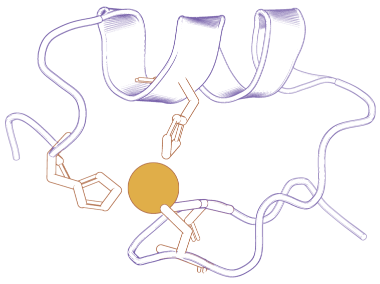
plinder

The Protein-Ligand Interactions Dataset and Evaluation Resource

Janani Durairaj*[1,2] , Yusuf Adeshina*[3], Zhonglin Cao[4], Xuejin Zhang[3], Vladas Oleinikovas[3], Thomas Duignan[3], Zachary McClure[4], Xavier Robin[1,2], Emanuele Rossi[3], Guoqing Zhou[4], Srimukh Veccham[4], Clemens Isert[3], Yuxing Peng[4], Prabindh Sundareson[4], Mehmet Akdel[3], Gabriele Corso[5], Hannes Stärk[5], Zachary Carpenter[3], Michael Bronstein[3,6], Emine Kucukbenli[4], Torsten Schwede[1,2], Luca Naef[3]

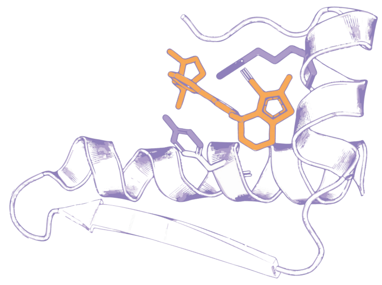
[1]Biozentrum University of Basel; [2]SIB Swiss Institute of Bioinformatics; [3]VantAI, [4]NVIDIA, [5]MIT CSAIL, [6]Oxford University

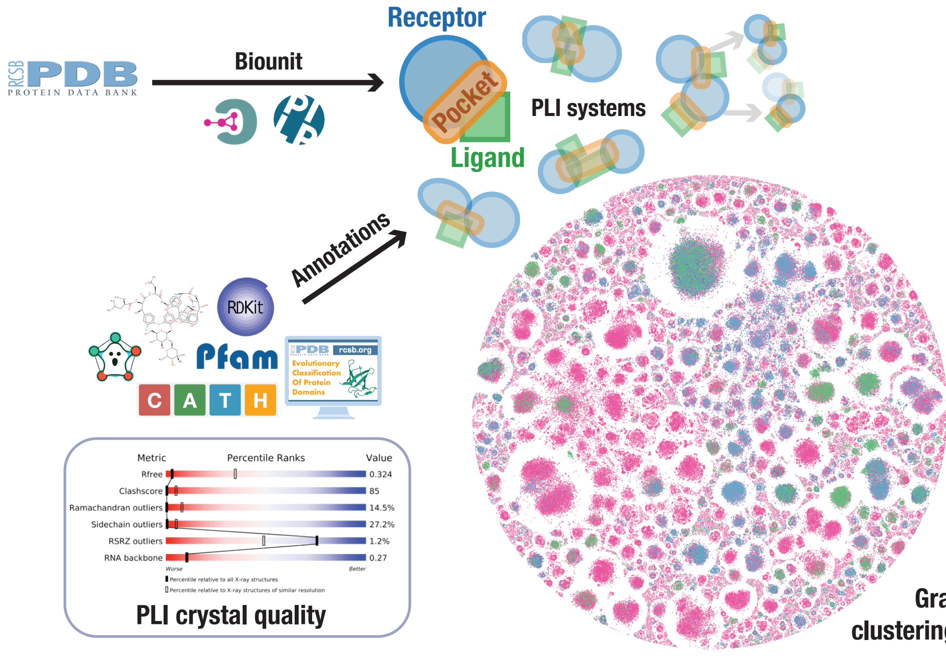
The effectiveness of protein-ligand complex prediction methods depends largely on the quality of the training and evaluation dataset



To create a high quality and reliable dataset, one needs to consider:

- ***Training set size and diversity*** to learn the underlying patterns instead of simple memorization
- ***Low information leakage*** between train and test to assess generalization and avoid overfitting
- ***Test set quality*** to avoid comparing prediction results to unreliable ground truth
- ***Test set diversity*** to showcase performance across a range of complex types and use-cases
- ***Realistic inference scenarios*** to move beyond "re-docking"





PLI crystal quality

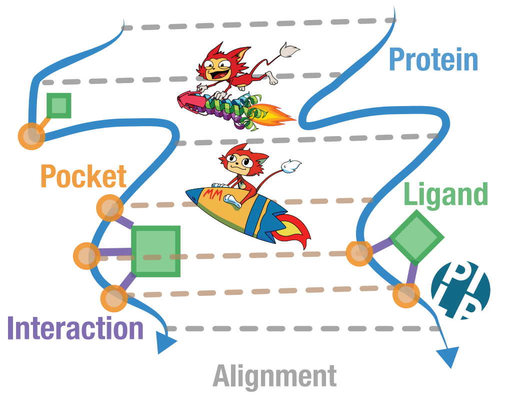
Annotated Dataset

>400,000 PLI systems
with 500+ annotations

	PLINDER	PDBBIND	DOCKGEN
PLI SYSTEMS	449,383	30,337	41,791
PDB IDS	110,791	19,007	16,881
PASS QUALITY	113,498	10,818	19,355
RECEPTORS	74,256	5,425	7,961
SMILES	51,573	15,279	9,174
CCD CODES	46,988	15,064	9,164
CATH	1,641	649	603
SCOP	11,154	2,423	2,817
ECOD	4,458	1,444	1,513
PROTEIN KINASE	297	184	174
KINASE INHIBITORS	48,064	4,682	5,605
APO LINKED	98,473	-	-
AFDB LINKED	205,300	-	-

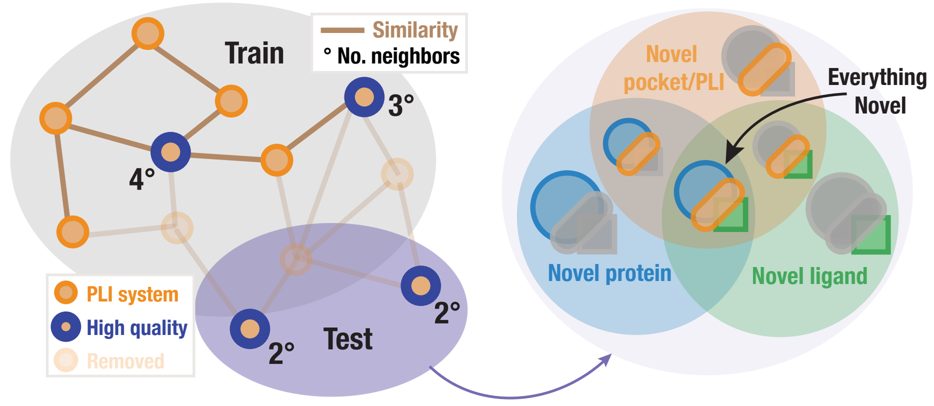
Graph-based similarity clustering across the whole PDB

Protein-Ligand Complex Similarity



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

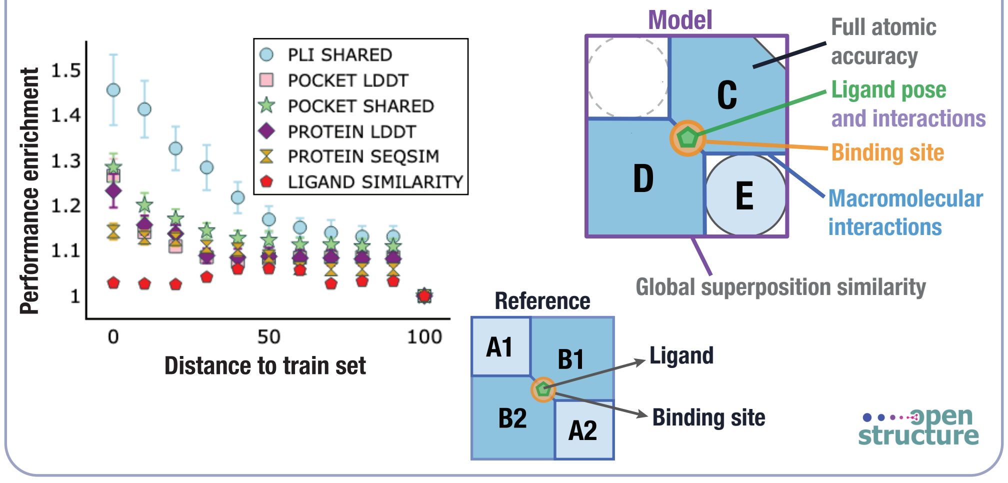
Relevant train/val/test sets



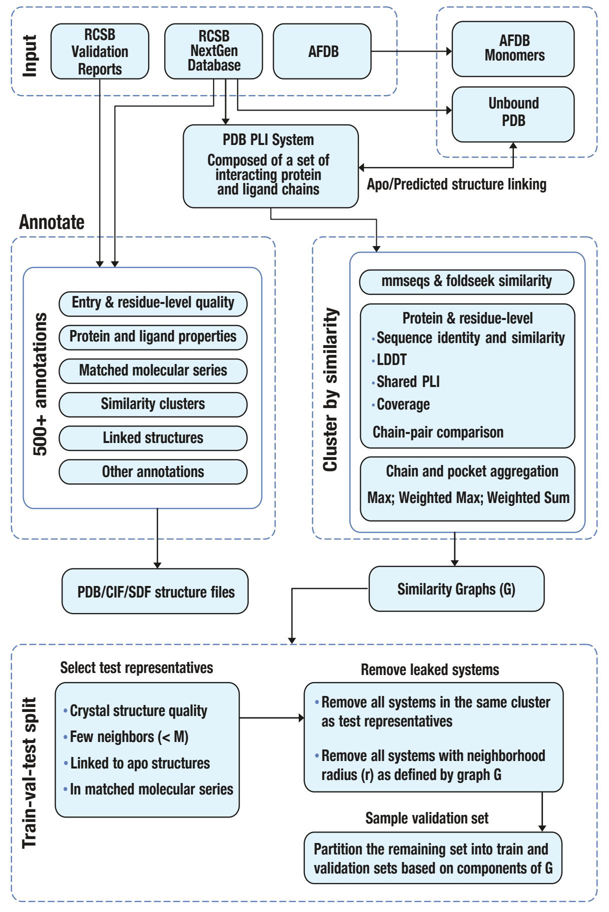
Maximizing test set quality and diversity while minimizing information leakage

Stratification of the test set for different tasks

Comprehensive Accuracy Evaluation



Workflow


Input
RCSB Validation Reports
RCSB NextGen Database
AFDB
AFDB Monomers
Unbound PDB
PDB PLI System
Composed of a set of interacting protein and ligand chains
Apo/Predicted structure linking
Annotate
500+ annotations
Entry & residue-level quality
Protein and ligand properties
Matched molecular series
Similarity clusters
Linked structures
Other annotations
Cluster by similarity
mmseqs & foldseek similarity
Protein & residue-level
·Sequence identity and similarity
·LDDT
·Shared PLI
·Coverage
Chain-pair comparison
Chain and pocket aggregation
Max; Weighted Max; Weighted Sum
PDB/CIF/SDF structure files
Similarity Graphs (G)
Train-val-test split
Select test representatives
•Crystal structure quality
•Few neighbors (< M)
•Linked to apo structures
•In matched molecular series
Remove leaked systems
•Remove all systems in the same cluster as test representatives
•Remove all systems with neighborhood radius (r) as defined by graph G
Sample validation set
Partition the remaining set into train and validation sets based on components of G

Re-training Results

NVIDIA BioNeMo

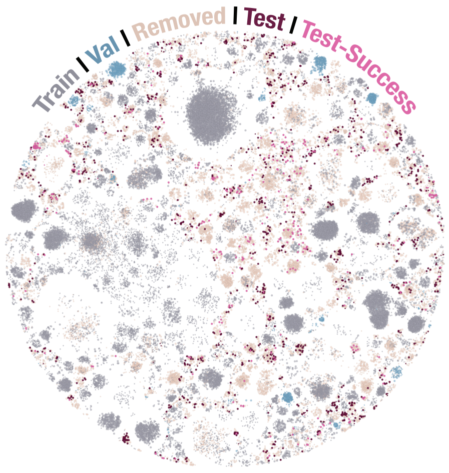
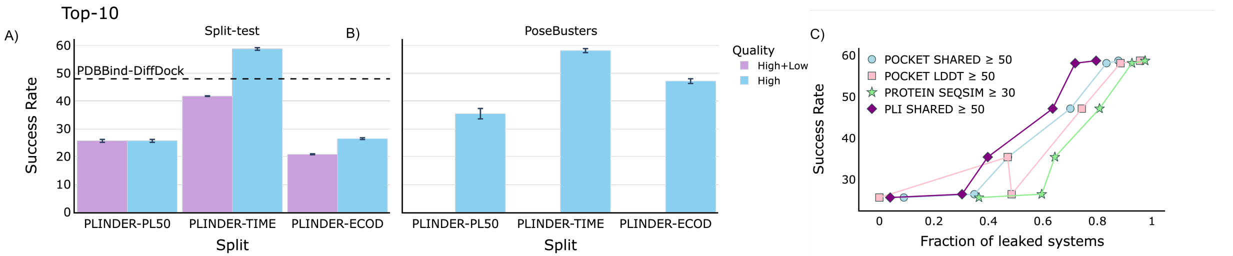


Table 2. Dataset train vs. test split/PoseBusters fraction of leaked systems

SPLIT SET	PLI SHARED ≥ 50	POCKET LDDT ≥ 50	POCKET SHARED ≥ 50	PROTEIN GLOBAL LDDT ≥ 50	PROTEIN SEQSIM ≥ 30	LIGAND SIMILARITY ≥ 30	NO. TRAIN / VAL / TEST	TEST PASS QUALITY%
VS. TEST SET								
PDBBIND-ORIGINAL	0.91	1.00	1.00	1.00	1.00	0.62	22,365 / 7,549 / 423	50.12
PDBBIND-DIFFDOCK	0.43	0.76	0.73	0.76	0.80	**0.43**	25,442 / 1,570 / 236	22.46
DOCKGEN	0.04	0.08	**0.05**	0.08	**0.18**	0.64	40,916 / 285 / 590	50.00
PDBBIND-LP	0.77	0.87	0.86	0.89	0.94	0.40	18,152 / 3,906 / 7,265	40.37
PLINDER-TIME	0.80	0.96	0.88	0.95	0.98	0.54	76,950 / 11,392 / 11,412	19.28
PLINDER-ECOD	0.30	0.49	0.35	0.49	0.60	0.52	77,411 / 10,169 / 12,174	20.81
PLINDER-PL50	**0.04**	**0.00**	0.09	**0.01**	0.37	0.58	57,602 / 3,453 / 3,729	**100.00**
VS. POSEBUSTERS								
PDBBIND-DIFFDOCK	0.52	0.69	0.65	0.70	0.78	0.59	25,442 / 1,570 / 308	**100.00**
PLINDER-TIME	0.72	0.88	0.83	0.88	0.93	0.66	76,950 / 11,392 / 308	**100.00**
PLINDER-ECOD	0.64	0.74	0.70	0.75	0.81	0.65	77,411 / 10,169 / 308	**100.00**
PLINDER-PL50	0.40	0.47	0.47	0.48	0.64	0.64	57,602 / 3,453 / 308	**100.00**



Ongoing Developments

- Measured and predicted **binding affinities**
- **Cryptic pockets** and promiscuous ligands
- Data **augmentation** strategies
- **Leaderboard** across tasks and use-cases

Availability

ICML'24 ML4LMS

gs://plinder

github.com/plinder-org

plinder.sh

Universität Basel
VANTAI
NVIDIA
SIB Swiss Institute of Bioinformatics
BIOZENTRUM Universität Basel The Center for Molecular Life Sciences
MIT CSAIL
Massachusetts Institute of Technology
EuroHPC Joint Undertaking
European Commission

This project has received funding from the European High-Performance Computing Joint Undertaking Joint Undertaking (JU) under grant agreement No 956137. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Italy, Sweden, Austria, Czech Republic, Switzerland.