Towards precision protein-ligand affinity prediction benchmark: A Complete and Modification-Aware DAVIS Dataset

Ming-Hsiu Wu[1], Ziqian Xie[1], Shuiwang Ji[2], Degui Zhi[1]

[1]McWilliams School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, USA

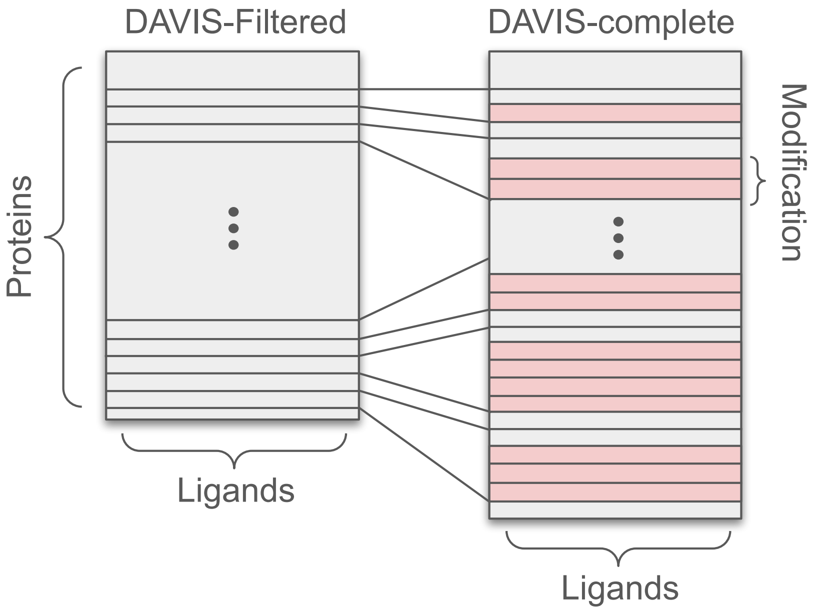
[2]Department of Computer Science and Engineering, Texas A&M University, USA

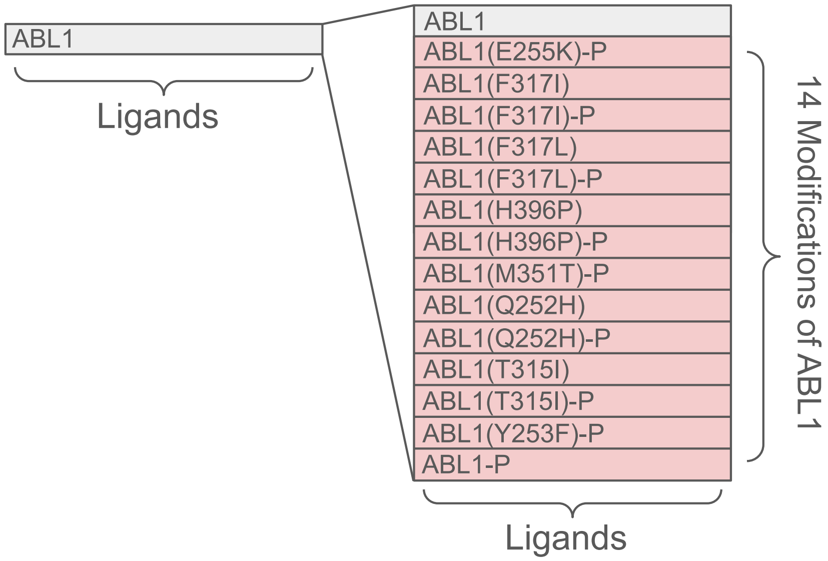
Background

Protein–ligand affinity modeling is limited by a lack of large, diverse, and experimentally homogeneous datasets that reflect real biology – especially protein modifications (substitutions, insertions, deletions, PTMs). Furthermore, most AI models focus on wild-type proteins or treat variants as equivalent (e.g., in **DAVIS**), leaving their ability to generalize to protein modifications largely unknown.

DAVIS-complete dataset

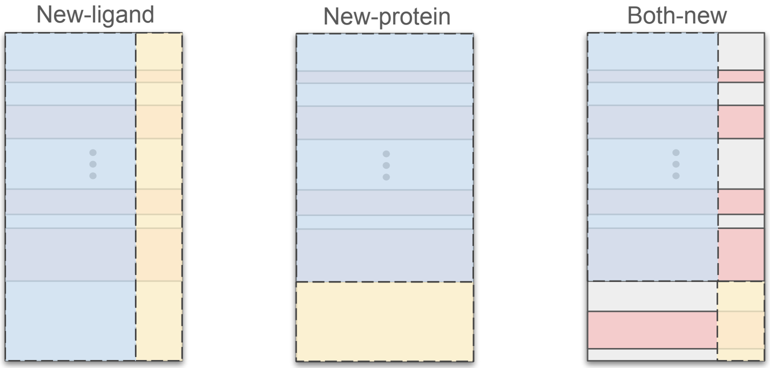
A curated, modification-aware version of the classic DAVIS kinase affinity dataset that now includes **4,032** additional kinase–ligand pairs covering substitutions, insertions, deletions, and phosphorylation across 56 modified sequences from 11 kinases.





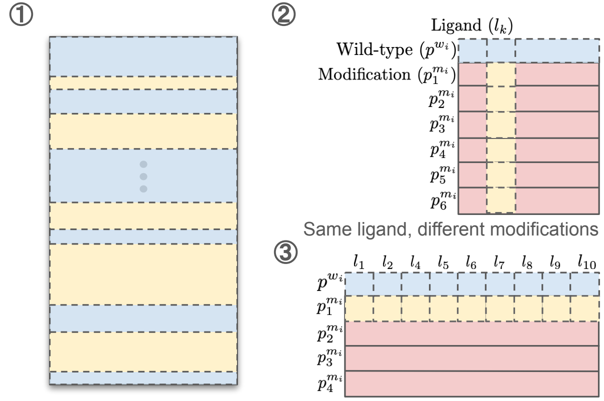
Benchmark Design

Augmented Dataset Prediction



🟦 Training data 🟨 Test data 🟩 Fine-tuning data

Wild-Type to Modification Generalization

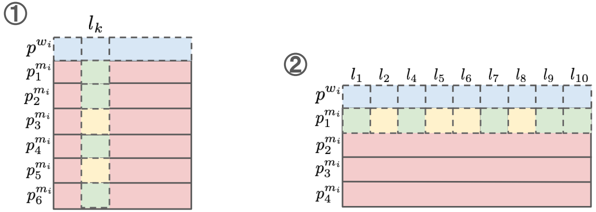


Same ligand, different modifications

Global modification generalization

Same modification, different ligands

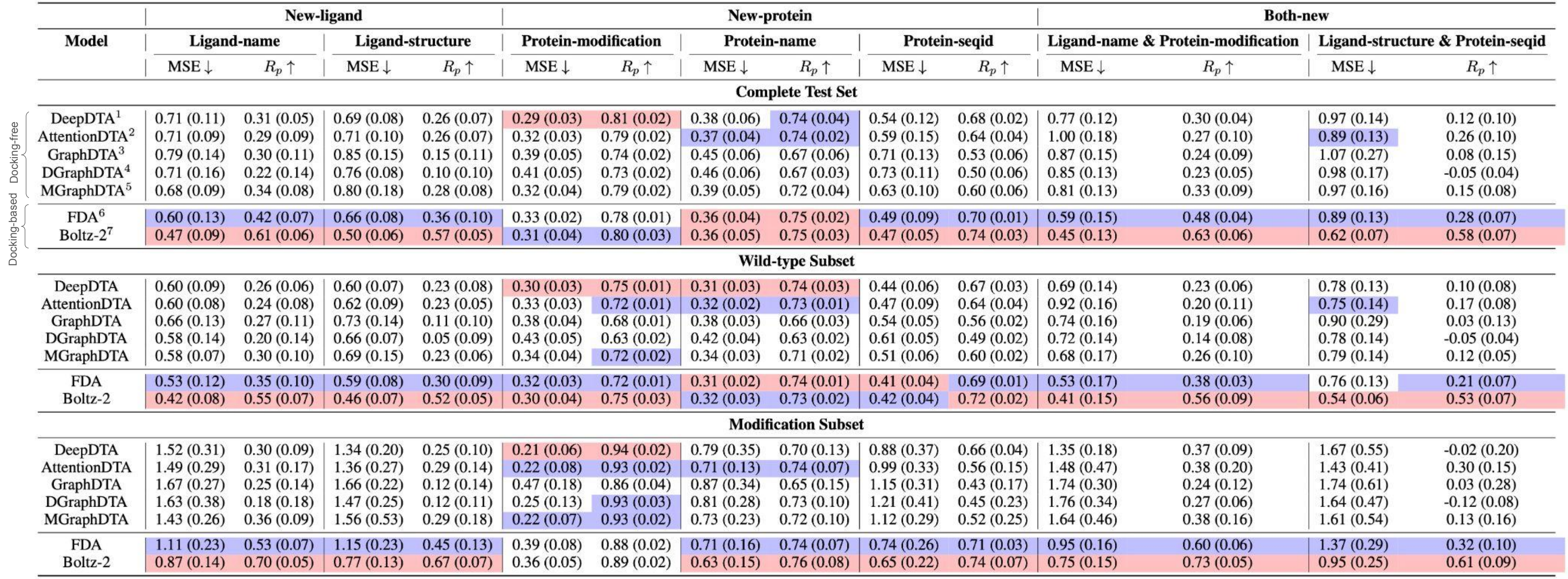
Few-Shot Modification Generalization



Same ligand, different modifications

Same modification, different ligands

Benchmark result – Augmented Dataset Prediction



		New-ligand				New-protein						Both-new			
	Model	**Ligand-name**		**Ligand-structure**		**Protein-modification**		**Protein-name**		**Protein-seqid**		**Ligand-name & Protein-modification**		**Ligand-structure & Protein-seqid**	
		MSE ↓	R_p ↑	MSE ↓	R_p ↑	MSE ↓	R_p ↑	MSE ↓	R_p ↑	MSE ↓	R_p ↑	MSE ↓	R_p ↑	MSE ↓	R_p ↑
							Complete Test Set								
Docking-free	DeepDTA[1]	0.71 (0.11)	0.31 (0.05)	0.69 (0.08)	0.26 (0.07)	0.29 (0.03)	0.81 (0.02)	0.38 (0.06)	0.74 (0.04)	0.54 (0.12)	0.68 (0.02)	0.77 (0.12)	0.30 (0.04)	0.97 (0.14)	0.12 (0.10)
	AttentionDTA[2]	0.71 (0.09)	0.29 (0.09)	0.71 (0.10)	0.26 (0.07)	0.32 (0.03)	0.79 (0.02)	0.37 (0.04)	0.74 (0.02)	0.59 (0.15)	0.64 (0.04)	1.00 (0.18)	0.27 (0.10)	0.89 (0.13)	0.26 (0.10)
	GraphDTA[3]	0.79 (0.14)	0.30 (0.11)	0.85 (0.15)	0.15 (0.11)	0.39 (0.05)	0.74 (0.02)	0.45 (0.06)	0.67 (0.06)	0.71 (0.13)	0.53 (0.06)	0.87 (0.15)	0.24 (0.09)	1.07 (0.27)	0.08 (0.15)
	DGraphDTA[4]	0.71 (0.16)	0.22 (0.14)	0.76 (0.08)	0.10 (0.10)	0.41 (0.05)	0.73 (0.02)	0.46 (0.06)	0.67 (0.03)	0.73 (0.11)	0.50 (0.06)	0.85 (0.13)	0.23 (0.05)	0.98 (0.17)	-0.05 (0.04)
	MGraphDTA[5]	0.68 (0.09)	0.34 (0.08)	0.80 (0.18)	0.28 (0.08)	0.32 (0.04)	0.79 (0.02)	0.39 (0.05)	0.72 (0.04)	0.63 (0.10)	0.60 (0.06)	0.81 (0.13)	0.33 (0.09)	0.97 (0.16)	0.15 (0.08)
Docking-based	FDA[6]	0.60 (0.13)	0.42 (0.07)	0.66 (0.08)	0.36 (0.10)	0.33 (0.02)	0.78 (0.01)	0.36 (0.04)	0.75 (0.02)	0.49 (0.09)	0.70 (0.01)	0.59 (0.15)	0.48 (0.04)	0.89 (0.13)	0.28 (0.07)
	Boltz-2[7]	0.47 (0.09)	0.61 (0.06)	0.50 (0.06)	0.57 (0.05)	0.31 (0.04)	0.80 (0.03)	0.36 (0.05)	0.75 (0.03)	0.47 (0.05)	0.74 (0.03)	0.45 (0.13)	0.63 (0.06)	0.62 (0.07)	0.58 (0.07)
							Wild-type Subset								
	DeepDTA	0.60 (0.09)	0.26 (0.06)	0.60 (0.07)	0.23 (0.08)	0.30 (0.03)	0.75 (0.01)	0.31 (0.03)	0.74 (0.03)	0.44 (0.06)	0.67 (0.03)	0.69 (0.14)	0.23 (0.06)	0.78 (0.13)	0.10 (0.08)
	AttentionDTA	0.60 (0.08)	0.24 (0.08)	0.62 (0.09)	0.23 (0.05)	0.33 (0.03)	0.72 (0.01)	0.32 (0.02)	0.73 (0.01)	0.47 (0.09)	0.64 (0.04)	0.92 (0.16)	0.20 (0.11)	0.75 (0.14)	0.17 (0.08)
	GraphDTA	0.66 (0.13)	0.27 (0.11)	0.73 (0.14)	0.11 (0.10)	0.38 (0.04)	0.68 (0.01)	0.38 (0.03)	0.66 (0.03)	0.54 (0.05)	0.56 (0.02)	0.74 (0.16)	0.19 (0.06)	0.90 (0.29)	0.03 (0.13)
	DGraphDTA	0.58 (0.14)	0.20 (0.14)	0.66 (0.07)	0.05 (0.09)	0.43 (0.05)	0.63 (0.02)	0.42 (0.04)	0.63 (0.02)	0.61 (0.05)	0.49 (0.02)	0.72 (0.14)	0.14 (0.08)	0.78 (0.14)	-0.05 (0.04)
	MGraphDTA	0.58 (0.07)	0.30 (0.10)	0.69 (0.15)	0.23 (0.06)	0.34 (0.04)	0.72 (0.02)	0.34 (0.03)	0.71 (0.02)	0.51 (0.06)	0.60 (0.02)	0.68 (0.17)	0.26 (0.10)	0.79 (0.14)	0.12 (0.05)
	FDA	0.53 (0.12)	0.35 (0.10)	0.59 (0.08)	0.30 (0.09)	0.32 (0.03)	0.72 (0.01)	0.31 (0.02)	0.74 (0.01)	0.41 (0.04)	0.69 (0.01)	0.53 (0.17)	0.38 (0.03)	0.76 (0.13)	0.21 (0.07)
	Boltz-2	0.42 (0.08)	0.55 (0.07)	0.46 (0.07)	0.52 (0.05)	0.30 (0.04)	0.75 (0.03)	0.32 (0.03)	0.73 (0.02)	0.42 (0.04)	0.72 (0.02)	0.41 (0.15)	0.56 (0.09)	0.54 (0.06)	0.53 (0.07)
							Modification Subset								
	DeepDTA	1.52 (0.31)	0.30 (0.09)	1.34 (0.20)	0.25 (0.10)	0.21 (0.06)	0.94 (0.02)	0.79 (0.35)	0.70 (0.13)	0.88 (0.37)	0.66 (0.04)	1.35 (0.18)	0.37 (0.09)	1.67 (0.55)	-0.02 (0.20)
	AttentionDTA	1.49 (0.29)	0.31 (0.17)	1.36 (0.27)	0.29 (0.14)	0.22 (0.08)	0.93 (0.02)	0.71 (0.13)	0.74 (0.07)	0.99 (0.33)	0.56 (0.15)	1.48 (0.47)	0.38 (0.20)	1.43 (0.41)	0.30 (0.15)
	GraphDTA	1.67 (0.27)	0.25 (0.14)	1.66 (0.22)	0.12 (0.14)	0.47 (0.18)	0.86 (0.04)	0.87 (0.34)	0.65 (0.15)	1.15 (0.31)	0.43 (0.17)	1.74 (0.30)	0.24 (0.12)	1.74 (0.61)	0.03 (0.28)
	DGraphDTA	1.63 (0.38)	0.18 (0.18)	1.47 (0.25)	0.12 (0.11)	0.25 (0.13)	0.93 (0.03)	0.81 (0.28)	0.73 (0.10)	1.21 (0.41)	0.45 (0.23)	1.76 (0.34)	0.27 (0.06)	1.64 (0.47)	-0.12 (0.08)
	MGraphDTA	1.43 (0.26)	0.36 (0.09)	1.56 (0.53)	0.29 (0.18)	0.22 (0.07)	0.93 (0.02)	0.73 (0.23)	0.72 (0.10)	1.12 (0.29)	0.52 (0.25)	1.64 (0.46)	0.38 (0.16)	1.61 (0.54)	0.13 (0.16)
	FDA	1.11 (0.23)	0.53 (0.07)	1.15 (0.23)	0.45 (0.13)	0.39 (0.08)	0.88 (0.02)	0.71 (0.16)	0.74 (0.07)	0.74 (0.26)	0.71 (0.03)	0.95 (0.16)	0.60 (0.06)	1.37 (0.29)	0.32 (0.10)
	Boltz-2	0.87 (0.14)	0.70 (0.05)	0.77 (0.13)	0.67 (0.07)	0.36 (0.05)	0.89 (0.02)	0.63 (0.15)	0.76 (0.08)	0.65 (0.22)	0.74 (0.07)	0.75 (0.15)	0.73 (0.05)	0.95 (0.25)	0.61 (0.09)

[1](Öztürk et al., 2018), [2](Zhao et al., 2023), [3](Nguyen et al., 2020), [4](Jiang et al., 2020), [5](Yang et al., 2022), [6](Wu et al., 2025), [7](Passaro et al., 2025)

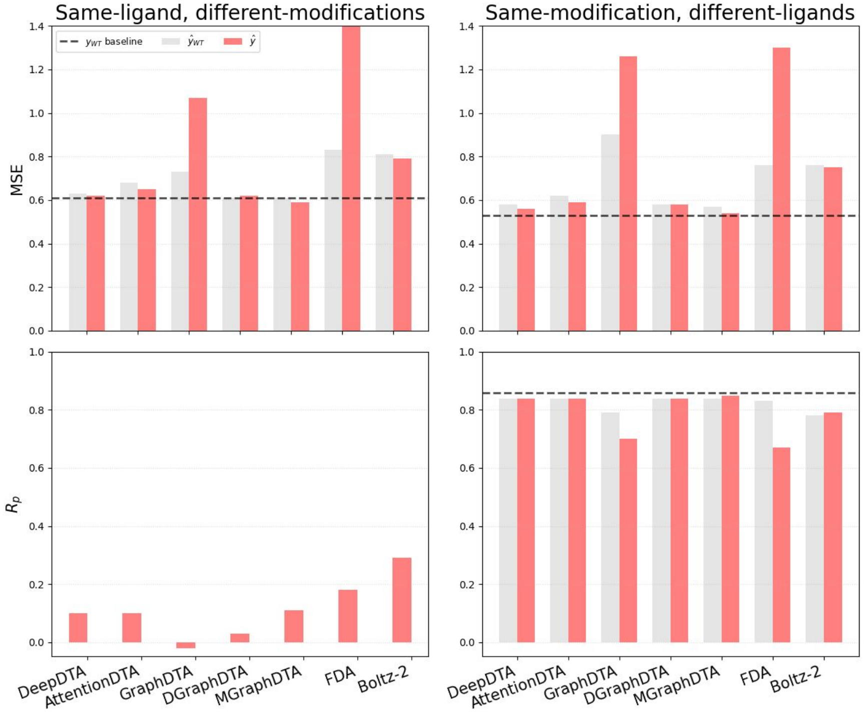
Red: Best, Blue: Second Best

Ligand-name: no ligand name overlaps between training and test sets; **Ligand-structure**: ligands in the test set have a Tanimoto similarity ≤ 0.5 (computed using Morgan fingerprints) to any ligand in the training set; **Protein-modification**: treating different modification variants of the same kinase as distinct unseen proteins (e.g., training on ABL1(Q252H) and testing on ABL1(T315I)); **Protein-name**: excluding all variants (including wild-type) of a protein from the test set if any variant appears in training; **Protein-seqid**: ensuring that kinases in the training set share ≤ 50% sequence identity with any kinase in the test set.

Benchmark result – WT-Modification Generalization

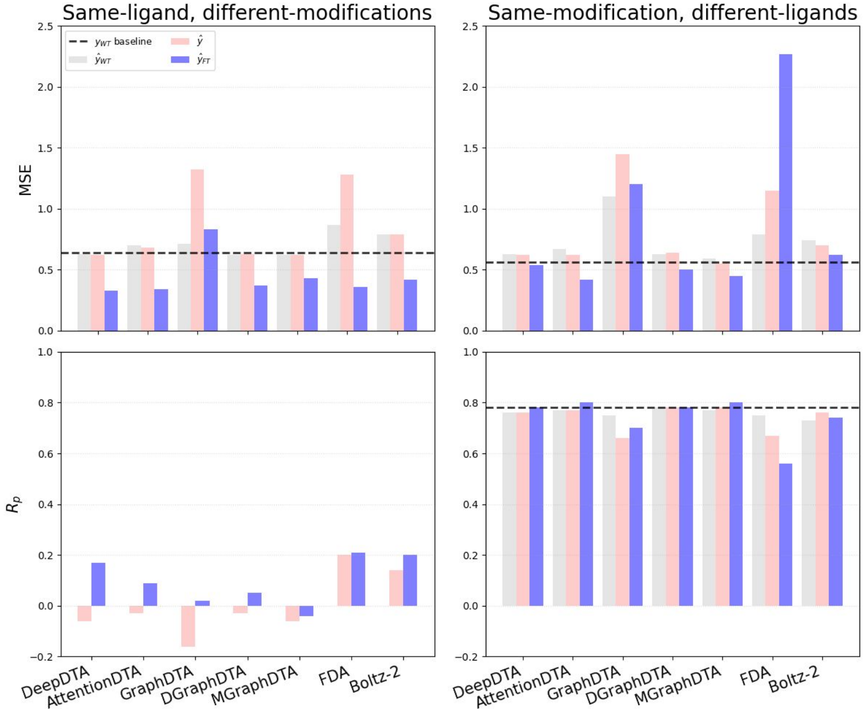
Zero-shot

Few-shot



Same-ligand, different-modifications

Same-modification, different-ligands



Same-ligand, different-modifications

Same-modification, different-ligands

Conclusion

- Based on experiment results, no model can effectively distinguish wild-type and modification proteins.
- Docking-based models show better generalizability in zero-shot scenarios, while docking-free methods tend to overfit to wild-type proteins.
- Few-shot fine-tuning can effectively improve docking-free models performance.