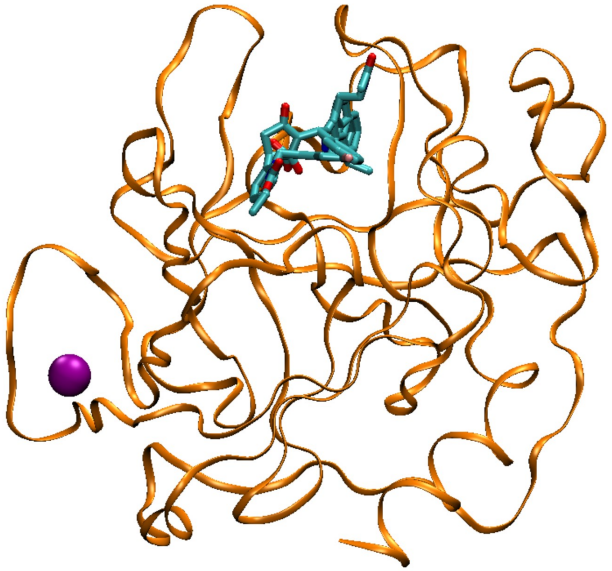
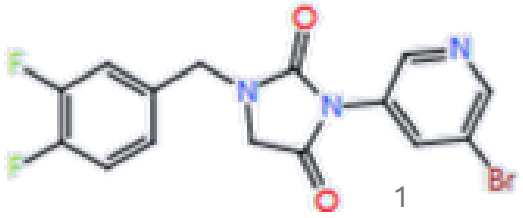


CASP16 Pharmaceutical Protein-Ligand Challenge

ligand poses and affinities

Mike Gilson
mgilson@ucsd.edu

UC San Diego
SKAGGS SCHOOL OF PHARMACY
AND PHARMACEUTICAL SCIENCES



Acknowledgements

Data Contributors

Hoffman La Roche
Christian Kramer
Andreas Tosstorff

Idorsia Pharmaceuticals Ltd
Florent Chevillard
Julien Hazemann
Aengus MacSweeney

Structural Genomics Consortium
Aled Edwards
Matthieu Schapira
Levon Halabelian

CASP16 Participants

Analysts and Organizers

Jerome Eberhardt, U Basel
Xavier Robin, U Basel

Sebastian Bittrich, RCSB PDB, UCSD
Jose Duarte , RCSB PDB, UCSD

Andriy Kryshtafovych, UC Davis
John Moult, U. Maryland

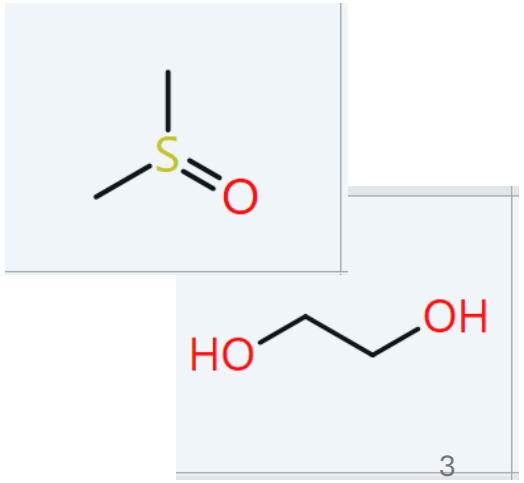
Four Challenge Types

Poses of pharma ligands: 3D structure of small molecules in protein binding sites
Also the structure of the binding site, but this ends up being secondary

Affinities of pharma ligands: binding affinity of protein with a drug-like molecule

Poses of incidental ligands in pharma structures (e.g. DMSO)

Self-assessment of pose-prediction accuracy



Outline

Mike Gilson – 40 min/slides

Challenge overview
Pose prediction results
Affinity prediction Results
Reliability
Speaker selection

Jerome Eberhardt – 10 min/slides

Performance of off-the-shelf “baseline” methods

CASP16 Pharma Challenge Targets

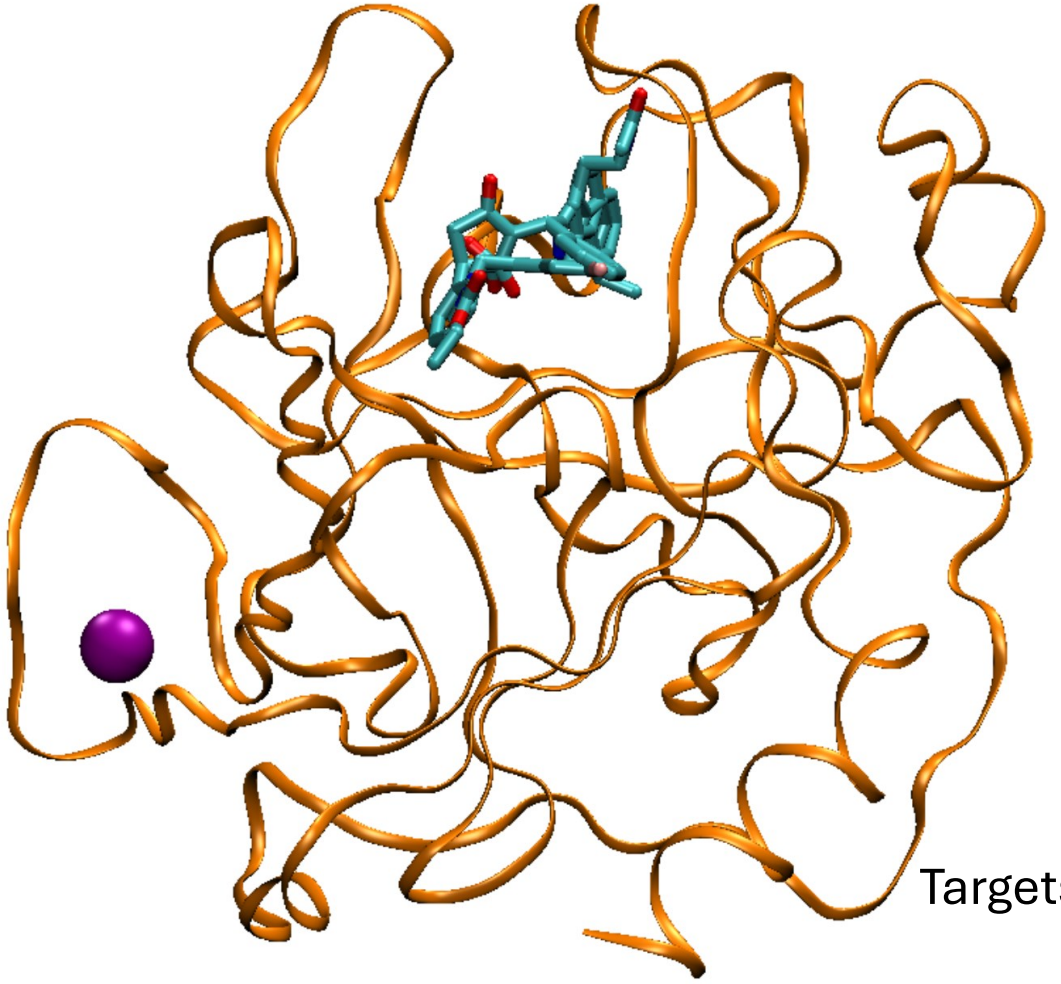
Supertarget ID	Protein	N_{Pose}	Resol (A)	$N_{Affinity}$	Affinities	Notes
L1000 (Roche)	**Chymase** (human)	17	1.1-2.2 mean 1.7	Stage 1: 17 Stage 2: 17	1-400 nM -12.4 to -8.8 kcal/mol	Racemic ligands
L2000 (Roche)	**Cathepsin G** (human)	2	1.2, 1.2	0		Racemic ligands
L3000 (Roche)	**Autotaxin** (rat)	189	1.3-2.7 mean 1.9	Stage 1: 123 Stage 2: 93	1nM-10uM* -12.4 to -6.9 kcal/mol	• Zn enzyme • No stereo centers • **5 incidental ligands included in challenge** • Some structures had two independent ligand poses. Others had alternative (A/B) confs of the ligand. We scored all predictions against all xtal poses and gave the one best score.
L4000 (Idorsia)	**Mpro** (SARS-CoV-2)	24**	1.3-1.9 mean 1.7	0	NA	• Stereochem provided • Four ligands covalently bound to Cys. Details provided to participants • **16 incidental ligands included in challenge**
L5000 (SGC)	**WDR55** (human)	1	1.95	0	NA	• Pilot from SGC's Target 2035 project, aiming for a small molecule modulator of every human protein by 2035. Scored separately due to lateness.

*: includes 11 reported K_D values of 10 uM, which are likely non-quantitative

**: was 25 but one target excluded because ligand L4006 was outside binding site at an xtal contact

Chymase

Supertarget 1000, 17 poses and affinities

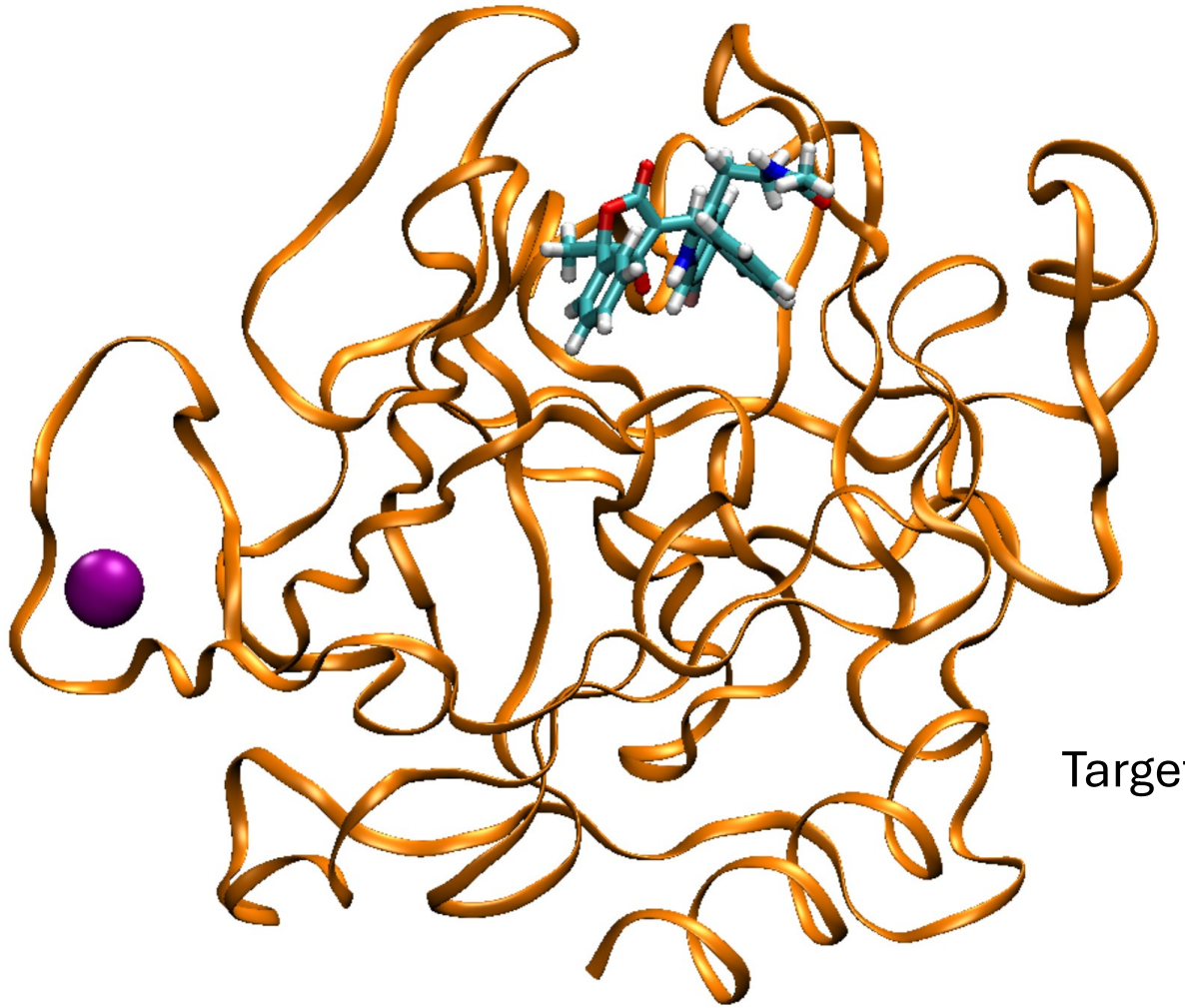


Targets L1001, L1002

structural Zn

Cathepsin G

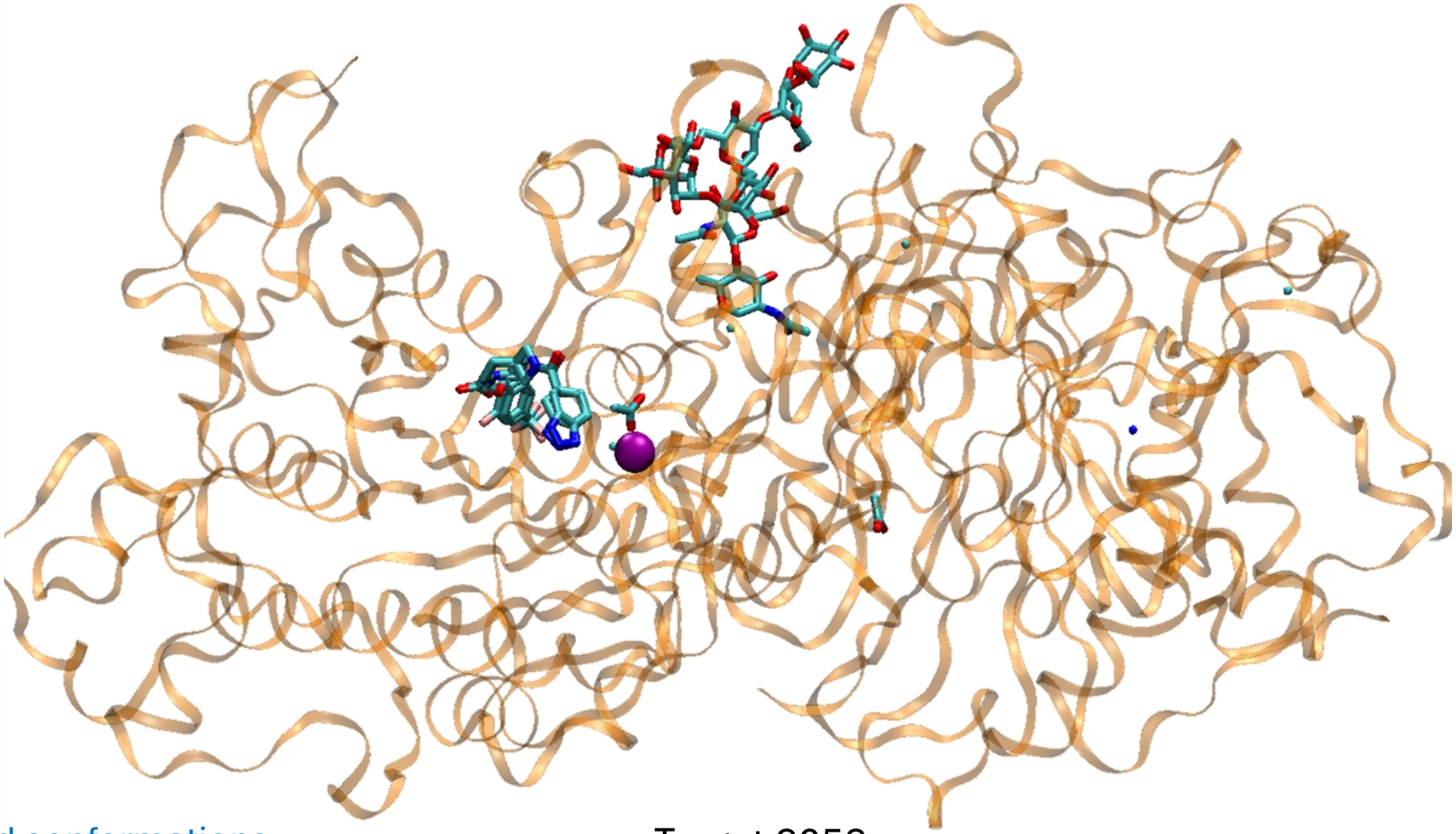
Supertarget 2000, 2 poses



Target 2001

Autotaxin

Supertarget 3000, 189 poses, 123 affinities (93 in common)

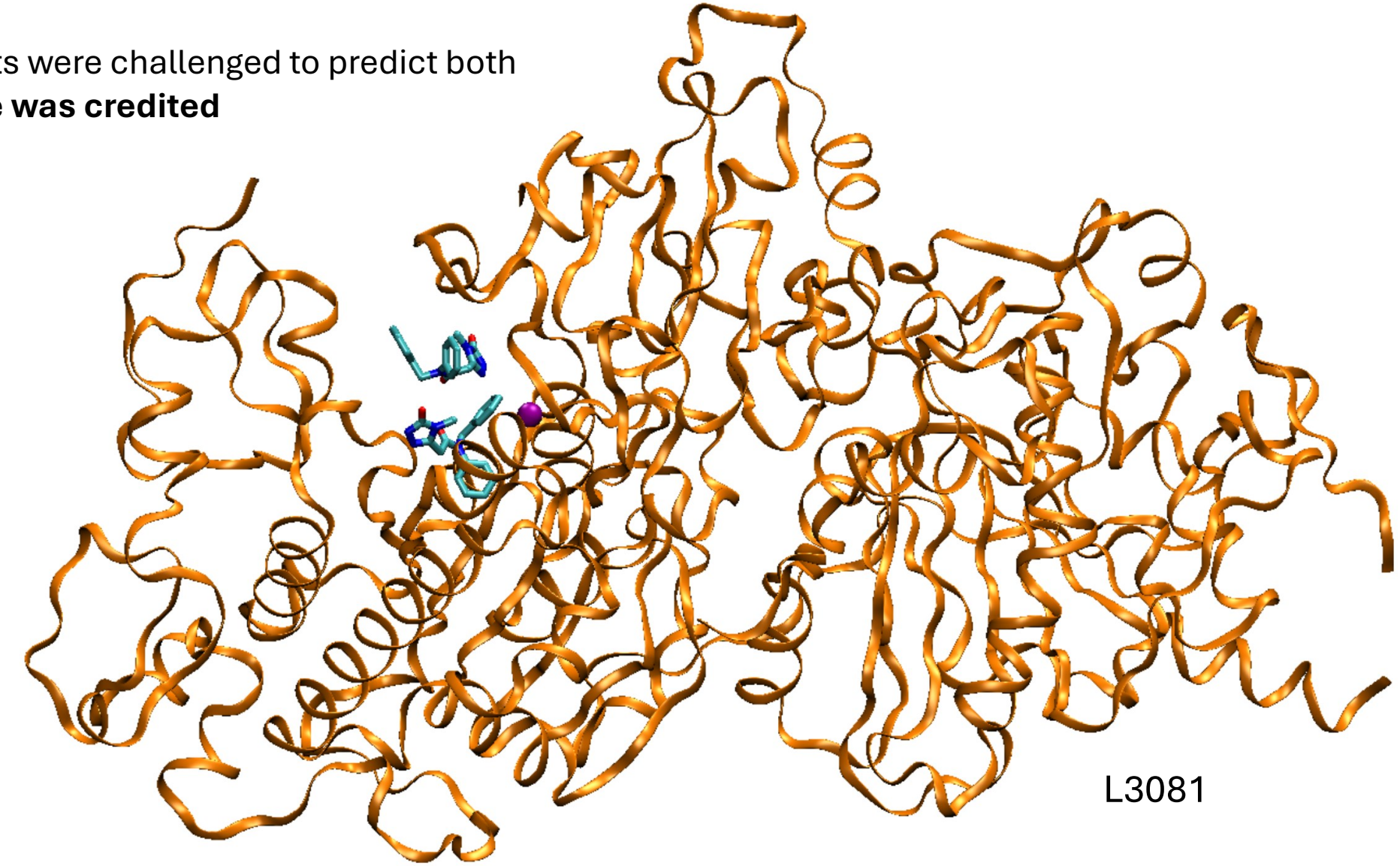


Target 3053

Alternate ligand conformations
Zn in/near binding pocket
Mannose chain

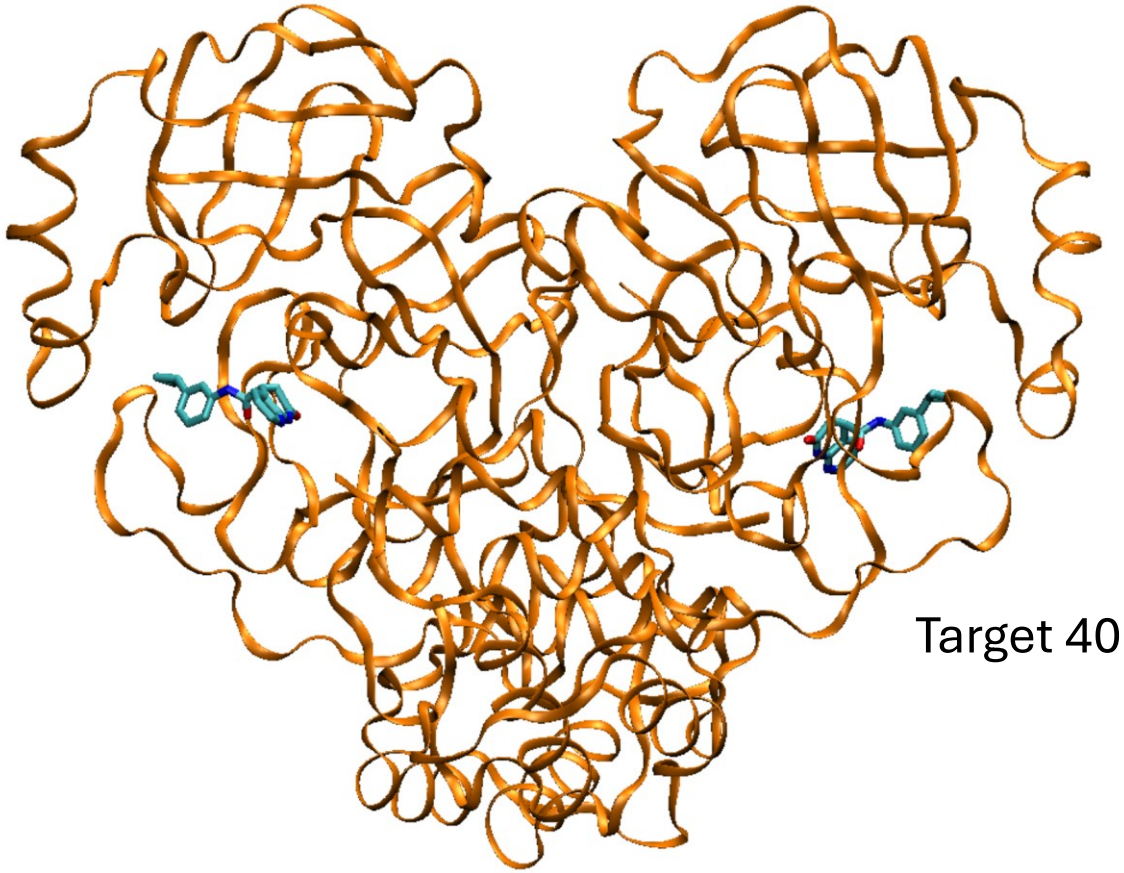
Six Autotaxin Targets Had Two Independent Ligand Poses

participants were challenged to predict both
best score was credited



SARS-Cov-2 Mpro

Supertarget 4000, 24 poses

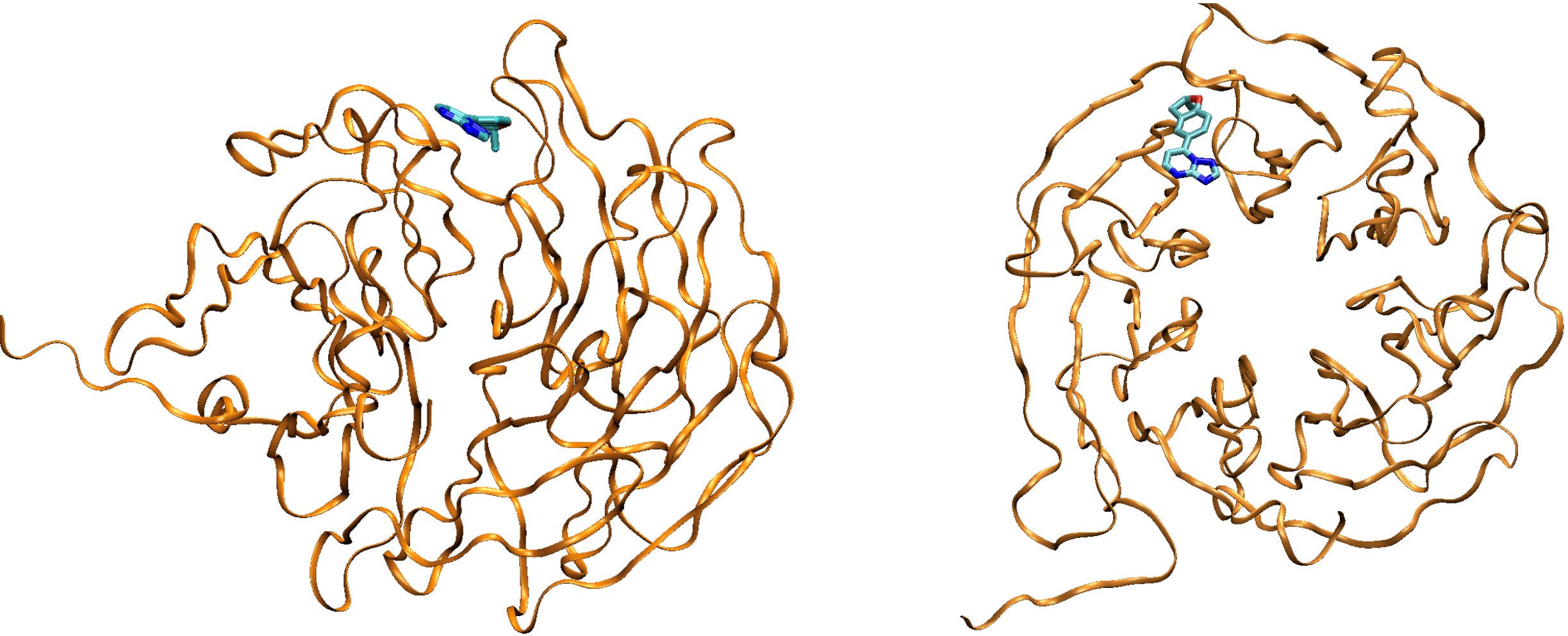


Target 4020

- Two chemically equivalent, but not crystallographically equivalent, chains, each with a bound ligand
- CASP invited docking to both sites but some participants submitted only one. We credited the best pose of two across both sites when two were provided, and the best score of both sites when only one submitted

WDR55

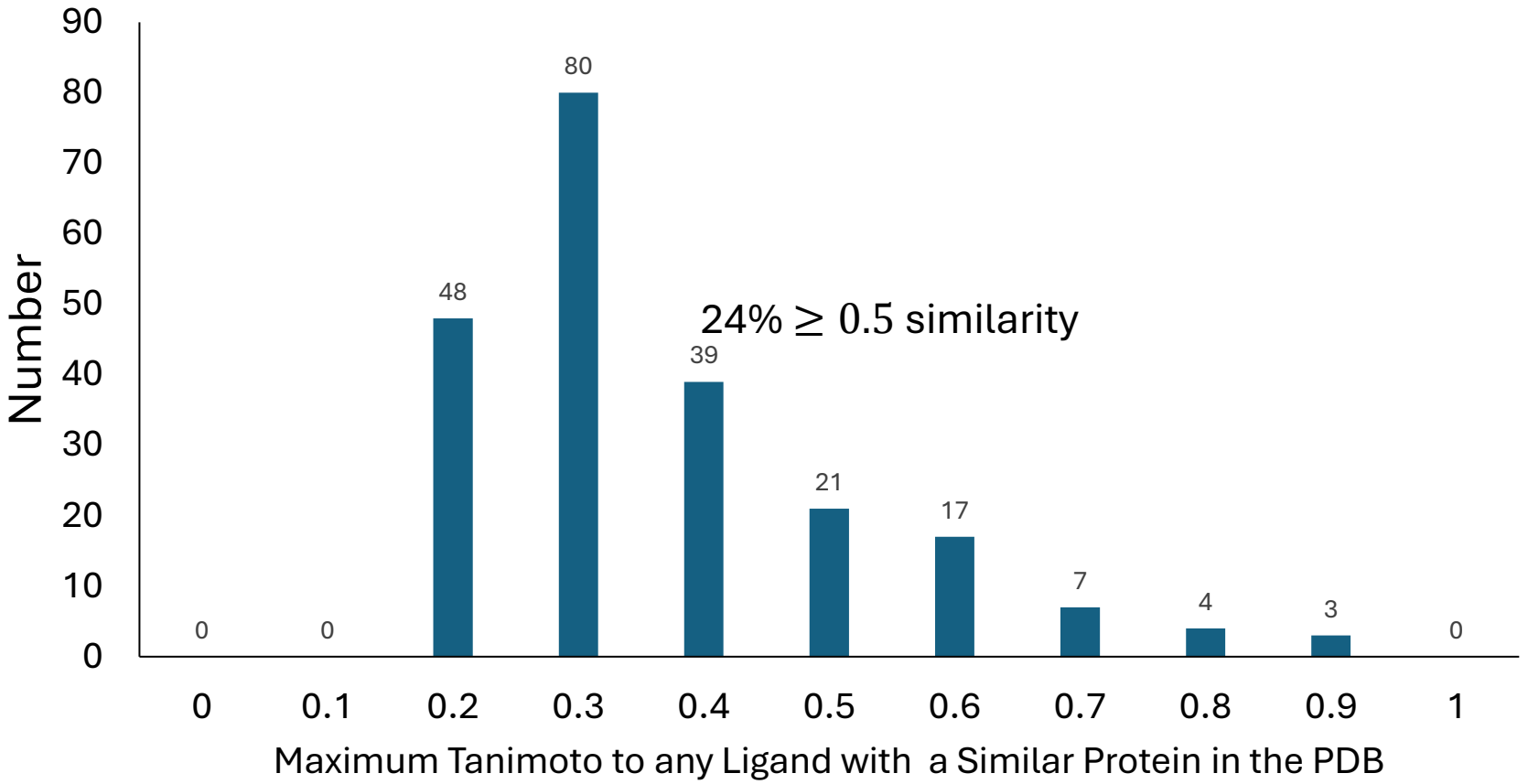
Supertarget 5000, 1 pose



Target 5001 (two viewpoints)

How Hard Are These Challenges?

For each CASP16 autotaxin ligand, the maximum Tanimoto similarity to any ligand in a protein in the PDB with ≥70% sequence similarity



Sebastian Bittrich, Jose Duarte, RCSB PDB

Main Information Provided to Participants

Data
Protein sequence
Ligand SMILES string
Whether to predict pose, affinity, or both
List of any nearby (<4.5 A) incidental ligands

Instructions: submit up to five model predictions, with Model 1 your favorite

Metrics

Poses

- **lDDT-PLI**: Fraction of correctly predicted ligand-protein interatomic distances for 6 Å radius, with penalty for incorrect close contacts. (0-1)
- If multiple binding sites or poses, we credited the most accurate
- Symmetry-corrected RMSD (Å) after binding-site superposition, and "success" if RMSD≤2.5 Å
  - RMSD can overweight bad poses
  - Dependent on superposition method
  - Success rate depends on success cutoff

Binding site structures

- bb_RMSD (Å):
- lDDT_LP (lDDT of ligand pocket)

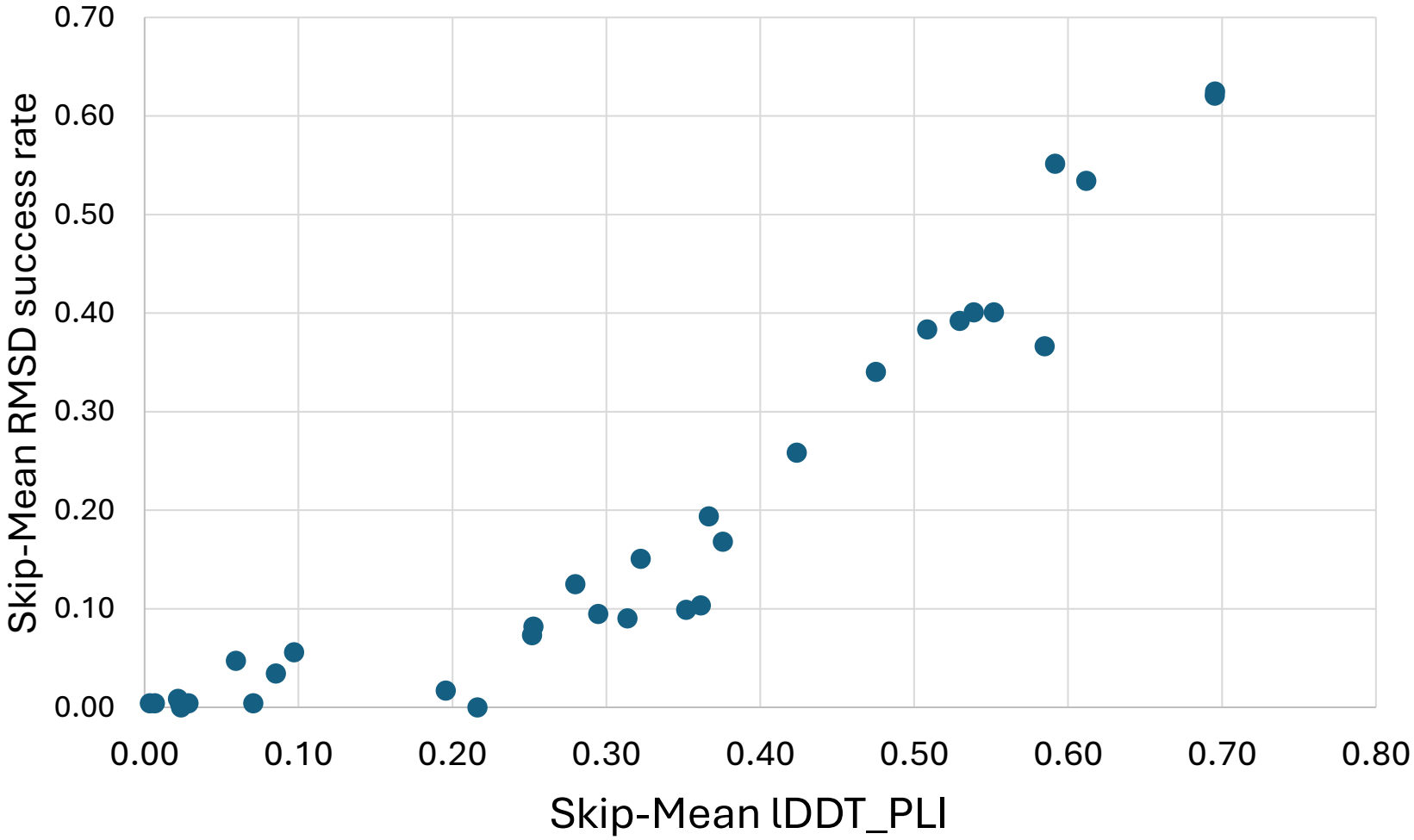
Affinities

Kendall's tau (applicable to all prediction types)

Structural metrics computed with OST (openstructure.org)

Comparison of RMSD <2.5 Success rate and lDDT_PLI using skip-penalized means, just for Model 1

each point a group

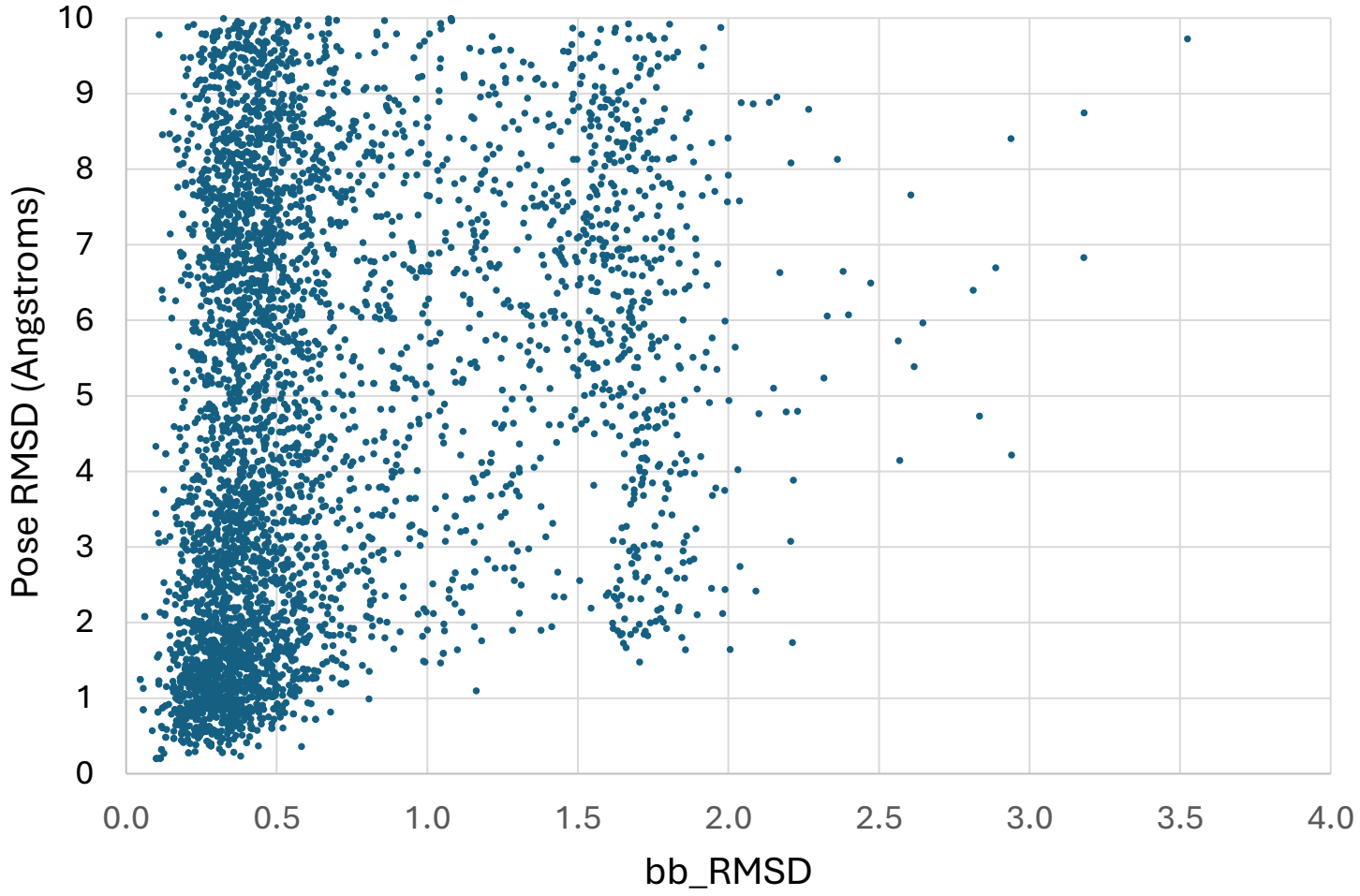


These two statistics tell the same story, at least for any predictions worth much attention (i.e., not the very lower left of the graph).

Pose Accuracy vs Binding Site Accuracy

Model 1 Pose Predictions

- Low RMSD implies low bb_RMSD
- Low bb_RMSD does not imply low RMSD



Methods Used by Participants

- Physics-based, AI-based, template-based, etc.
- Often hard to categorize; e.g., might use a ligand template if available and physics-based docking otherwise, then score poses with an AI method
- Mix of home-made software and other open-source and commercial software, often in one method; e.g. one that includes OMEGA2, Chimera, AutoDock Vina, Glide, and at least two others
- One method relied essentially entirely on citizen-scientist docking with DockIt

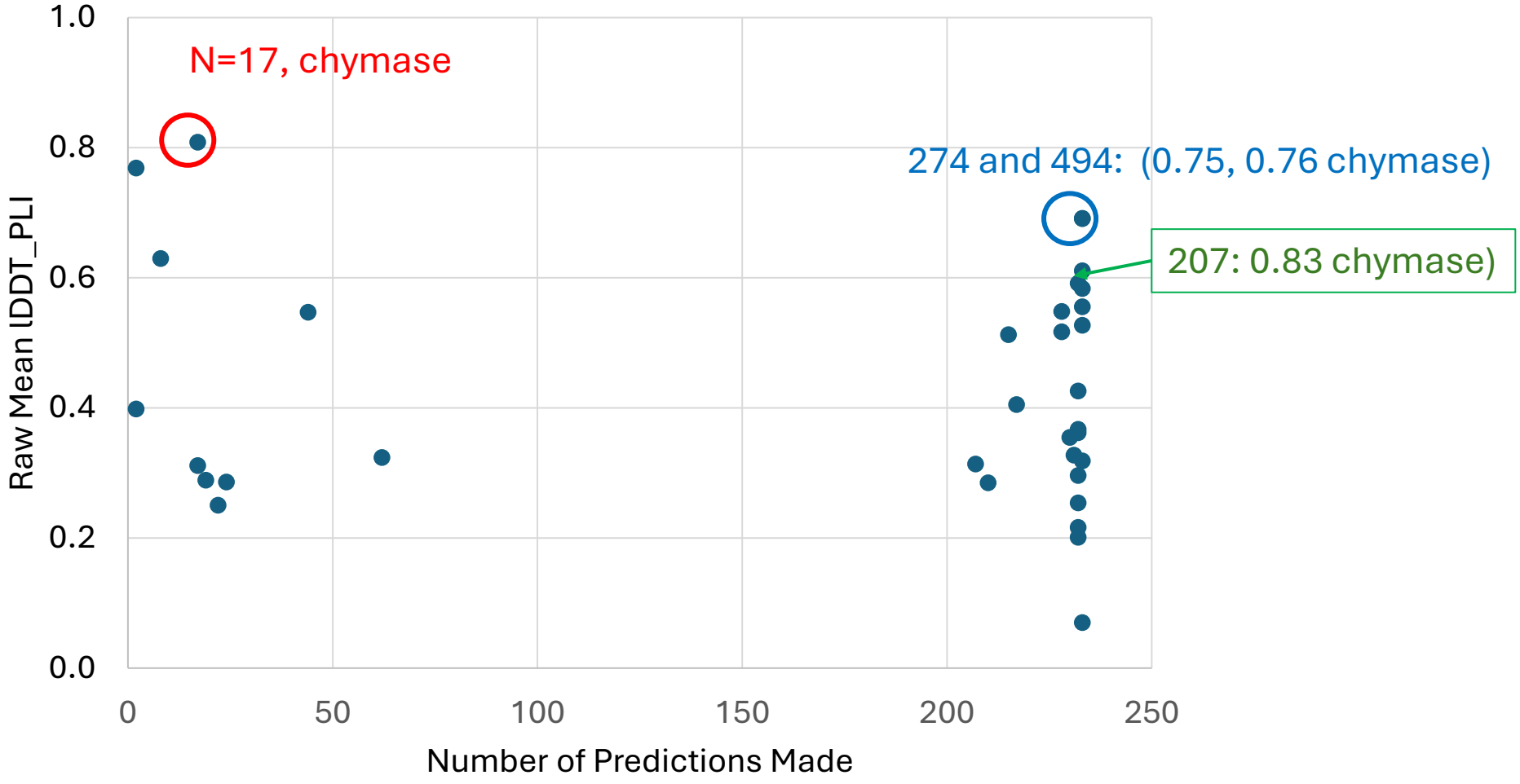
Assessment of Pose Predictions

initial focus on Model 1 predictions

34 CASP groups

32 distinct research groups

Accuracy of Predictions vs Number of Targets Predicted

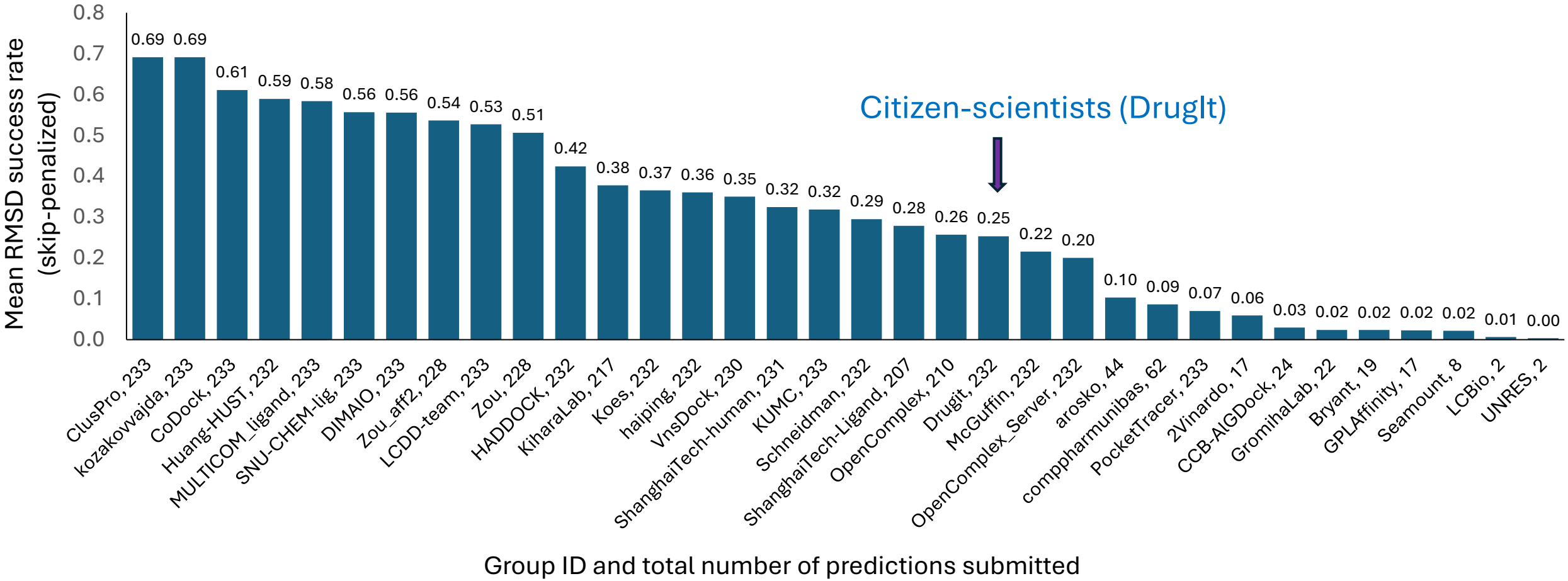


The top-scoring groups on the left should be ranked well below the top-scoring groups on the right

➔ skipped predictions may be treated as “unsuccessful” predictions (lDDT_PLI=0, RMSD>2.5)

Mean lDDTI_PLI for each CASP Group

Model 1, skip-penalized mean across all 233 pose targets

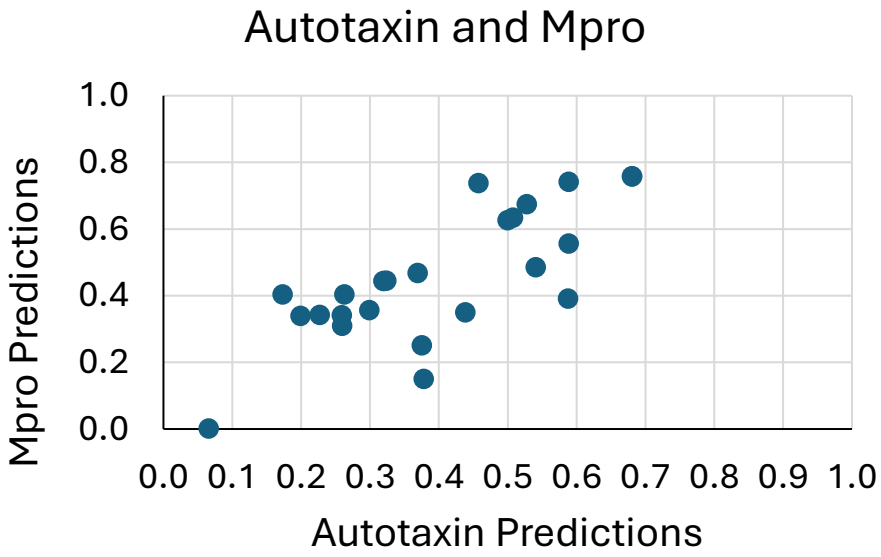


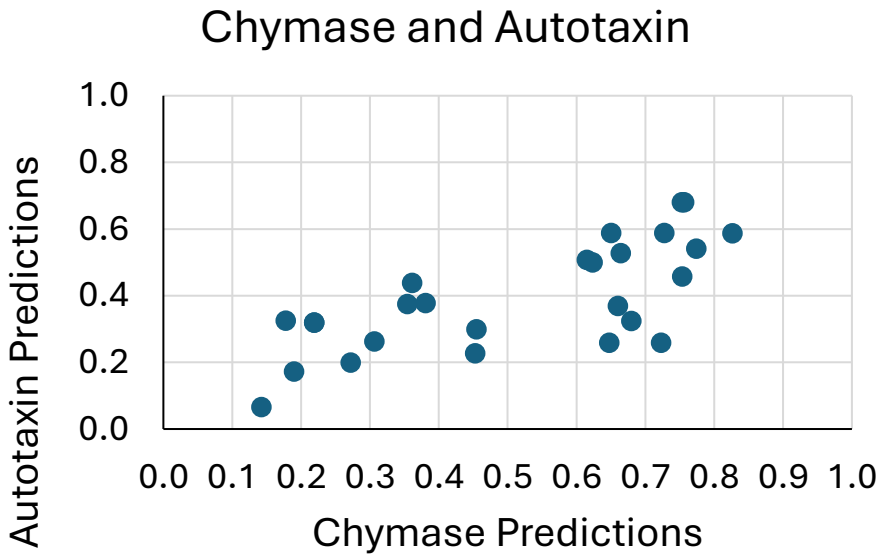
Group ID and total number of predictions submitted

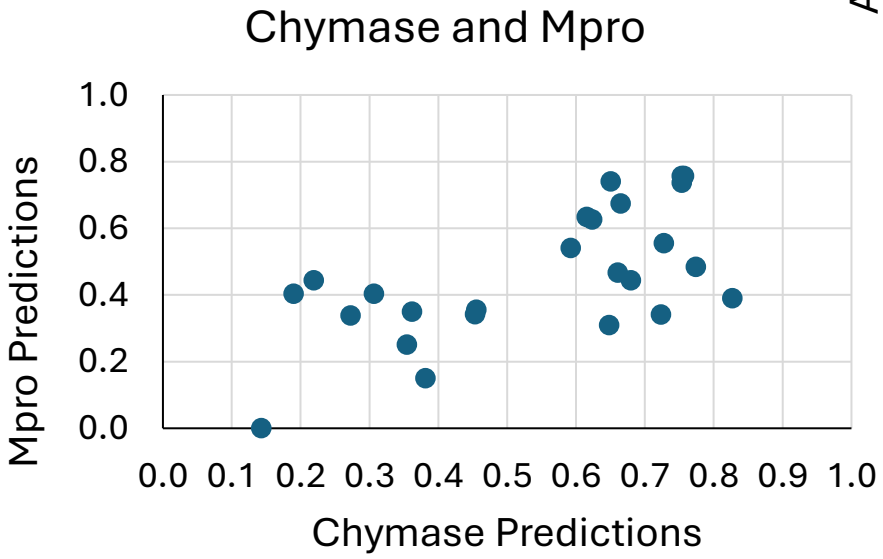
Top Groups: ClusPro, Kozakov/vajda, CoDock, Huang-HUST, MULTICOM_ligand

Consistency of Model 1 Performance Across Supertargets

Focus on chymase (N=17), autotaxin (N=189), Mpro (N=24)

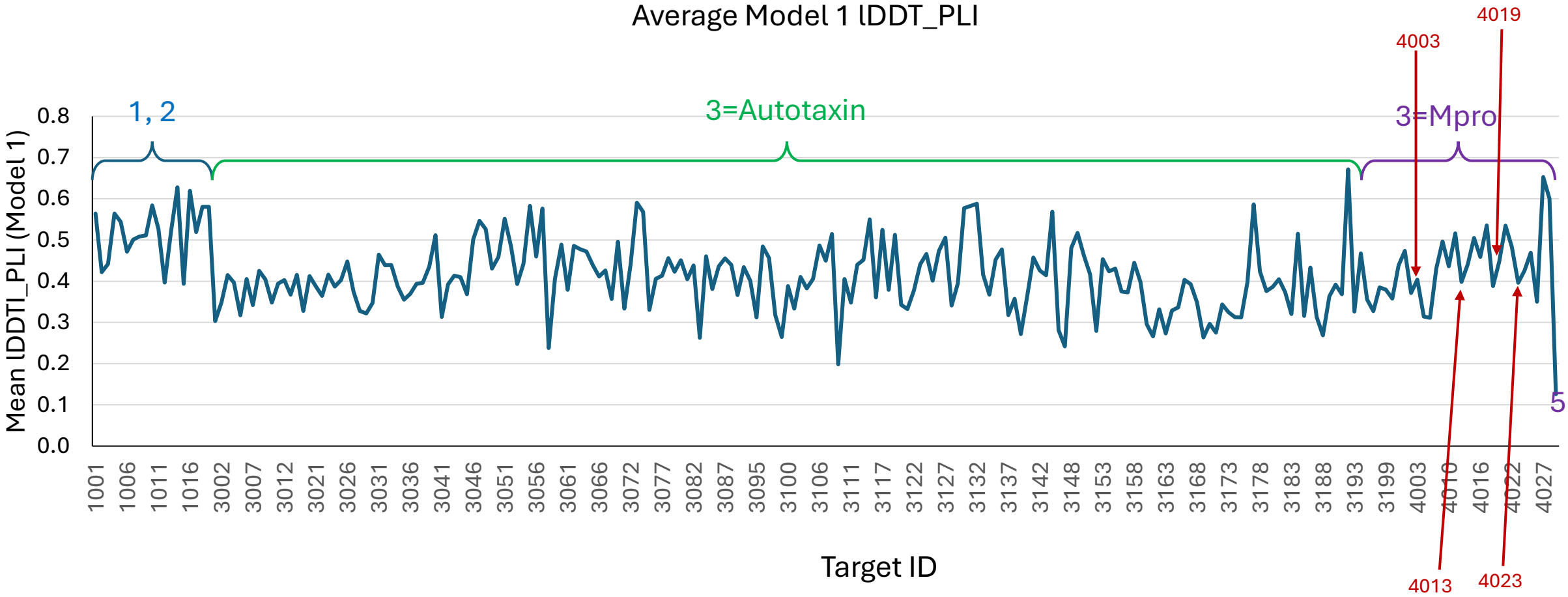






Raw mean lDDTI_PLI

Are Some Pose-Prediction Targets More Difficult?



1: chymase
2: cathepsin G
3: autotaxin
4: Mpro
5: WDR55

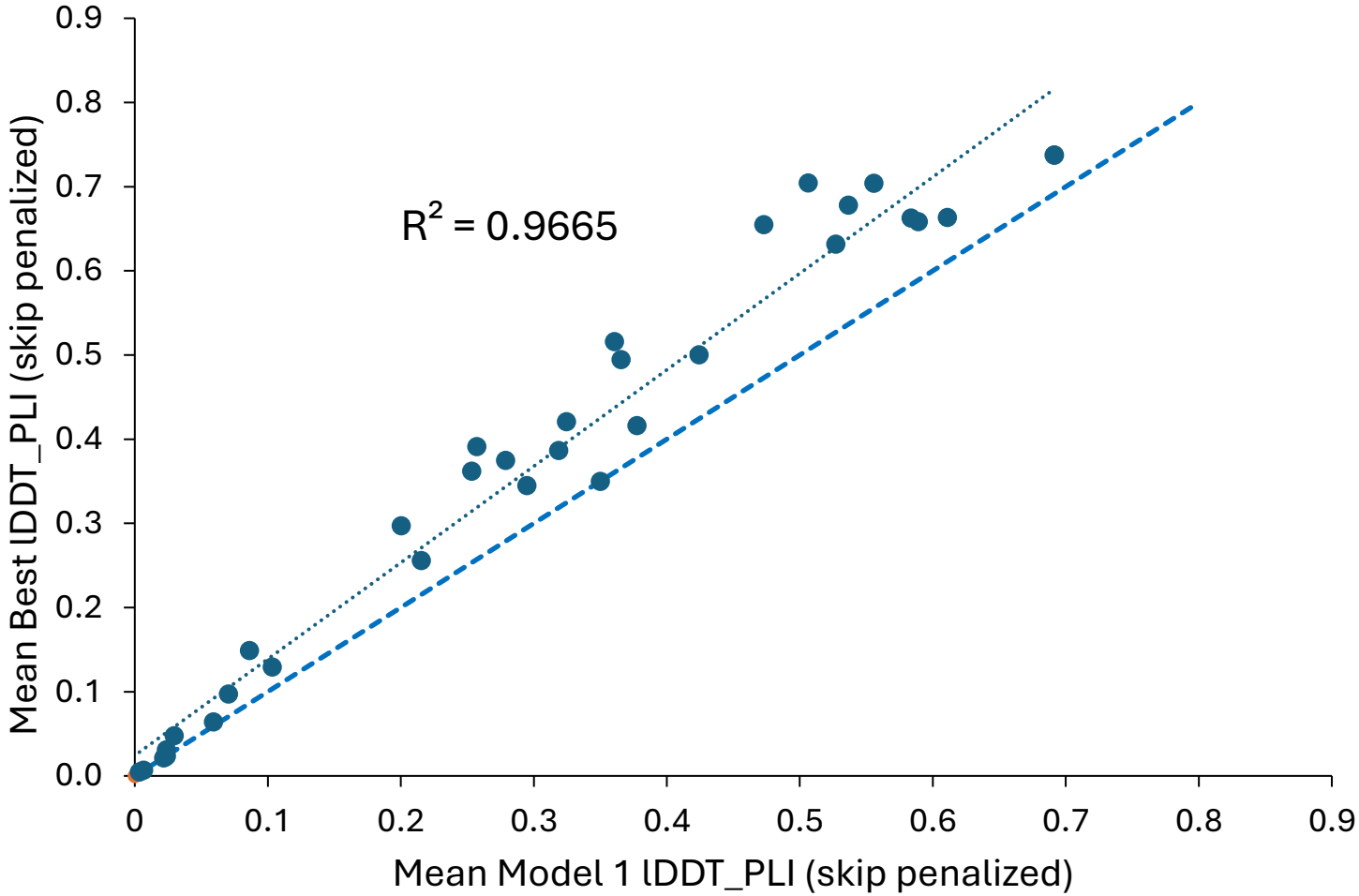
Covalent ligands

No skip penalty

Mean Best-of-5-Models Predictions vs Model 1 Predictions

skip mean lDDT_PLI

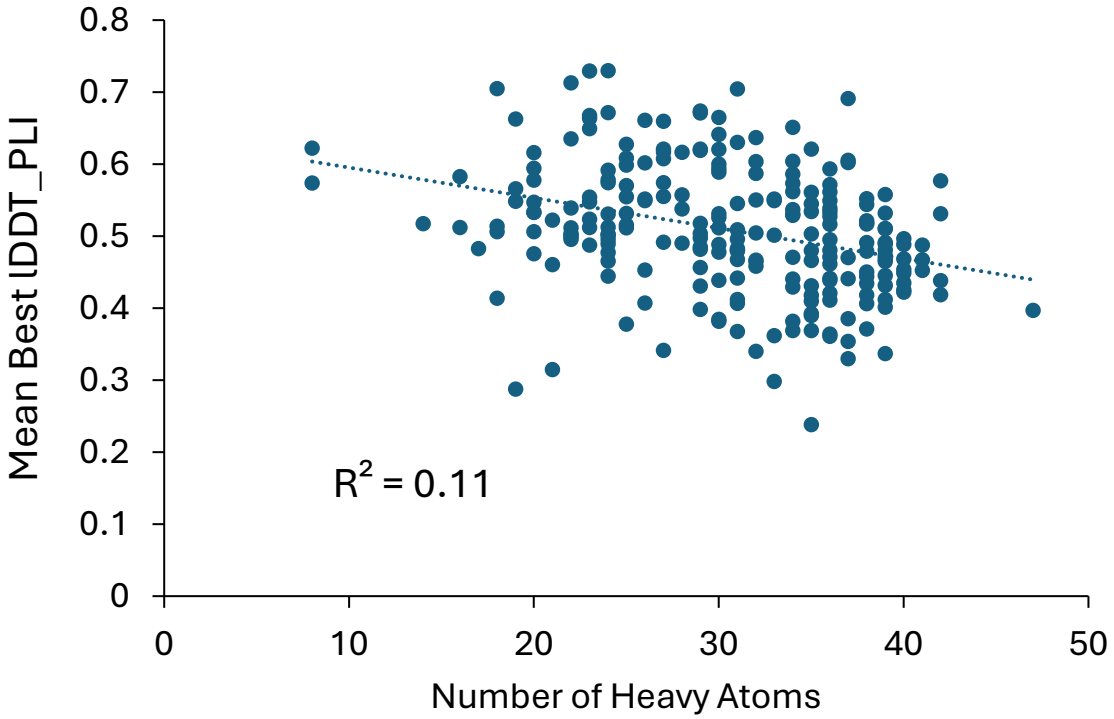
mean improvement from Model 1 alone is 0.08

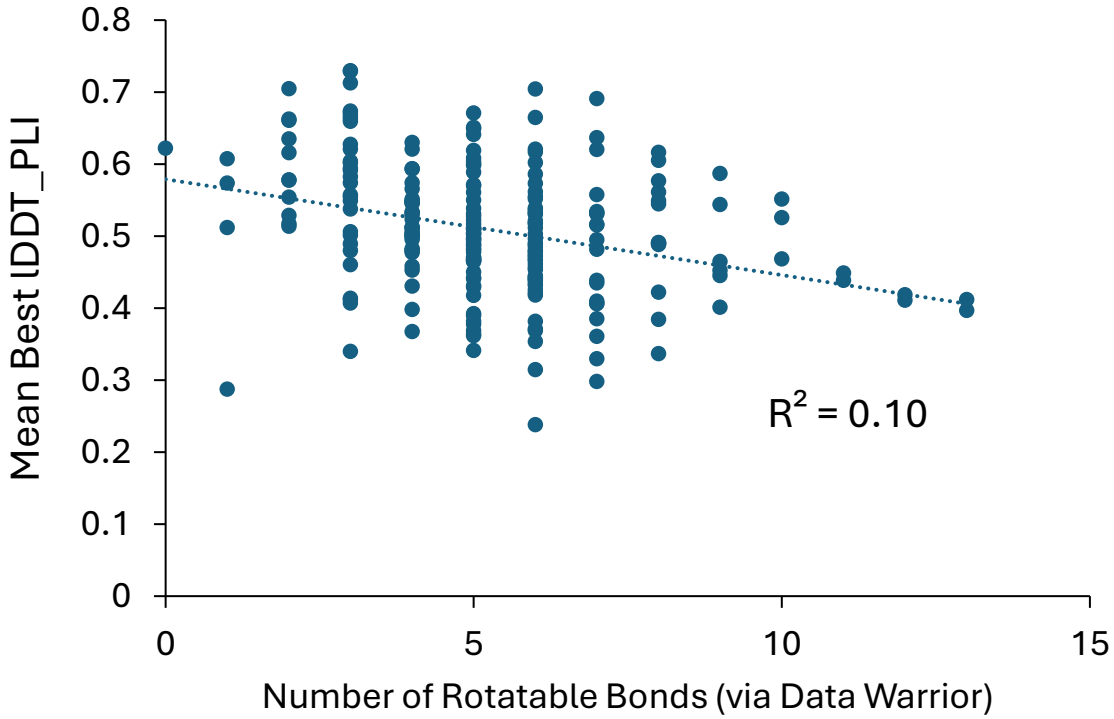


N.B.: not all groups submitted >1 model

Pose Accuracy vs Ligand Size and Flexibility

mean, raw, best-of-five lDDT_PLI

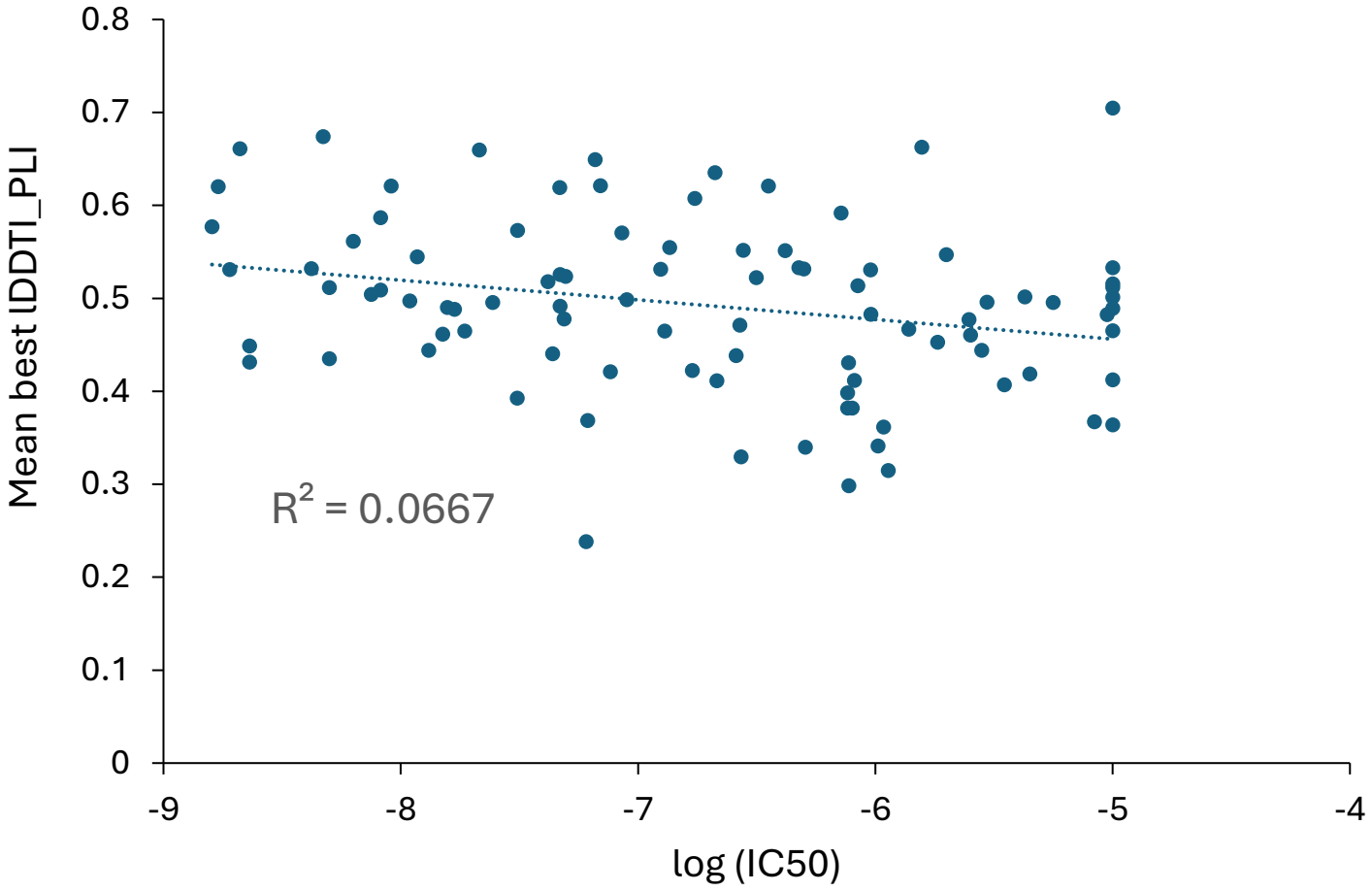




Raw means

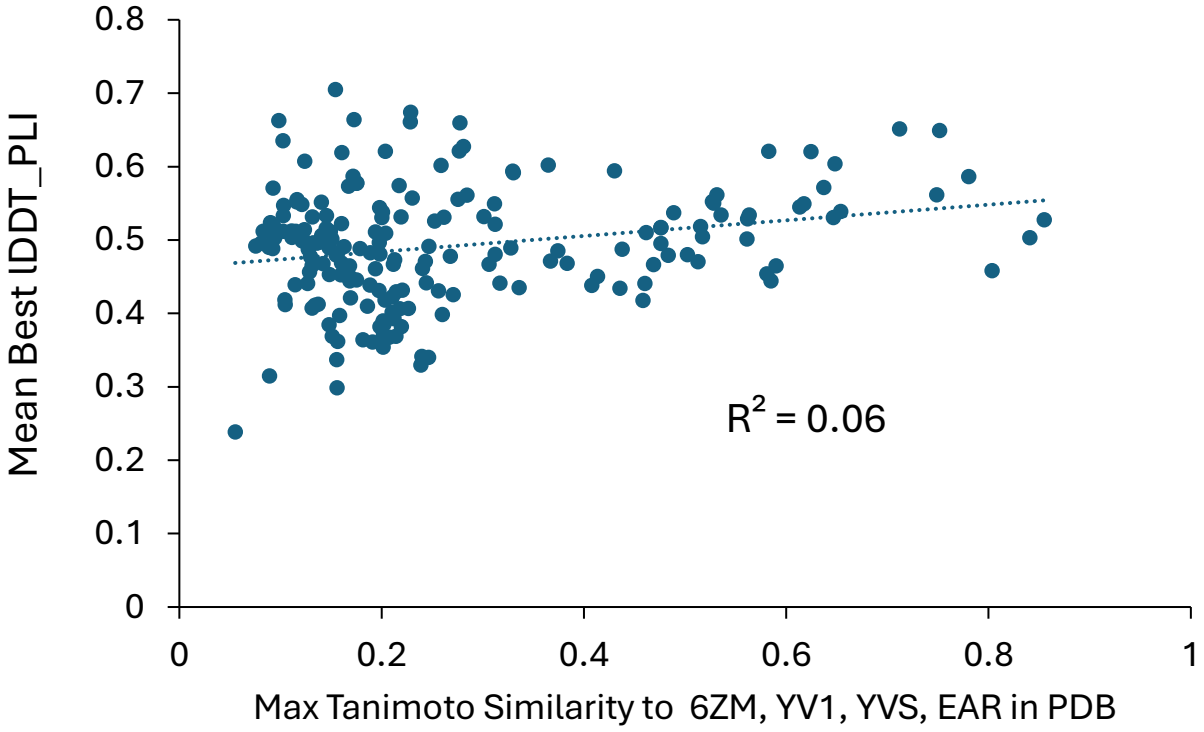
Pose Accuracy vs Ligand Affinity

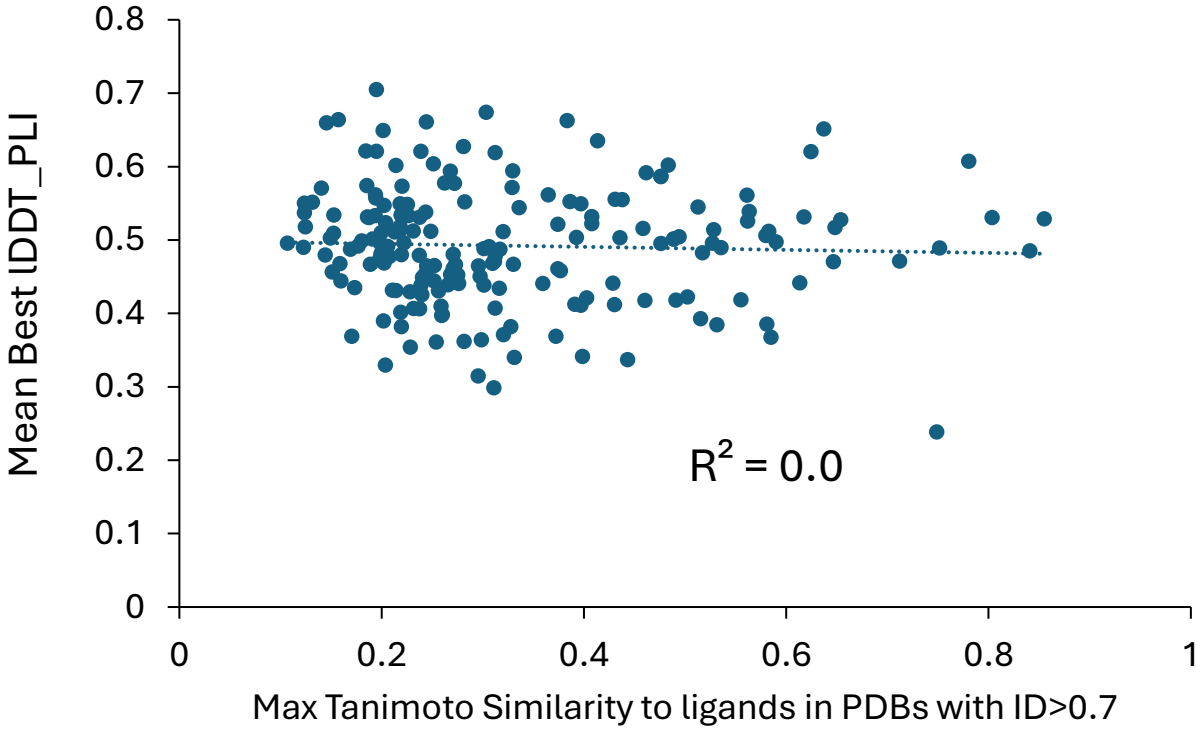
mean, raw, best-of-five lDDT_PLI



Pose Accuracy vs Available Template Similarity

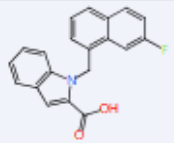
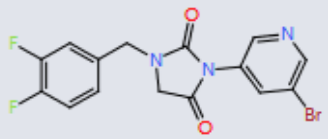
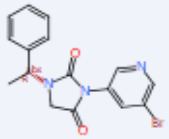
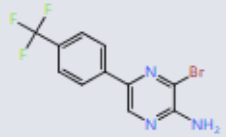
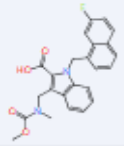
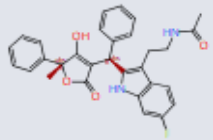
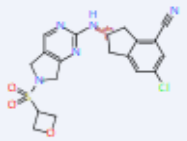
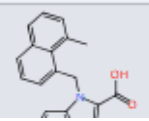
Autotaxin, mean, raw, best-of-five lDDT_PLI

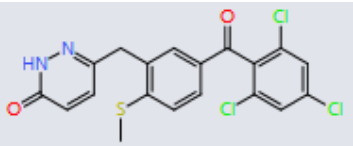
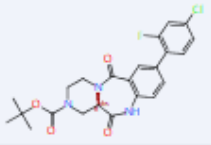
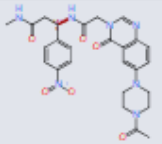
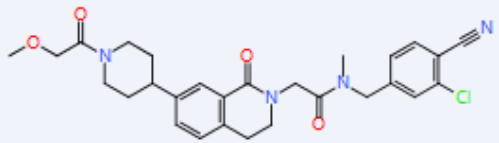
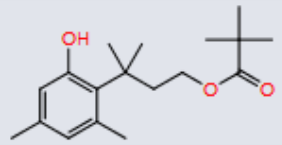
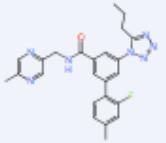
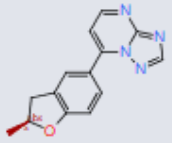
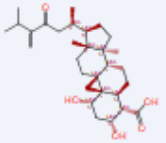




PDB ligands of rat and mouse autotaxin structures
5LOB, 5S9M, 5S9N, 6LEH

Easiest

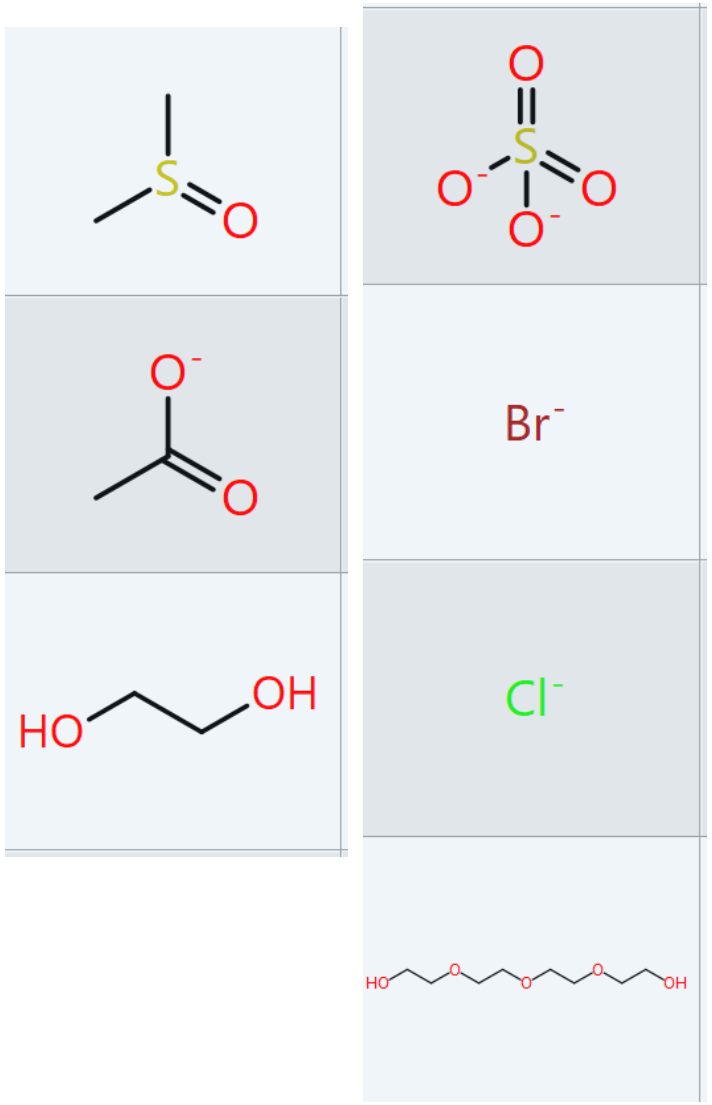
Mean best ld...	unique targ...	Structure of SMILES
0.72986226	1010	
0.729179172	4028	
0.712940909	4027	
0.704875708	3192	
0.70432511	2002	
0.691145691	2001	
0.673908044	3132	
0.671551303	1014	

0.341196914	3095	
0.339759556	3099	
0.336797476	3146	
0.329586634	3188	
0.314680698	3083	
0.298332049	3058	
0.287381562	5001	
0.238360422	3109	

Hardest

NonPharma (“Incidental”) Ligands in Autotaxin and Mpro

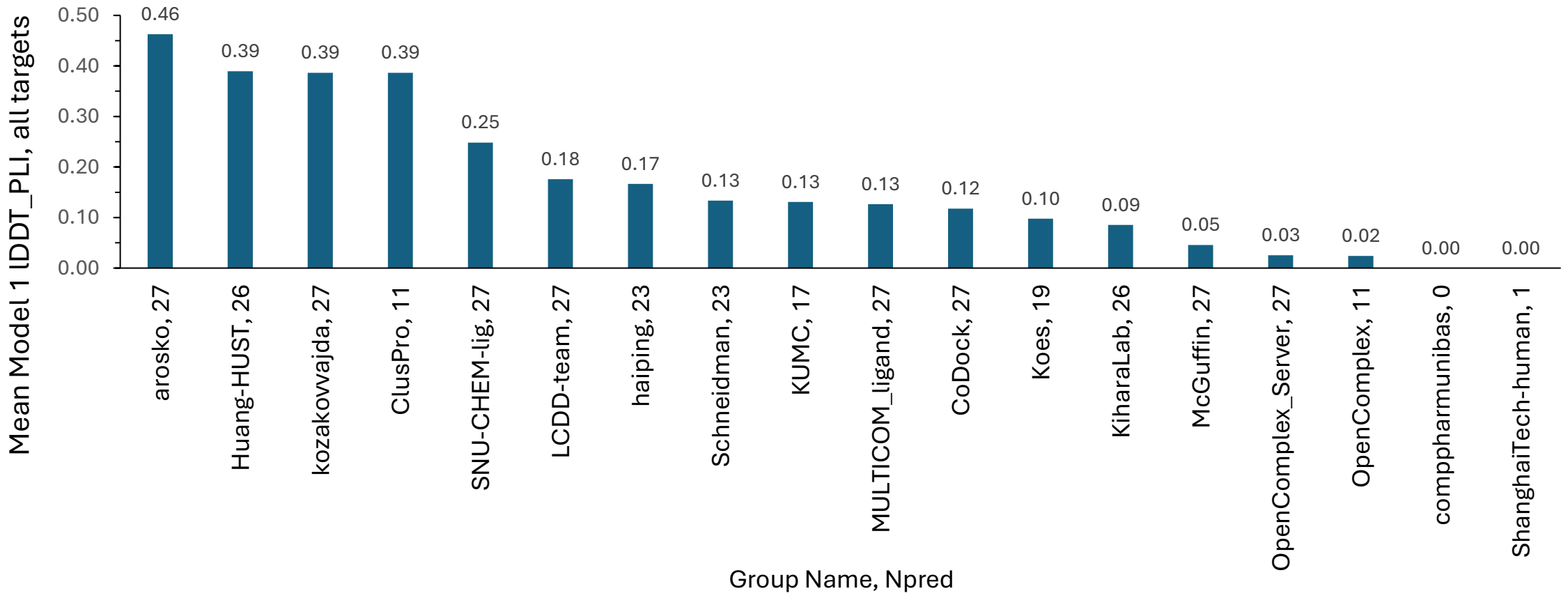
pose prediction



Incidental Ligands in Pharma Targets

Accuracy of Model 1 Pose Predictions, not skip-penalized

(maximum was 0.69 for pharma ligands)



Top Groups: arosko, Huang-HUST, kozakovvajda, ClusPro

Assessment of Affinity Predictions

28 unique groups participated

Included Stage 1 (no exptl poses) and Stage 2 (exptl poses provided)

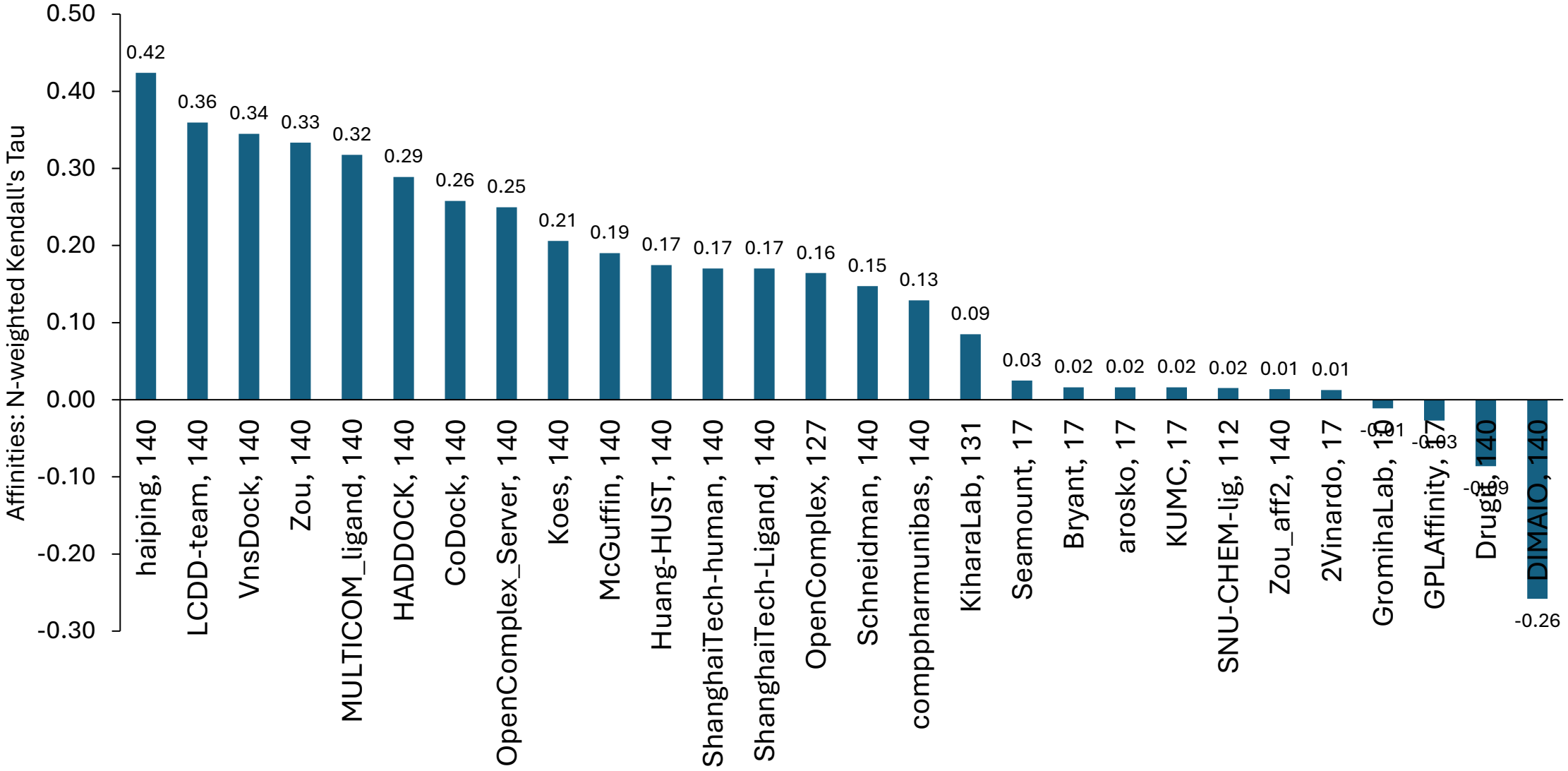
Focus here on Model 1 only, for brevity

Kendall's tau allows comparison across prediction types

- absolute affinities
- relative affinities
- affinity rankings

Ranking of Stage 1 Affinity Predictions for Both Targets by Group

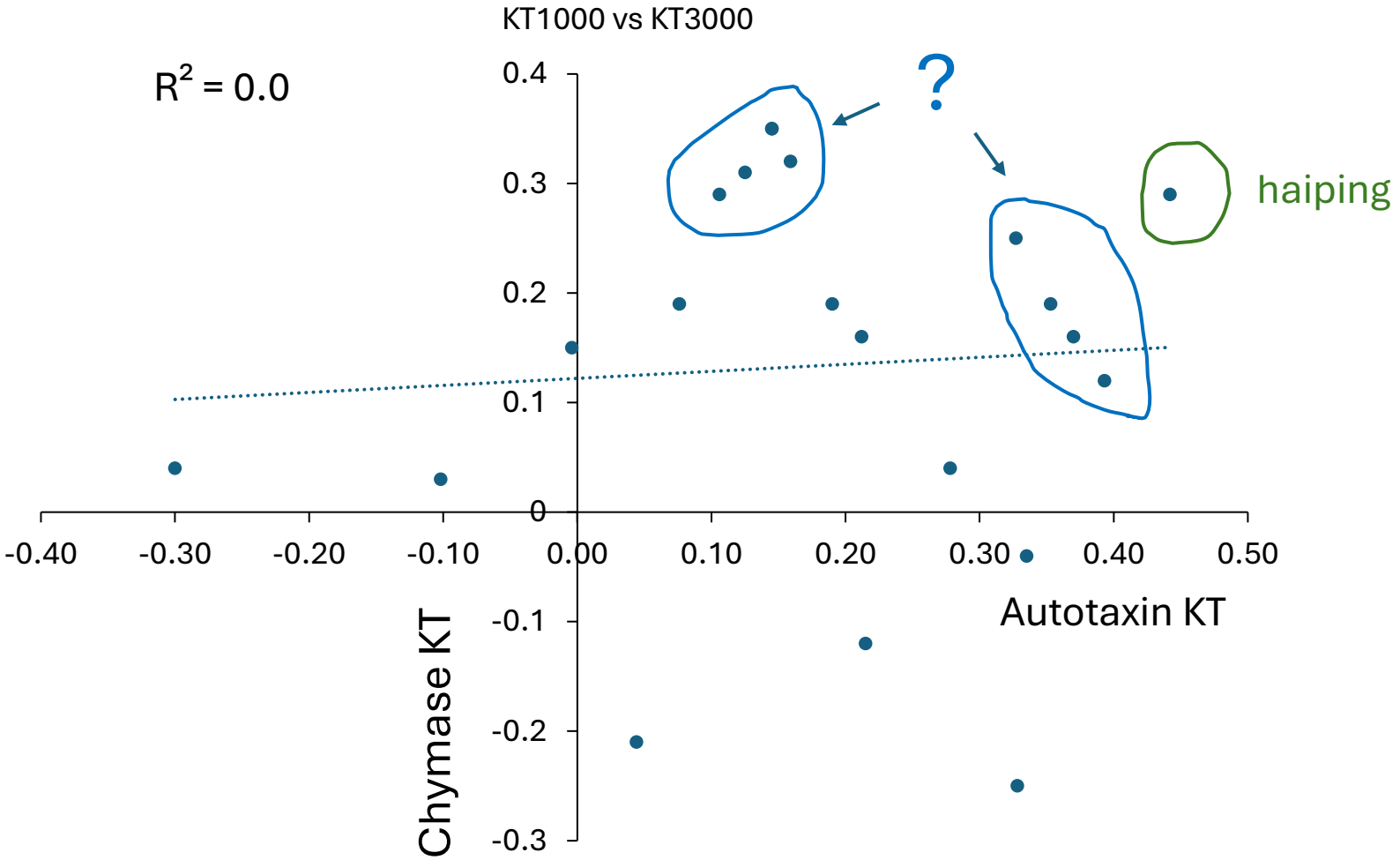
N- weighted Kendall's tau: $(N_aKT_a + N_cKT_c)/140$



Top Groups: haiping, LCDD-team, VnsDock, Zou, MULTICOM_ligand

No Correlation of Model 1 KT Values for Chymase and Autotaxin

stage 1; each point a group



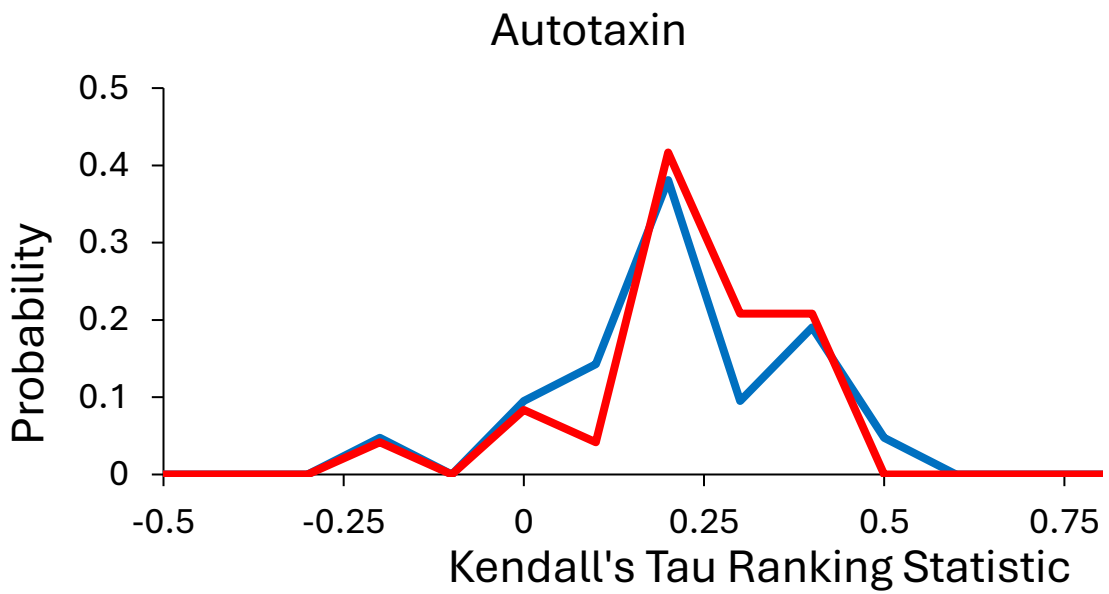
Stage 2 Affinity Predictions were made after release of experimental poses

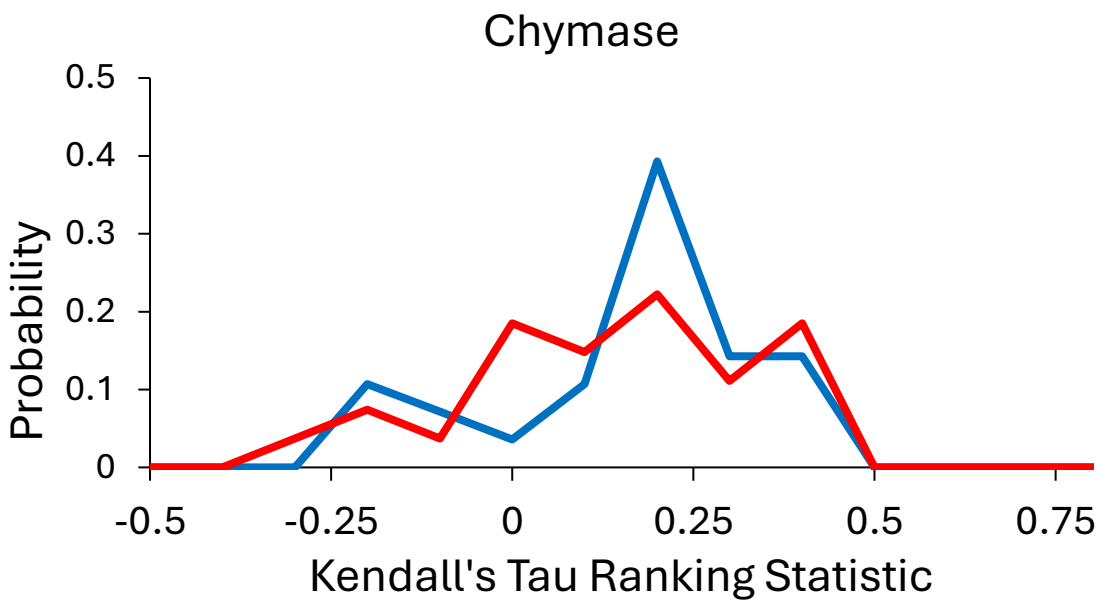
- 93 autotaxin affinity predictions and 17 chymase predictions made with knowledge of the 93 & 17 respective cocrystal structures

Not fully representative of a real-world application
Presumably easier--?

Stage 1 vs Stage 2 Accuracy

no improvement





21 stage 1 and 24 stage 2 submissions
93 affinity targets

28 stage1 groups, 27 stage 2 groups
17 affinity targets

Reliability scores (LSCORE)

Each group predicts their own accuracy across their submitted models

Group	N	Kendall's τ
HADDOCK	1292	0.43
McGuffin	1295	0.29
Bryant	85	0.24
MULTICOM_ligand	1295	0.19
haiping	1295	0.11
Zou	1225	0.11
Zou_aff2	1225	0.11
Koes	959	0.06
CoDock	1295	0.04
GromihaLab	22	-0.16

Looked for Groups…

Best at pose-prediction
Best at affinity prediction
Best overall, including LScore

Summary Scores

- **Pose:** 0.75 pharma-lDDT_PLI + 0.25 incidental-lDDT_PLI
- **Affinity:** 0.75 Stage 1 + 0.25 Stage 2
  Each stage: 1/3 Chymase + 2/3 autotaxin
- **Overall**: Overall pose + Overall affinity + 0.25 LSCORE

Selection of Groups By Three Criteria

	Group	Pharma lDDT_PLI	Incidental lDDT_PLI	S1 Chymase	S1 Autotaxin	S1 Summary (1/3 chymase, 2/3 autotaxin)	S2 Chymase	S2 Autotaxin	S2 Summary (1/3 chymase, 2/3 autotaxin)	Affinity Summary : 0.75 S1 +0.25 S2	L_{score}	Summary statistic: 3/4 * p-lDDT_PLI + 1/4i-lDDT_PLI + 3/4 S1 Summary + 1/4 S2 Summary + 0.25* Lscore	Summary normalized to max of 1
POSE Best p-lDDT_PLI	494	0.70	0.39			0.00			0.00	0.00		0.62	0.28
	274	0.70	0.39			0.00			0.00	0.00		0.62	0.28
	262	0.61	0.12	-0.25	0.33	0.14	-0.21	0.15	0.03	0.11	0.11	0.63	0.28
	91	0.59	0.39	-0.12	0.22	0.11	-0.03	0.20	0.12	0.11		0.65	0.29
	207	0.58	0.13	0.25	0.33	0.30	0.19	0.28	0.25	0.29	0.30	0.83	0.37
	432	0.55		0.04	-0.30	-0.19	0.03	-0.03	-0.01	-0.14		0.27	0.12
	420	0.54		-0.21	0.04	-0.04	-0.24	0.14	0.01	-0.03	0.16	0.42	0.19
	55	0.53	0.18	0.12	0.39	0.30			0.00	0.22		0.67	0.30
AFFINITY Best Overall Affinity	16	0.36	0.17	0.29	0.44	0.39	0.21	0.28	0.26	0.36	0.15	0.71	0.31
	207	0.58	0.13	0.25	0.33	0.30	0.19	0.28	0.25	0.29	0.30	0.83	0.37
	204	0.51		0.19	0.35	0.30	0.24	0.18	0.20	0.27	0.13	0.69	0.31
	82	0.35		0.16	0.37	0.30			0.00	0.23		0.49	0.22
	55	0.53	0.18	0.12	0.39	0.30			0.00	0.22		0.67	0.30
	386	0.28		0.35	0.15	0.22	0.35	0.16	0.22	0.22		0.43	0.19
	298	0.32	0.00	0.35	0.15	0.22	0.35	0.16	0.22	0.22		0.46	0.20
	167	0.25	0.02	0.32	0.16	0.21	-0.07	0.32	0.19	0.21		0.40	0.18
OVERALL Best Summary Score	207	0.58	0.13	0.25	0.33	0.30	0.19	0.28	0.25	0.29	0.30	0.83	0.37
	16	0.36	0.17	0.29	0.44	0.39	0.21	0.28	0.26	0.36	0.15	0.71	0.31
	204	0.51		0.19	0.35	0.30	0.24	0.18	0.20	0.27	0.13	0.69	0.31
	55	0.53	0.18	0.12	0.39	0.30			0.00	0.22		0.67	0.30
	91	0.59	0.39	-0.12	0.22	0.11	-0.03	0.20	0.12	0.11		0.65	0.29
	262	0.61	0.12	-0.25	0.33	0.14	-0.21	0.15	0.03	0.11	0.11	0.63	0.28
	494	0.70	0.39			0.00			0.00	0.00		0.62	0.28
	274	0.70	0.39			0.00			0.00	0.00		0.62	0.28

Speakers

Sandor Vajda, ClusPro (494)& kozakovvajda (274): Template based, both whole-ligand and ligand fragments.
(Ryota Ashizawa, Omeir Khan, Sergei Kotelnikov, Maria Lazou, Xiaognang Li, UsmanGhani, Dzmitry Padhorny, Dmitri Beglov, Sandor Vajda, Dima Kozakov)

Shan Chang, CoDock, (262): template-guided docking, AI-based scoring.
(author list unavailable)

Shengyou Huang, Huang-HUST (91): template based, traditional docking if no template.
(Keqiong Zhang, Qilong Wu, and Sheng-You Huang)

Haiping Zhang, haiping (16): Conventional docking followed by a graph-based neural network method to identify the best docked poses.
(author list unavailable)

Alex Morehead, MULTICOM_ligand (207) : PoseBench script invoking multiple DL methods (e.g. DiffDock and Neuralplexer) and then some method of choosing models.
(Alex Morehead, Jian Liu, Pawan Neupane, Jianlin Cheng)

Lessons Learned

It is difficult to make predictions, especially about the future.
-- *Berra? Bohr?*

It is difficult to make predictions about the future, but it is impossible to make predictions about anything else.
-- *Gilson*

Performance of Baseline Methods

Jerome Eberhardt, U Basel

Without thinking too much, how far would we go?

- Fully automated method v.s. manual tweaking / expert knowledge
- Gives an idea on how hard or easy was this CASP ligand category
- To know if it is time to change the baseline…

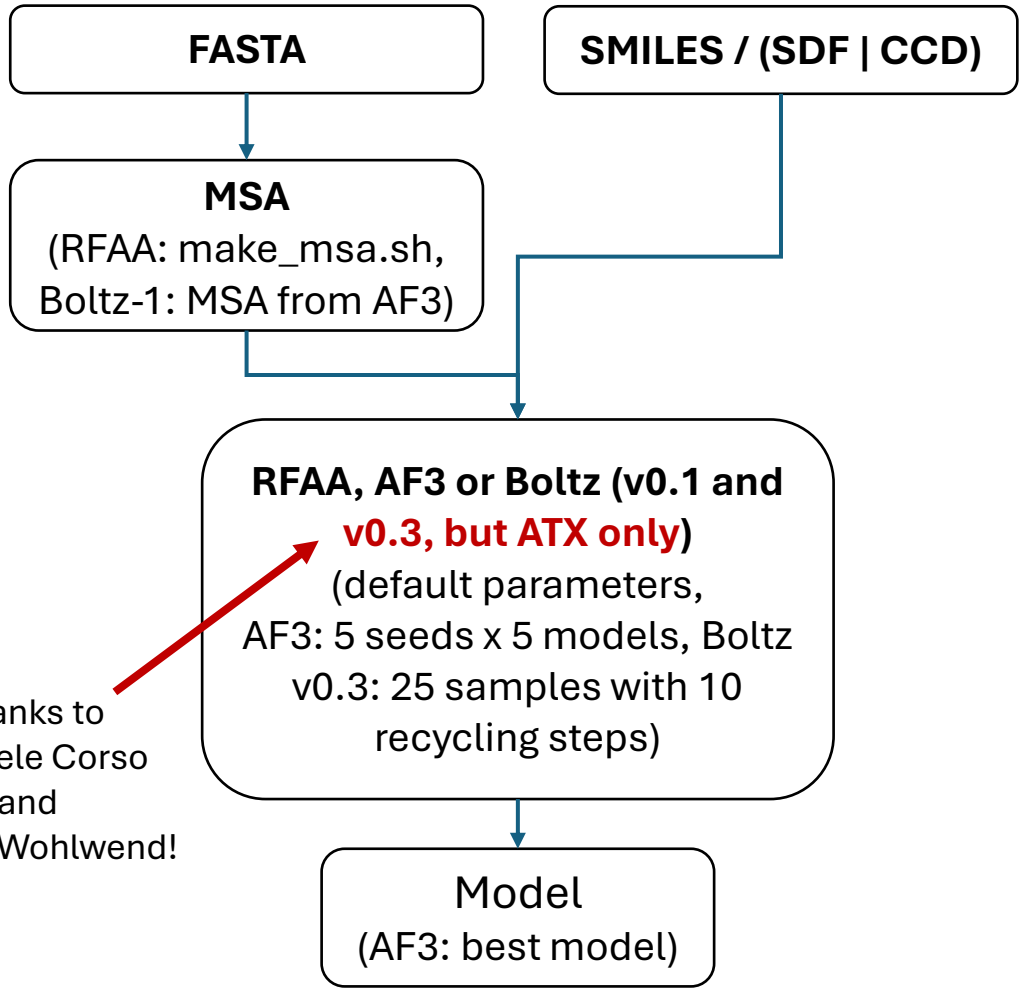
Janani Durairaj

- Pose prediction baselines:
  - Stage 1: AlphaFold3, Boltz-1, RosettaFold All-Atom, ~~Chai-1~~ and naïve docking
- Affinity prediction baselines:
  - Stage 1: Molecular weight, LogP, ML (public data) or AutoDock-Vina scores
  - Stage 2: AutoDock-Vina or GNINA docking scores

Peter Škrinjar

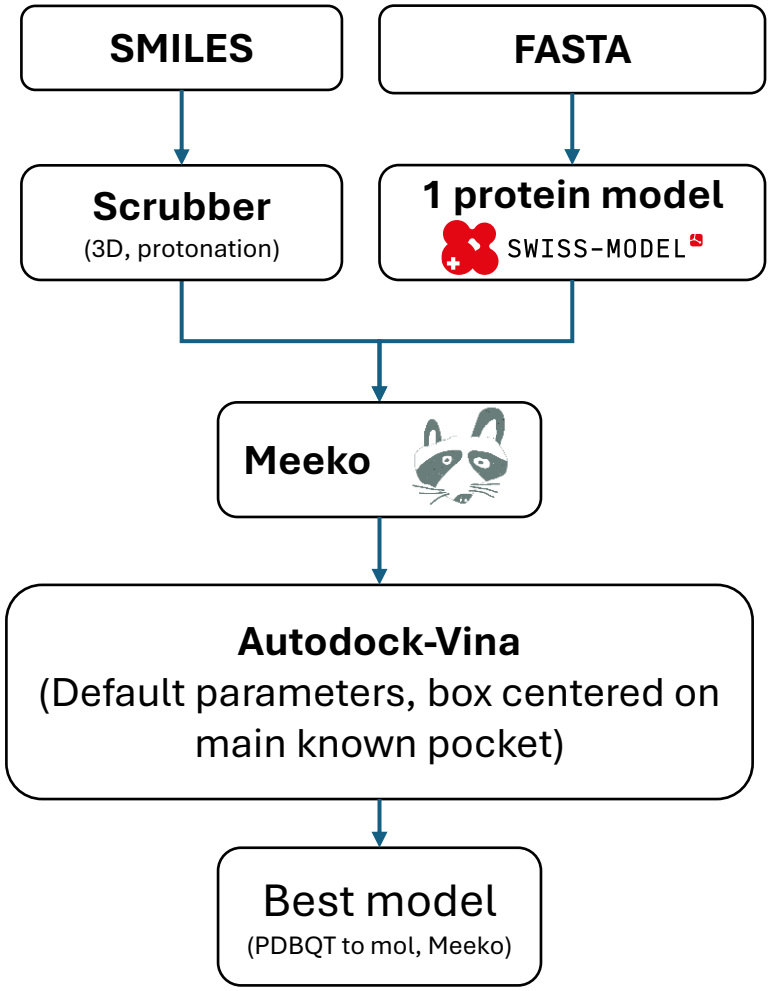
Pose prediction baselines

RosettaFold All-Atom, AF3 & Boltz-1



Thanks to Gabriele Corso and Jeremy Wohlwend!

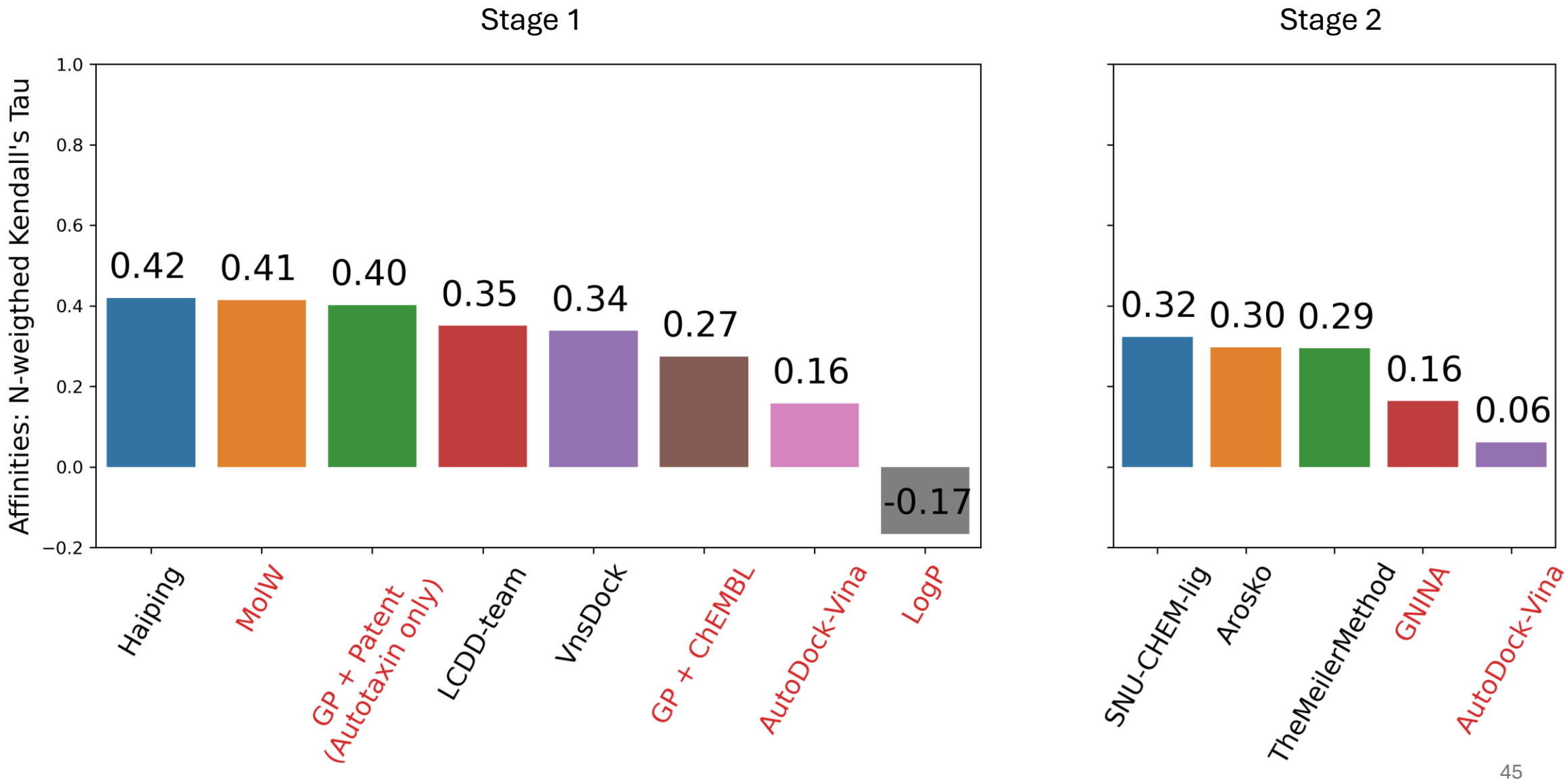
"Naive" docking



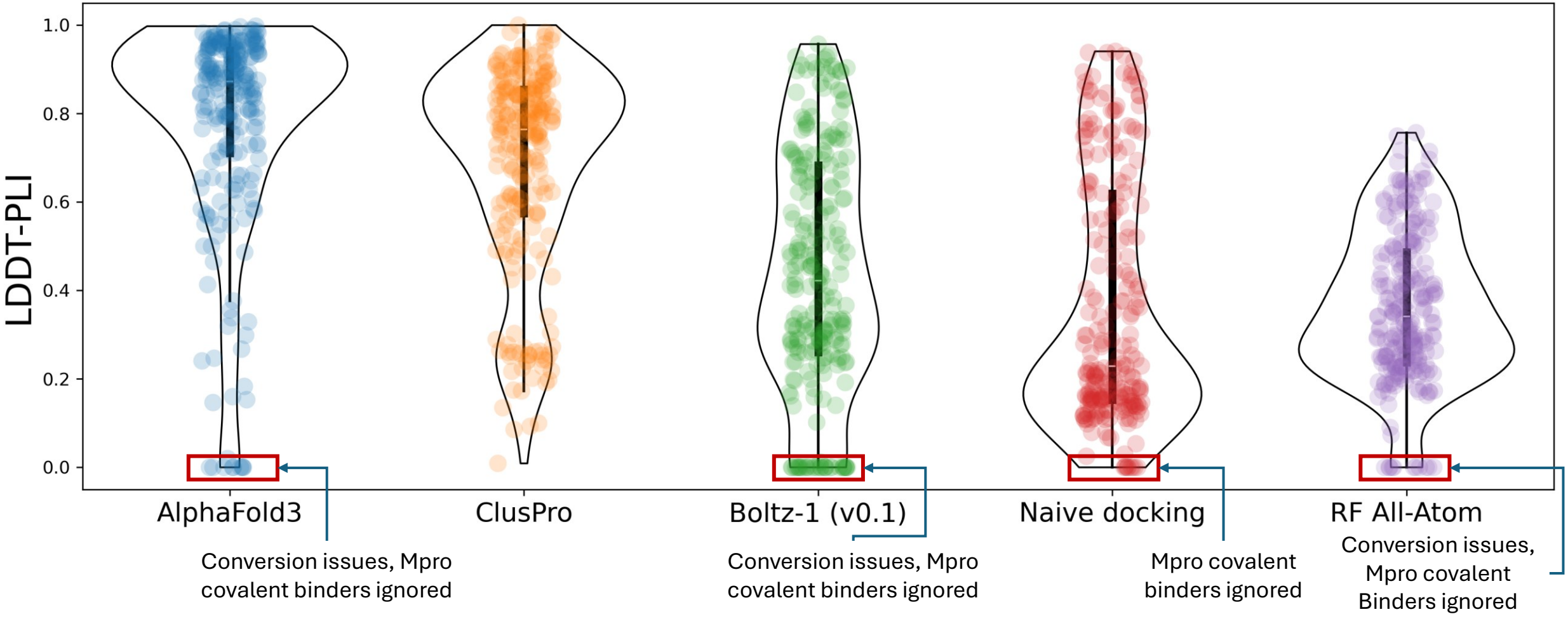
Affinity prediction baselines

- Naïve baselines (stage 1):
  - Molecular weight (interaction surface) or LogP (hydrophobicity)
- Machine-Learning baselines (stage 1):
  - Gaussian Process + RBF kernel + MAP4 fingerprint (Capecchi *et al.*, 2020)
  - Either trained on ChEMBL or patent data (P task) (Autotaxin only)
- Molecular docking baselines (stages 1 and 2):
  - Stage 1: AutoDock-Vina docking scores from the “naïve docking” baseline
  - Stage 2: AutoDock-Vina and GNINA scoring functions

Molecular weights is correlated to binding affinities (L1000/L3000)

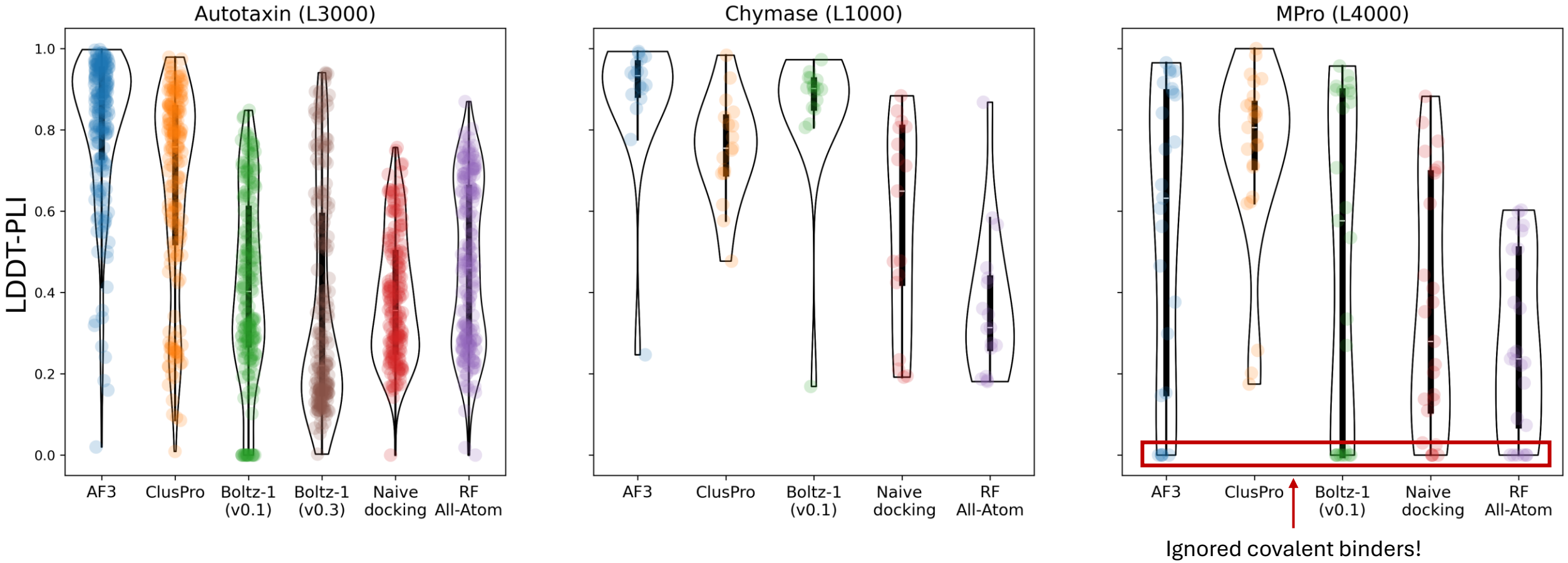


Overall LDDT-PLI performances per docking method



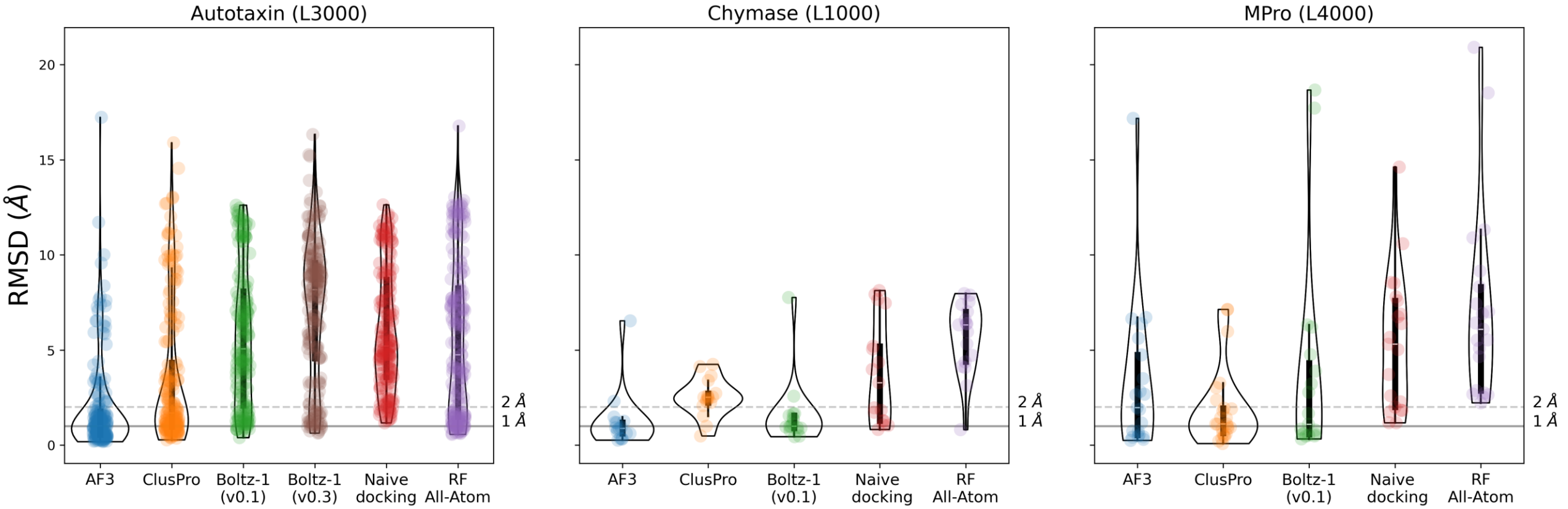
IT DOES NOT MEAN THAT IT IS GENERALIZABLE!

Overall LDDT-PLI performances per target protein



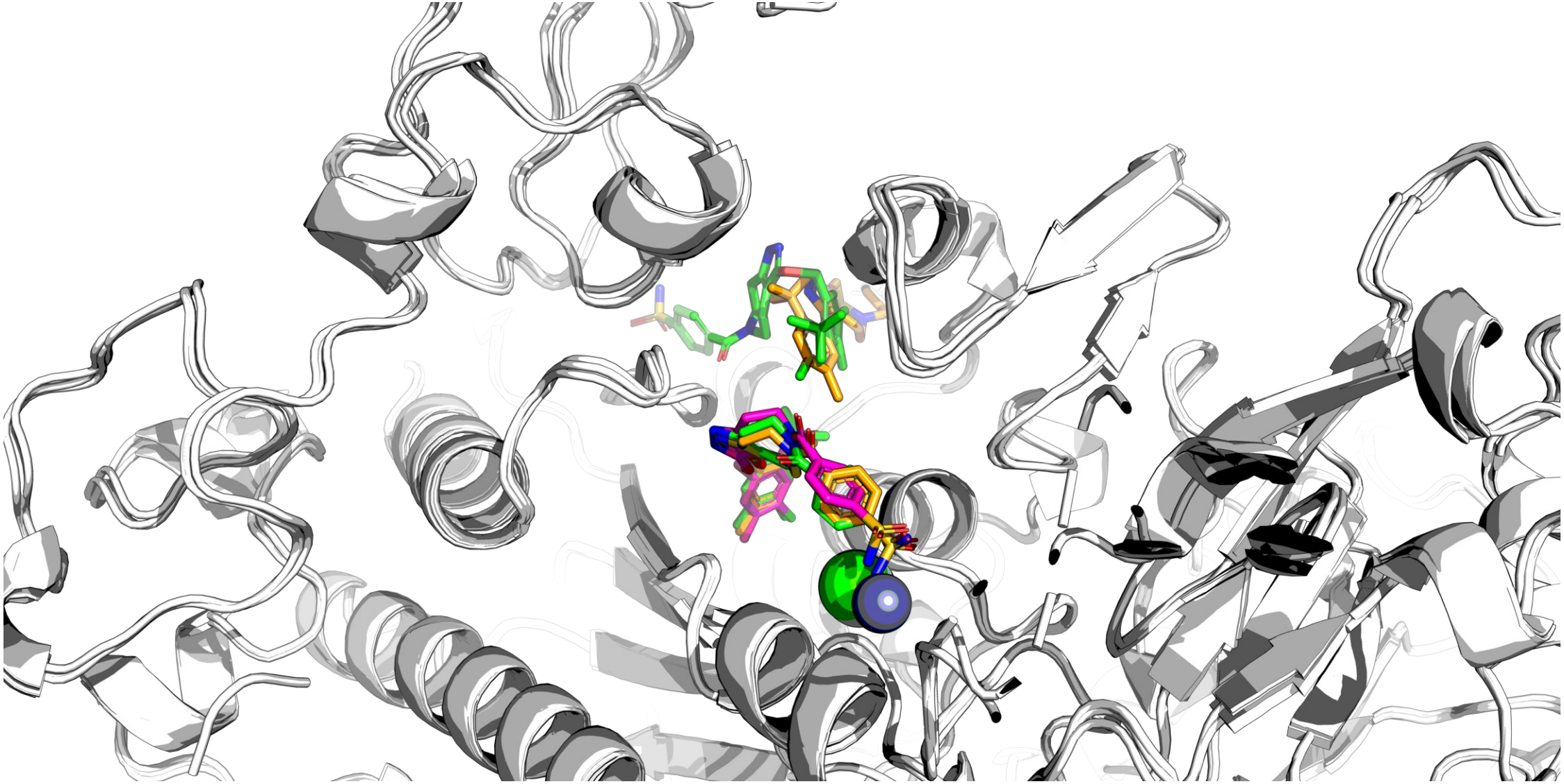
REALLY, IT DOES NOT MEAN IT IS GENERALIZABLE!

Overall RMSD performances per target protein



I GUESS YOU GOT IT NOW, DOES NOT SHOW IT GENERALIZES!

One example from the Autotaxin dataset



Target ID: L3019 – 2 copies of the same ligand (X-Ray: green, AF3: orange, ClusPro: pink)

What did we learn from these baselines?

- Template-based methods are strong baselines!

- Why is AlphaFold3 performing so well on the Autotaxin target?
  - 39 targets < 0.5 Å, 93 targets < 1 Å
  - Different time cutoff?
  - What is different between the paper version, the server and the public release?

- AF3 and template-based can be complementary methods (ex: Mpro)
  - If experimental data with close templates are available, use that
  - Otherwise AF3/Boltz can save you, but proceed with caution