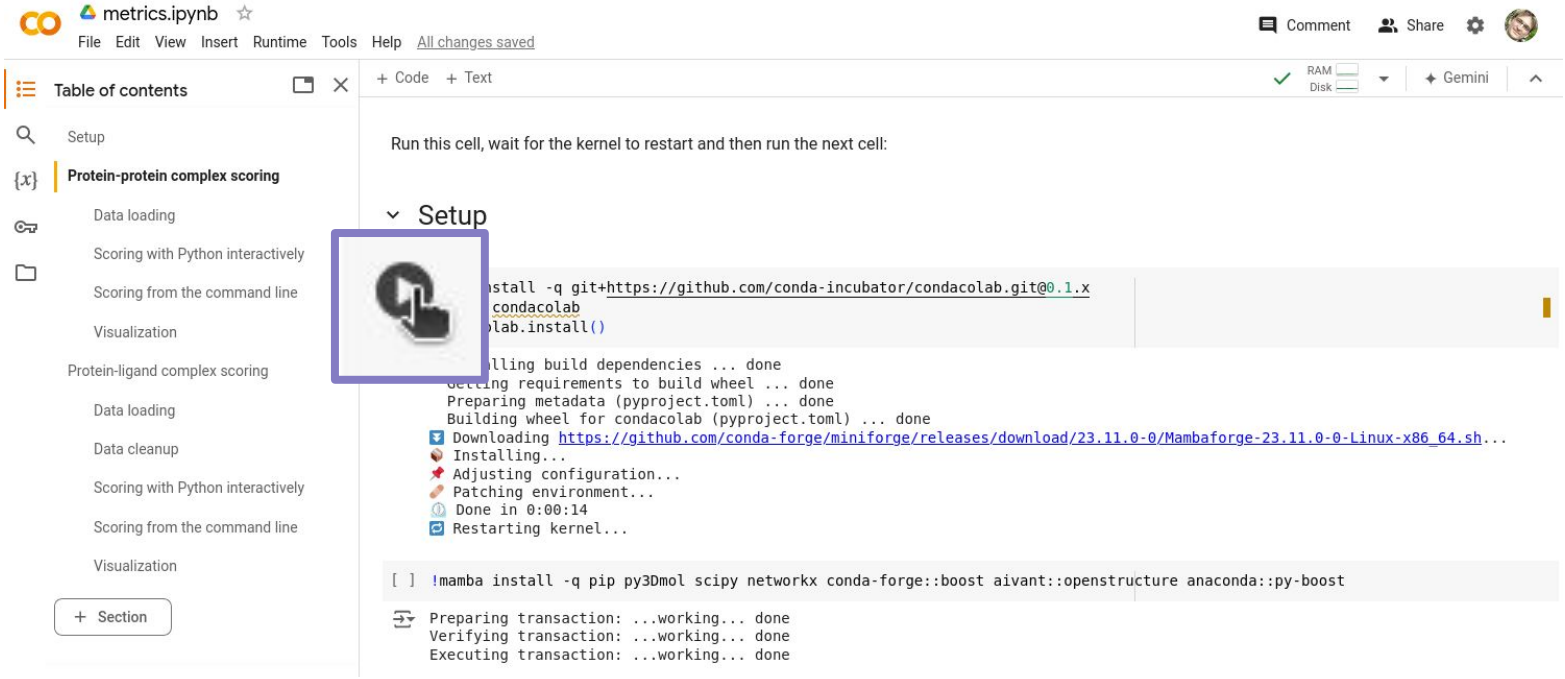
Hands on

Links available in the Discord chat and from GitHub



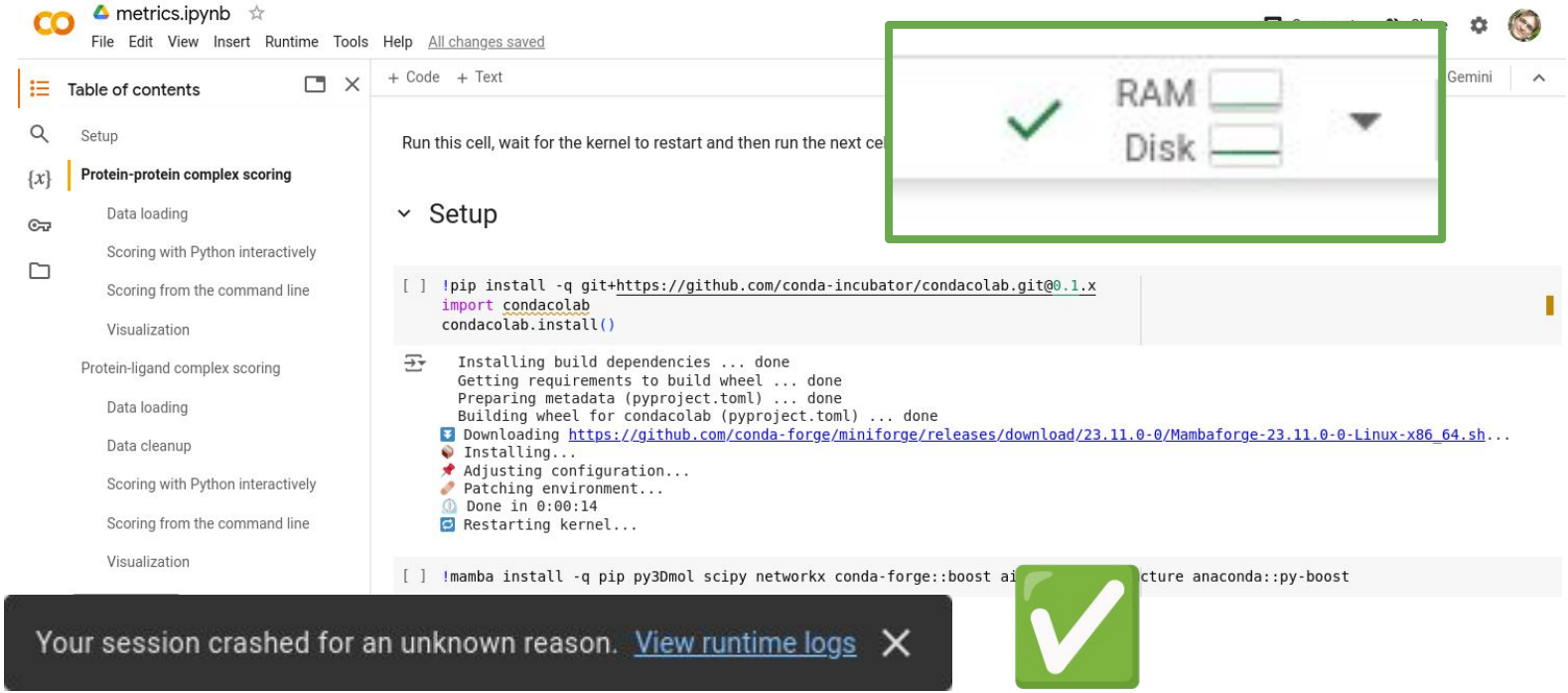
Hands on

Links available in the Discord chat and from GitHub



Hands on

Links available in the Discord chat and from GitHub



Hands on

Links available in the Discord chat and from GitHub