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]The University of Texas Health Science Center at Houston
[2]Texas A&M University

{ming.hsiu.wu,ziqian.xie,degui.zhi}@uth.tmc.edu, {sji}@tamu.edu

Abstract

Advancements in AI for science unlocks capabilities for critical drug discovery tasks such as protein-ligand binding affinity prediction. However, current models overfit to existing oversimplified datasets that does not represent naturally occurring and biologically relevant proteins with modifications. In this work, we curate a complete and modification-aware version of the widely used DAVIS dataset by incorporating 4,032 kinase–ligand pairs involving substitutions, insertions, deletions, and phosphorylation events. This enriched dataset enables benchmarking of predictive models under biologically realistic conditions. Based on this new dataset, we propose three benchmark settings—Augmented Dataset Prediction, Wild-Type to Modification Generalization, and Few-Shot Modification Generalization—designed to assess model robustness in the presence of protein modifications. Through extensive evaluation of both docking-free and docking-based methods, we find that docking-based model generalize better in zero-shot settings. In contrast, docking-free models tend to overfit to wild-type proteins and struggle with unseen modifications but show notable improvement when fine-tuned on a small set of modified examples. We anticipate that the curated dataset and benchmarks offer a valuable foundation for developing models that better generalize to protein modifications, ultimately advancing precision medicine in drug discovery. The benchmark is available at: https://github.com/ZhiGroup/DAVIS-complete

1 Introduction

Measuring protein-ligand binding affinity is a critical task in drug development, as it directly determines the therapeutic efficacy and selectivity of potential drug candidates [37]. AI breakthrough has revolutionized protein folding [4, 11, 16, 22], protein design [19, 18], and even protein-ligand binding [4, 7, 46, 8]. However, even with breakthroughs like AlphaFold [4, 11, 16] and DiffDock [7, 8], the protein-ligand affinity prediction problem is not solved yet. First, structural predictions from models like AlphaFold are AI estimations, not always equivalent to experimentally solved crystal structures [33, 17]. Second, even with solved protein structures, co-crystalized ligand-bound structures are often unavailable. Third, factors beyond direct structural complementarity, such as pH [30] and solvent effects [3], also significantly influence binding affinity. Current AI-driven affinity prediction methods are still in what might be termed a 'pre-AlphaFold era'; a significant portion operate as 'structure-free' (using 1D protein amino acid sequences) [55, 54, 50, 27] or 'docking-free' even with the incorporation of structural information [25, 15, 44].

A more pressing challenge in the field is the lack of large, diverse, and experimentally homogeneous training datasets [20] that adequately capture biological realities. In particular, protein modifica-

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

tions—such as substitutions, insertions, deletions, and post-translational modifications (PTMs)—can drastically alter protein structure and ligand interactions [38, 26, 48]. While generating such comprehensive datasets is challenging, current AI-driven models [21, 2, 34, 14, 52, 13, 53, 44, 28] only focus on wild-type proteins, overlooking modified protein versions or applying a simplistic "one-size-fits-all" approach to variants within datasets like DAVIS [10]. This oversight creates a significant gap in understanding how these models perform in real-world biological contexts, where proteins naturally undergo structural or chemical modifications. Models trained solely on such data may overfit to wild-type proteins and fail to generalize to more complex, yet practical, scenarios.

This study aims to bridge this critical gap. We introduce DAVIS-complete, a curated and complete version of the DAVIS dataset [10] that explicitly accounts for protein modifications, as illustrated in Fig. 1(a). Building upon this, we design three novel benchmark frameworks inspired by realistic drug discovery scenarios: (1) Augmented Dataset Prediction: We augment the previously used DAVIS dataset with modified protein–ligand pairs and evaluate model performance across three standard train-test splits, assessing general predictive capability in a diverse setting (Fig. 1(c)). (2) Wild-Type to Modification Generalization: This benchmark assesses a model's ability to generalize from wild-type proteins to unseen modified variants in a zero-shot setting, reflecting practical cases where experimental data for modified proteins are unavailable (Fig. 1(d)). (3) Few-Shot Modification Generalization: We further evaluate model adaptability by fine-tuning on a small number of modified protein–ligand pairs(Fig. 1(e)). This scenario mirrors precision medicine applications, where individualized therapies frequently rely on accurately predicting drug responses from limited genetic or proteomic data unique to each patient. Together, these benchmarks, for the first time, provide a more comprehensive and biologically relevant framework for evaluating binding affinity prediction models.

Our contributions are:

- The curation and public release of DAVIS-complete, a comprehensive dataset incorporating protein modifications for binding affinity prediction benchmark.
- The design and proposal of three biologically relevant benchmarks built upon DAVIS-complete.
- An extensive evaluation of existing state-of-the-art methods using these new benchmarks, highlighting current limitations (e.g., overfitting to wild-type proteins) and demonstrating the potential improvement (e.g., through fine-tuning strategies).

2 Related Works

2.1 Docking free-based models

In scenarios where high-resolution co-crystallized three-dimensional protein structures are unavailable, most existing deep learning approaches for predicting protein-ligand binding affinity operate without considering explicit binding poses—commonly referred to as docking-free methods. These models often represent proteins using amino acid sequences or predicted protein contact maps, while ligands are depicted as SMILES strings or molecular graphs. Deep neural networks are then employed to extract latent features from these representations to predict binding affinities. Notable models in this category include DeepDTA [55], AttentionDTA [54], GraphDTA [27], DGraphDTA [15], and MGraphDTA [50]. These methods have significantly advanced the field by circumventing the high cost of experimentally determining protein–ligand binding conformations. However, due to the nature of their input representations, these models are inherently limited in their ability to capture structural alterations caused by protein mutations or PTMs. This limitation is particularly important, as such modifications frequently occur in biological systems and could substantially influence protein-ligand binding affinity.

2.2 Docking-based models

When high-resolution co-crystallized three-dimensional structures are available, docking-based approaches have also made strides in modeling protein-ligand interactions by explicitly considering atom-level interaction details. Unlike docking-free models, these methods incorporate spatial information about the binding pose, allowing for a more accurate depiction of the interaction landscape.

Notable examples include SchNet [36], EGNN [35], and GIGN [51], which leverage 3D convolutional networks or equivariant graph neural networks to process molecular structures and predict binding affinity directly from geometric configurations. However, the applicability of docking-based methods is limited by the availability of high-quality co-crystallized 3D structures.

Building on this direction, the Folding-Docking-Affinity (FDA) [47] provides a framework for binding affinity prediction in scenarios where experimentally determined co-crystallized structures are unavailable. It unifies protein structure prediction, molecular docking, and binding affinity estimation into a single pipeline. FDA employs predicted 3D protein structures (e.g., from AlphaFold [11, 4]) and ligand binding poses generated through docking methods (e.g., DiffDock [7]) to construct realistic protein-ligand complexes at scale. Despite potential noise, FDA explicitly models atom-level interactions within predicted complexes to capture spatial information for binding affinity prediction. In parallel, recent co-folding models that jointly fold proteins and ligands—such as AlphaFold3 [4], Chai-1 [41], Protenix [40], and Boltz-1 [46]—have advanced binding structure generation. Building on this line, Boltz-2 [32] augments its predecessor with an affinity module to predict protein–ligand binding affinity.

2.3 Datasets

Two widely used datasets for evaluating the performance of deep learning-based protein-ligand affinity prediction models are DAVIS [10] and KIBA [39]. The DAVIS dataset focuses on kinase–ligand interactions and provides binding affinity values measured as dissociation constants (K_d). These measurements offer an experimentally homogeneous, high-quality, and biologically meaningful ground truth, making the dataset suitable for assessing model performance in kinase-targeted drug discovery. On the other hand, the KIBA dataset aggregates various bioactivity measurements, including K_i, K_d, and IC_{50} values, into a unified KIBA score, providing a broader yet noisier representation of drug-target interactions across a diverse set of kinases and compounds. These datasets have served as the standard benchmarks for docking free-based models like DeepDTA [55], GraphDTA [27], and their variants [54, 50, 25], allowing for consistent performance comparisons across different protein-ligand representation and model architectures.

2.4 Related Datasets on Modification-aware Binding

The PSnpBind dataset [5] offers a resource for studying the impact of single-point mutations at protein binding sites on ligand binding affinity. However, its reliance on traditional molecular docking methods may hinder its acceptability, as experimental assays are still the gold standard. The predicted binding conformations are static and may not capture the dynamic, context-dependent effects of mutations. Additionally, the empirical scoring functions used in docking often fail to accurately reflect changes in binding affinity, particularly in mutated proteins [31].

Another large-scale dataset is BindingDB [24], which contains approximately 3 million experimentally measured binding affinity data points, including both modified and unmodified proteins. Despite its scale, the dataset suffers from heterogeneity in assay types, experimental conditions, and reporting formats, leading to inconsistencies that impede data integration and limit its utility for predictive modeling. For example, a study by Landrum et al. [20] have demonstrated that combining IC_{50} or K_i values from different sources introduces significant noise.

In contrast, the DAVIS dataset offers an experimentally homogeneous, high-quality resource on kinase protein–ligand interactions. In addition to wild-type proteins, it also includes numerous data points involving modified kinase proteins. Previous studies using this dataset [55, 54, 27, 15, 50, 25], however, typically treated modified and unmodified kinases as equivalent or excluded the proteins with modifications altogether. Such modifications could significantly affect binding affinity predictions in certain cases. Therefore, indiscriminately incorporating them into predictive models without accounting for their differences-or simply discarding them-may not leveraging the full value of this dataset. Worse, models trained on such oversimplified DAVIS dataset may even overfit the wild-type proteins. Of note, this study aims to curate a complete version of the DAVIS dataset that accounts for all the modified kinases mentioned. Furthermore, the complete dataset is utilized to benchmark previously proposed docking-free methods as well as the recently published docking-based approach, Folding-Docking-Affinity (FDA) [47] and Boltz-2 [32].

3 A complete version of DAVIS dataset

The DAVIS dataset [10] covers interactions of 442 kinase proteins with 72 kinase inhibitors. Kinase proteins are represented by their Entrez Gene Symbols and corresponding names, with modification annotations included when applicable. This protein collection primarily focuses on catalytically active human protein kinase domains across the eight major typical kinase groups, representing over 80% of the human protein kinome. The dataset was primarily curated to analyze the selectivity of kinase inhibitors by examining small molecule-kinome interaction patterns. 31,824 binding affinity measurements (K_d) were simultaneously determined using a biochemical assay panel developed for this purpose. The consistency of the assay conditions minimizes variations in the experimental settings, which is a common concern found in other heterogeneous datasets [23, 39, 20].

This comprehensive assay provides critical insights into how protein modifications affect kinase binding affinity, which is vital for drug discovery. For example, the T790M mutation in EGFR reduces Lapatinib binding affinity by approximately 360-fold compared to the wild-type, demonstrating the significant impact of single-point mutations. Similarly, kinase conformational states, regulated by phosphorylation, influence inhibitor binding, as seen with Imatinib, a type II inhibitor, which binds more strongly to the inactive conformation of ABL1 kinase than its active state. These findings highlight the importance of considering kinase conformational dynamics in designing targeted therapies.

To include these modified kinase proteins, Entrez Gene Symbols in the dataset were mapped to UniProt IDs, and the corresponding amino acid sequences were retrieved from the UniProt database [1]. We then manually curated 56 modified amino acid sequences for 11 kinase proteins based on available annotations, including substitutions, insertions, deletions, phosphorylations, or any combinations. This process added 4,032 new modified protein-ligand pair data points (56 sequences * 72 ligands) to the dataset. Notable examples include ABL1 variants (e.g., T315I, H396P, F317I) with or without Tyr393 phosphorylation (Fig. 1(b)), EGFR mutations (L858R, T790M), and the FLT3-ITD found in the MV4;11 AML cell line [45]. Moreover, we refined existing entries for 11 kinase proteins (such as JAK, TYK2, and RSK family members) to include annotations of multiple specific domains rather than merely full-length sequences—a distinction also overlooked in previous studies [55, 54, 27, 15, 50], either. We updated these sequences by meticulously selecting domain boundaries based on relevant literature [12] and UniProt annotations [1]. Details of protein modifications are provided in Table. S1.

To formalize our extension of the DAVIS dataset by including modified kinase proteins, we introduce the following notation: Let $P^w = \{p^{w_i} \mid i = 1, 2, 3, \ldots, |P^w|\}$ denote the set of wild-type kinase proteins from the DAVIS dataset. For kinase proteins with modification variants, define: $P^m = \{p^{m_i} \mid i = 1, 2, 3, \ldots, |P^m|\}$ where P^m encompasses all modified variants across proteins, and each p^{m_i} specifically denotes the set of modified variants for a given protein. Each modified kinase variant within p^{m_i} is represented by $p_j^{m_i}$, corresponding to a specific type of modification (e.g., mutation, deletion, post-translational modification, or combinations thereof). Thus: $p^{m_i} = \{p_j^{m_i} \mid j = 1, 2, 3, \ldots |p^{m_i}|\}$, p^{m_i} can be the empty set if no modified variants are available in the dataset. To denote the combined set including both wild-type and modified kinase proteins, we introduce: $P^* = P^w \cup P^m$. The ligand set is denoted as: $L = \{l_k \mid k = 1, 2, 3, \ldots, |L|\}$, where each ligand l_k represents a distinct chemical compound in the dataset. $A(p, l)$ denotes the binding affinity between a protein p and a ligand l.

The DAVIS affinity distribution is dominated by capped measurements: approximately 70% of pairs are reported at $K_d > 10\mu M$ ($pK_d = 5$), over-representing weaker interactions. Among uncapped values, affinities center at $pK_d = 6.48 \pm 1.05$; median 6.24; IQR 5.68–7.08; range $[5.00, 10.80]$). To assess modification effects, we quantify the affinity alternation $\Delta pK_d = A(p_j^{m_i}, l_k) - A(p^{w_i}, l_k)$, which has mean -0.21 ± 0.84; median -0.04; IQR -0.59–0.30; range $[-4.49, 3.02]$. However, due to the $10\mu M$ cap, the exact ΔpK_d is unobservable for 60% of modified pairs, where the WT, the modification, or both exceed this threshold. Additional details are provided in section S2-S3.

4 Benchmark Design

We aim to assess whether the proposed state-of-the-art deep learning models can accurately predict binding affinity by distinguishing subtle differences among protein modifications. To reflect realistic

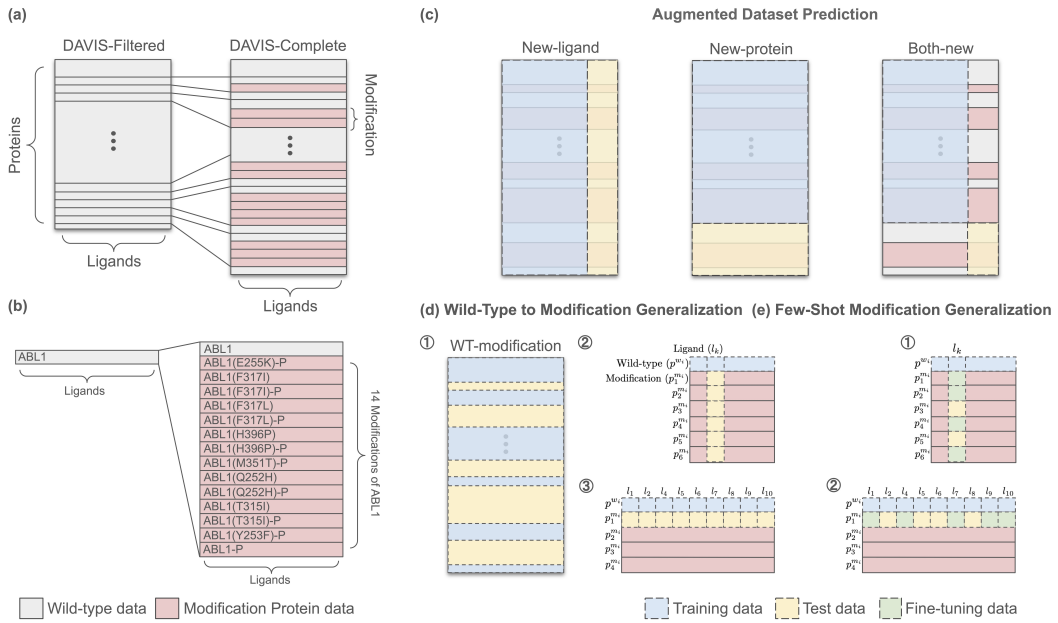

Figure 1: (a) DAVIS-Complete is curated by adding modified kinase protein–ligand pairs previously excluded from DAVIS-Filtered. (b) Example of dataset extension: 14 modifications of the kinase ABL1 are incorporated alongside its wild-type form. (c) Augmented Dataset Prediction benchmark: Wild-type and modified protein–ligand pairs are combined and evaluated under three main splits—new-drug, new-protein, and both-new—each with corresponding sub-splits. (d) Wild-Type to Modification Generalization benchmark: models trained on wild-type pairs are evaluated across (1) global modification generalization, (2) same-ligand different-modifications, and (3) same-modification different-ligands. (e) Few-Shot Modification Generalization: models fine-tuned on limited modified pairs to assess generalization to unseen variants.

drug discovery scenarios, we design three distinct benchmarking settings—summarized in Table 1—to evaluate the model's predictive performance.

Augmented Dataset Prediction We augment the DAVIS dataset used in prior studies [55, 54, 27, 15, 50, 25] by adding modified proteins that were previously ignored. Following prior work [25, 47], we evaluate model performance under three main train–test splits (Figure 1(c)). In all cases, both wild-type and modified protein–ligand pairs are included and mixed in the training and test sets, denoted as P^*L. Each main split has corresponding sub-splits. For the new-ligand split, the ligand-name setting ensures no ligand name overlaps between training and test sets, whereas the stricter ligand-structure setting requires that ligands in the test set have a Tanimoto similarity ≤ 0.5 (computed using Morgan fingerprints) to any ligand in the training set. For the new-protein split, the protein-modification setting treats different modification variants of the same kinase as distinct unseen proteins (e.g., training on ABL1(Q252H) and testing on ABL1(T315I)); the protein-name setting excludes all variants (including wild-type) of a protein from the test set if any variant appears in training; and the protein-seqid setting, the strictest version, ensures that kinases in the training set share $\leq 50\%$ sequence identity with any kinase in the test set. Combining the new-ligand and new-protein strategies yields six both-new configurations. Our benchmark includes the most lenient (ligand-name & protein-modification) and the strictest (ligand-structure & protein-seqid) configurations. The train/validation/test split ratio is kept as close as possible to 70%/10%/20%. The details of model training can be found in Table. S3. Binding affinity prediction performances are evaluated using mean squared error (MSE) and Pearson correlation coefficient (R_p).

Wild-Type to Modification Generalization To assess how well models transfer binding-affinity prediction from wild type to modified proteins, we train each model exclusively on wild-type protein–ligand pairs (P^wL) and evaluate under three biologically motivated settings. We report MSE, R_p, and C-index, and compare against two informative baselines: (i) wild-type ground truth

(y_{WT}), which predicts a modified pair's affinity by reusing the measured affinity of its corresponding wild-type pair; and (ii) wild-type prediction ($\hat{y}_{\text{WT}}$), which reuses the model's prediction for the wild-type pair. Models that do not surpass these baselines fail to capture modification-specific effects beyond what is already implied by the wild type. The evaluation settings are: (1) Global modification generalization: The model is evaluated on a broad set of modified protein-ligand pairs (Figure 1(d-1)). It reflects the challenge of predicting binding affinity across diverse protein variants arising from genetic mutations, deletions, or PTM—common in cancer, infectious diseases, and personalized medicine contexts. (2) Same-ligand, different-modifications: The model is tested on multiple distinct modifications of a single kinase, all bound to the same ligand (Figure 1(d-2)). This setting mimics drug resistance studies, where a therapeutic compound must be evaluated across different mutation profiles of a known target protein (e.g., EGFR inhibitors in lung cancer [9, 29]). (3) Same-modification, different ligands: The model predicts binding affinity for a set of ligands against a single modified kinase (Figure 1(d-3)). This scenario supports modification-specific drug screening, where the goal is to identify new compounds that effectively bind a disease-relevant mutant protein and potentially overcome resistance to existing therapies. The section S7 provides further details for the baseline calculation.

Few-Shot Modification Generalization Building on the same-ligand, different-modifications and same-modification, different-ligands scenarios from the Wild-Type to Modification Generalization benchmark, we further examine model adaptability by fine-tuning on a limited set of modified protein-ligand pairs, as illustrated in Figure. 1(e). 80% of the available modified protein–ligand pairs are used for model fine-tuning, and the remaining 20% for evaluation. The details of model fine-tuning can be found in Table. S4. This few-shot generalization scenario closely aligns with precision medicine contexts, where personalized treatments often depend on accurately predicting drug responses from sparse, patient-specific genetic or proteomic data. Enhancing model performance in such settings is essential for effectively guiding individualized therapeutic decisions.

Table 1: Summary of benchmarks, sub-tasks, and dataset splits. Training, fine-tuning, and test sets are represented using the introduced notations.

Benchmark	Sub-task	Training	Fine-tuning	Test
Augmented Dataset Prediction	New-ligand	P^*L	-	P^*L'
	New-protein	P^*L	-	$P^{*\prime}L$
	Both-new	P^*L	-	$P^{*\prime}L'$
Wild-Type to Modification Generalization	Global modification generalization	P^wL	-	P^mL
	Same-ligand, different-modifications	P^wL	-	$p^{m_i}l_k$
	Same-modification, different-ligands	P^wL	-	$p_j^{m_i}L$
Few-Shot Modification Generalization	Same-ligand, different-modifications	P^wL	$p_j^{m_i}l_k$	$p_{j'}^{m_i}l_k$
	Same-modification, different-ligands	P^wL	$p_j^{m_i}l_k$	$p_j^{m_i}l_{k'}$

5 Experiments

We benchmark five docking-free models—DeepDTA [55], AttentionDTA [54], GraphDTA [27], DGraphDTA [15], and MGraphDTA [50]—and two docking-based models, FDA [47] and Boltz-2 [32], on the curated, complete DAVIS dataset. Details of input preprocessing and all models are provided in Section S4-S5. All models are trained from scratch except Boltz-2, which we evaluate in inference-only mode on the test sets (no fine-tuning or hyperparameter optimization). Because of this mismatch in training protocol, Boltz-2 is not directly comparable; its numbers are reported for reference only and excluded from model rankings. Our evaluation covers following three benchmarks:

5.1 Augmented Dataset Prediction

We define seven train–test split settings to evaluate prediction performance: ligand-name, ligand-structure, protein-modification, protein-name, protein-seqid, ligand-name & protein-modification, and ligand-structure & protein-seqid. Table 2 reports results on the complete test set, as well as on two subsets: one containing wild-type protein–ligand pairs (wild-type subset) and the other containing modified protein–ligand pairs (modification subset). Across all splits except protein-modification and protein-name, the FDA method consistently outperforms docking-free models, achieving higher

R_p and lower MSE values. This trend holds for both the wild-type and modification subsets. In the most challenging both-new split (ligand-structure & protein-seqid), AttentionDTA performs on par with FDA. In the protein-modification split, all models except Boltz-2 generally demonstrate stronger performance compared to other train-test splits (MSE < 0.5, R_p > 0.6). However, the FDA model loses its top-ranked position and is outperformed by three comparably performing docking-free models—DeepDTA, AttentionDTA, and MGraphDTA—on the complete test set. On the modification subset, FDA exhibits a more pronounced decline in ranking, placing second to last—only ahead of GraphDTA.

The observation from these splits suggests that binding affinity prediction performance depends strongly on whether proteins or ligands are seen during training. For new-ligand tests, R_p is consistently higher under the ligand-name split than under the stricter ligand-structure split, reflecting the added difficulty of enforcing structural novelty. In the new-protein splits, similarly, performance declines as the test proteins become more dissimilar to those in training (protein-modification $\rightarrow$ protein-name $\rightarrow$ protein-seqid). Comparing new-ligand and new-protein splits, models generally perform worse in new-ligand, indicating a higher dependency on ligand familiarity. Overall, models perform worse in the both-new setting than in the corresponding new-ligand or new-protein splits, with the strictest ligand-structure & protein-seqid configuration yielding the lowest performance.

In particular, we found that this dependency is even more evident among docking-free methods. By examining the degree of performance decline across different splits, it is clear that docking-free models suffer sharper drops in accuracy when proteins, ligands, or both are not present in the training data, highlighting their stronger reliance on seen training examples, which is consistent with previous findings [43, 6, 49, 42]. Besides, when proteins and ligands are included in the training set, most of docking-free methods consistently outperform the docking-based FDA model. This suggests that docking-free approaches may be better at learning direct mappings between known protein-ligand pairs and their binding affinities, whereas the docking-based FDA model, which relies on binding conformation, may not benefit as much from the simple presence of proteins or ligands in the training phase.

5.2 Wild-Type to Modification Generalization

To assess the models' ability to generalize from wild-type to modified kinase proteins, we train each model exclusively on wild-type protein-ligand pairs (P^wL). We then evaluate their performance across three distinct test scenarios. In the first scenario, termed Global modification generalization (P^mL), all modified kinase proteins are included. We additionally stratify results into four subsets defined by whether the wild-type (WT) and modification affinities are capped or uncapped (Details in Section S3). Results are reported in Table 3. In the WT-uncapped & modification-uncapped subset, DeepDTA, AttentionDTA, DGraphDTA, and MGraphDTA perform similarly well on MSE, R_p, and C-index, while GraphDTA and FDA lag behind. However, for these docking-free models the predictions for modification pairs are highly correlated with their own WT predictions (high $R_p(\hat{y}, \hat{y}_{\text{WT}})$), whereas this correlation is much lower for the docking-based FDA. Notably, about 84% of affinity changes lie within $[-1, 1]$ in this category (Fig. S3(a)). For the stronger docking-free models, R_p is nearly identical to $R_p(y, \hat{y}_{\text{WT}})$, indicating overfitting to WT: because the modification–WT differences are majorly small, simply echoing the seen WT prediction yields seemingly strong performance. By contrast, in the WT-capped & modification-uncapped and WT-uncapped & modification-capped subsets, models can no longer rely on guessing the WT value; both R_p and C-index drop markedly, the advantage of WT-overfitting models diminishes, and FDA becomes relatively stronger. Finally, in the WT-capped & modification-capped subset, WT overfitting docking-free models again appear to perform well, mirroring the pattern observed in the WT-uncapped & modification-uncapped case.

Furthermore, in real-world biological scenarios, a kinase protein often exhibits multiple distinct mutations across different populations, potentially leading to varied binding affinities for the same ligand. To capture this biologically relevant variability, we introduce a second evaluation scenario—same-ligand, different-modifications—to examine whether models pre-trained solely on wild-type proteins can effectively distinguish variations in binding affinity caused by diverse protein modifications when interacting with the same ligand. Notably, in contrast to the global setting that mixes multiple ligands and kinases, this benchmark isolates a fixed kinase–ligand pair ($p^{m_i}l_k$) and restricts evaluation to WT-uncapped & modification-uncapped pairs, varying only the kinase modification to test fine-grained sensitivity. Results are shown in Table 4(a). In terms of MSE, DeepDTA, AttentionDTA,

DGraphDTA, and MGraphDTA perform comparably, whereas GraphDTA and FDA show weaker performance. Notably, only MGraphDTA nominally exceeds the y_{WT} baseline; however, given the large standard deviation, this difference is not meaningful. Simply using the wild-type ground-truth affinity (y_{WT}) matches or exceeds these models. Furthermore, the consistently low R_p (below 0.2) and marginally better-than-random C-index (just above 0.5) suggest that current models fail to capture or generalize protein modifications from the wild-type training data. Among these approaches, docking-free methods fare worse than docking-based ones. A case study on EGFR variants with staurosporine (Fig. S4) illustrates this: the docking-free MGraphDTA overfits to the wild type and produces nearly identical predictions across variants, whereas the docking-based FDA better tracks the affinity trends.

In another biologically relevant scenario, we may need to rank different ligands for a modified protein. To assess this, we introduce the third scenario—same-modification, different-ligands—which tests whether models trained only on wild-type proteins can distinguish ligand affinities for the same modified kinase. This benchmark fixes a kinase–ligand pair ($p^{m_i} l_k$) and also evaluates only WT-uncapped & modification-uncapped cases, varying only the ligand to probe sensitivity. The results of this evaluation are summarized in Table 4(b). The models perform notably better in the same-modification, different-ligands scenario compared to the same-ligand, different-modifications setting, particularly in terms of R_p and C-index. However, in the case of EGFR(L858R, T790M) with various ligands (Fig. S5), MGraphDTA predictions closely follow the binding affinity trend of the wild-type, again reflecting its tendency to overfit to wild-type data. Additionally, in most cases (44 out of 55), such as EGFR(G719C) (Fig. S6), we observe strong consistency between wild-type and modified protein-ligand affinity profiles, with R_p values above 0.8. This suggests that ligand often plays a more dominant role than protein modification, and the effect of modification on binding affinity is generally smaller. Consequently, the WT–overfitting docking-free models can still outperform the docking-based method in this scenario. Nonetheless, a model's ability to surpass the y_{WT} baseline remains a meaningful indicator of its sensitivity to subtle affinity shifts.

Table 2: Performance comparison of docking-free and docking-based methods on the complete test set, wild-type subset, and modification subset across seven train–test splits. Results are reported as mean (std) over five random splits. Pearson correlation coefficient (R_p) and Mean Squared Error (MSE) are computed from predicted vs. true pK_d values. Boltz-2 was evaluated in inference-only mode (no training on our dataset). Its results are shown for reference and are excluded from rankings.

	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														
DeepDTA	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	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	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	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	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)
FDA	**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	1.18 (0.08)	0.41 (0.08)	1.11 (0.13)	0.45 (0.05)	1.17 (0.03)	0.50 (0.01)	1.17 (0.03)	0.50 (0.03)	1.08 (0.03)	0.55 (0.02)	1.17 (0.10)	0.40 (0.09)	1.02 (0.11)	0.54 (0.06)
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	1.18 (0.08)	0.39 (0.08)	1.13 (0.13)	0.41 (0.05)	1.18 (0.02)	0.46 (0.02)	1.19 (0.04)	0.47 (0.02)	1.12 (0.04)	0.52 (0.01)	1.17 (0.09)	0.38 (0.09)	1.06 (0.12)	0.49 (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	1.18 (0.07)	0.53 (0.10)	0.97 (0.11)	0.63 (0.05)	1.10 (0.07)	0.64 (0.03)	1.12 (0.31)	0.64 (0.09)	0.96 (0.05)	0.66 (0.05)	1.21 (0.18)	0.50 (0.12)	0.84 (0.10)	0.69 (0.09)

5.3 Few-Shot Modification Generalization

In the same-ligand, different-modifications setting, the evaluation results are summarized in Table. 5(a), which reports model performance before ($\hat{y}$) and after fine-tuning ($\hat{y}_{\text{FT}}$) across three metrics: MSE, R_p, and C-index. All docking-free models show improved performance after fine-tuning, demonstrating the value of even limited modified kinase data in enhancing generalization at the protein modification level. However, in terms of R_p and C-index, all models still exhibit low performance—remaining below 0.6, which is often considered the threshold for effective prediction.

In the same-modification, different-ligands setting, the benchmark results are shown in Table. 5(b). All models except the docking-based FDA model similarly show noticeable improvements across

Table 3: Performance comparison of docking-free models and a docking-based method trained exclusively on wild-type protein–ligand pairs (P^wL) and evaluated on modified kinase protein–ligand pairs (P^mL). The P^mL test set is partitioned into four distinct subsets depending on whether affinity values are capped or not. Results are reported as mean (standard deviation) over five independent runs using identical train–test splits but different model parameter initialization. MSE, R_p, and C-index are computed between predicted and true pK_d values. Boltz-2 was evaluated in inference-only mode (no training on our dataset); its results are shown for reference and excluded from rankings.

Model	MSE ↓	R_p ↑	C-index ↑	$R_p(y, \hat{y}_{\text{WT}})$	$R_p(\hat{y}, \hat{y}_{\text{WT}})$	MSE ↓	R_p ↑	C-index ↑	$R_p(y, \hat{y}_{\text{WT}})$	$R_p(\hat{y}, \hat{y}_{\text{WT}})$
	WT-uncapped & modification-uncapped					**WT-capped & modification-uncapped**				
DeepDTA	0.63 (0.04)	0.79 (0.01)	0.79 (0.01)	0.79 (0.01)	1.00 (0.00)	0.35 (0.01)	0.11 (0.05)	0.53 (0.03)	0.08 (0.02)	0.87 (0.20)
AttentionDTA	0.66 (0.07)	**0.80** (0.02)	**0.80** (0.01)	0.79 (0.02)	0.99 (0.01)	0.37 (0.01)	0.06 (0.04)	0.50 (0.02)	0.05 (0.04)	0.84 (0.20)
GraphDTA	1.17 (0.12)	0.64 (0.02)	0.74 (0.01)	0.76 (0.03)	0.87 (0.02)	0.34 (0.03)	0.00 (0.07)	0.51 (0.02)	0.03 (0.07)	0.89 (0.09)
DGraphDTA	0.64 (0.02)	**0.80** (0.01)	**0.80** (0.00)	0.80 (0.00)	1.00 (0.00)	0.33 (0.01)	0.03 (0.06)	0.51 (0.02)	0.04 (0.06)	0.97 (0.04)
MGraphDTA	**0.61** (0.04)	**0.80** (0.01)	**0.80** (0.01)	0.79 (0.01)	0.99 (0.01)	0.37 (0.01)	0.05 (0.07)	**0.54** (0.03)	0.02 (0.08)	0.92 (0.09)
FDA	1.47 (0.05)	0.62 (0.01)	0.72 (0.00)	0.78 (0.02)	0.58 (0.02)	**0.30** (0.01)	**0.13** (0.02)	0.53 (0.01)	0.07 (0.07)	0.09 (0.09)
Boltz-2	1.17 (0.00)	0.54 (0.00)	0.68 (0.00)	0.54 (0.00)	0.94 (0.00)	0.61 (0.00)	0.24 (0.00)	0.58 (0.00)	0.15 (0.00)	0.93 (0.00)
	WT-uncapped & modification-capped					**WT-capped & modification-capped**				
DeepDTA	1.89 (0.20)	–	–	–	0.99 (0.00)	**0.01** (0.00)	–	–	–	0.96 (0.03)
AttentionDTA	2.06 (0.40)	–	–	–	0.93 (0.13)	**0.01** (0.01)	–	–	–	0.92 (0.04)
GraphDTA	1.76 (0.14)	–	–	–	0.99 (0.00)	0.03 (0.01)	–	–	–	0.67 (0.03)
DGraphDTA	1.92 (0.14)	–	–	–	1.00 (0.00)	**0.01** (0.00)	–	–	–	0.98 (0.00)
MGraphDTA	1.51 (0.14)	–	–	–	0.90 (0.04)	**0.01** (0.00)	–	–	–	0.95 (0.04)
FDA	**0.65** (0.03)	–	–	–	0.50 (0.05)	0.05 (0.00)	–	–	–	0.11 (0.03)
Boltz-2	1.46 (0.00)	–	–	–	0.70 (0.00)	1.02 (0.00)	–	–	–	0.92 (0.00)

Table 4: Wild-Type to Modification Generalization benchmark (a) Same-ligand, different-modifications: models are trained on all wild-type pairs and evaluated on modified variants of the same kinase protein with a fixed ligand. (b) Same-modification, different-ligands: models are evaluated on distinct ligands for a fixed kinase modification. Metrics are mean (std) across kinase–ligand combinations. Boltz-2 was evaluated in inference-only mode (no training on our dataset). Its results are shown for reference and are excluded from rankings.

Model	**MSE ↓**			**R_p ↑**			**C-index ↑**		
	y_{WT}	$\hat{y}_{\text{WT}}$	$\hat{y}$	y_{WT}	$\hat{y}_{\text{WT}}$	$\hat{y}$	y_{WT}	$\hat{y}_{\text{WT}}$	$\hat{y}$
(a) Same-ligand, different-modifications									
DeepDTA	0.61 (0.72)	0.63 (0.59)	0.62 (0.57)	–	–	0.10 (0.31)	–	–	0.53 (0.11)
AttentionDTA	0.61 (0.72)	0.68 (0.76)	0.65 (0.73)	–	–	0.10 (0.28)	–	–	0.53 (0.10)
GraphDTA	0.61 (0.72)	0.73 (0.68)	1.07 (1.46)	–	–	-0.02 (0.31)	–	–	0.50 (0.12)
DGraphDTA	0.61 (0.72)	**0.61** (0.65)	0.62 (0.64)	–	–	0.03 (0.26)	–	–	0.52 (0.11)
MGraphDTA	0.61 (0.72)	**0.61** (0.63)	**0.59** (0.61)	–	–	0.11 (0.32)	–	–	0.53 (0.12)
FDA	0.61 (0.72)	0.83 (1.04)	1.41 (1.89)	–	–	**0.18** (0.46)	–	–	**0.56** (0.18)
Boltz-2	0.61 (0.72)	1.06 (1.27)	1.06 (1.30)	–	–	0.25 (0.48)	–	–	0.59 (0.20)
(b) Same-modification, different-ligands									
DeepDTA	0.53 (0.59)	0.58 (0.49)	0.56 (0.47)	0.86 (0.16)	**0.84** (0.16)	0.84 (0.15)	0.84 (0.08)	**0.83** (0.08)	0.83 (0.07)
AttentionDTA	0.53 (0.59)	0.62 (0.57)	0.59 (0.53)	0.86 (0.16)	**0.84** (0.16)	0.84 (0.15)	0.84 (0.08)	**0.83** (0.08)	**0.84** (0.08)
GraphDTA	0.53 (0.59)	0.90 (0.66)	1.26 (1.08)	0.86 (0.16)	0.79 (0.15)	0.70 (0.26)	0.84 (0.08)	0.82 (0.06)	0.79 (0.09)
DGraphDTA	0.53 (0.59)	0.58 (0.50)	0.58 (0.49)	0.86 (0.16)	**0.84** (0.15)	0.84 (0.15)	0.84 (0.08)	**0.83** (0.07)	0.83 (0.07)
MGraphDTA	0.53 (0.59)	**0.57** (0.53)	**0.54** (0.50)	0.86 (0.16)	**0.84** (0.16)	**0.85** (0.14)	0.84 (0.08)	**0.83** (0.08)	**0.84** (0.08)
FDA	0.53 (0.59)	0.76 (0.68)	1.30 (0.66)	0.86 (0.16)	0.83 (0.16)	0.67 (0.17)	0.84 (0.08)	**0.83** (0.08)	0.75 (0.09)
Boltz-2	0.53 (0.59)	1.23 (0.46)	1.23 (0.48)	0.86 (0.16)	0.50 (0.23)	0.50 (0.24)	0.84 (0.08)	0.69 (0.14)	0.69 (0.14)

evaluation metrics—MSE, R_p, and C-index—after fine-tuning on few-shot samples. Among all models, AttentionDTA achieves the best overall performance, with its MSE decreasing from 0.62 to 0.42, R_p increasing from 0.77 to 0.80, and C-index improving from 0.81 to 0.82. These results suggest that AttentionDTA is particularly effective at adapting to ligand-induced variability. In contrast, the FDA model is the only method that does not benefit from fine-tuning; rather than improving, its performance deteriorates after incorporating the few-shot examples, suggesting a need for more effective fine-tuning strategies.

6 Limitation

Despite the addition of modified protein–ligand pairs, bringing the DAVIS dataset to 31,824 entries, its size remains limited for training data-intensive deep learning models. Its kinase-centric focus further restricts generalizability, as kinases represent only a fraction of the proteome. A more intrinsic

Table 5: Few-shot Modification Generalization benchmark. (a) Same-ligand, different-modifications: models are fine-tuned on limited modified protein–ligand pairs and evaluated on additional variants of the same kinase with a shared ligand. (b) Same-modification, different-ligands: models are fine-tuned and tested on distinct ligands targeting the same kinase modification. Metrics are mean (std) across kinase–ligand combinations.

Model	MSE ↓				R_p ↑				C-index ↑			
	y_{WT}	$\hat{y}_{\text{WT}}$	$\hat{y}$	$\hat{y}_{\text{FT}}$	y_{WT}	$\hat{y}_{\text{WT}}$	$\hat{y}$	$\hat{y}_{\text{FT}}$	y_{WT}	$\hat{y}_{\text{WT}}$	$\hat{y}$	$\hat{y}_{\text{FT}}$
(a) Same-ligand, different-modifications												
DeepDTA	0.64 (0.94)	0.63 (0.75)	**0.62** (0.73)	**0.33** (0.43)	–	–	-0.06 (0.51)	0.17 (0.52)	–	–	0.46 (0.24)	**0.56** (0.23)
AttentionDTA	0.64 (0.94)	0.70 (0.94)	0.68 (0.92)	0.34 (0.45)	–	–	-0.03 (0.54)	0.09 (0.61)	–	–	0.47 (0.26)	0.53 (0.28)
GraphDTA	0.64 (0.94)	0.71 (0.81)	1.32 (2.09)	0.83 (1.18)	–	–	-0.16 (0.68)	0.02 (0.71)	–	–	0.43 (0.33)	0.50 (0.34)
DGraphDTA	0.64 (0.94)	**0.62** (0.82)	0.63 (0.81)	0.37 (0.49)	–	–	-0.03 (0.46)	0.05 (0.51)	–	–	0.48 (0.21)	0.52 (0.25)
MGraphDTA	0.64 (0.94)	**0.62** (0.83)	**0.62** (0.82)	0.43 (0.55)	–	–	-0.06 (0.46)	-0.04 (0.50)	–	–	0.45 (0.23)	0.48 (0.22)
FDA	0.64 (0.94)	0.87 (1.22)	1.28 (1.80)	0.36 (0.45)	–	–	**0.20** (0.75)	**0.21** (0.70)	–	–	**0.60** (0.35)	**0.56** (0.33)
(b) Same-modification, different-ligands												
DeepDTA	0.56 (0.87)	0.63 (0.74)	0.62 (0.67)	0.54 (0.41)	0.78 (0.28)	0.76 (0.28)	0.76 (0.26)	0.78 (0.25)	0.82 (0.14)	0.80 (0.13)	0.80 (0.13)	0.81 (0.12)
AttentionDTA	0.56 (0.87)	0.67 (0.89)	0.62 (0.75)	**0.42** (0.45)	0.78 (0.28)	0.77 (0.27)	0.77 (0.26)	**0.80** (0.24)	0.82 (0.14)	**0.81** (0.13)	**0.81** (0.13)	**0.82** (0.12)
GraphDTA	0.56 (0.87)	1.10 (1.09)	1.45 (1.44)	1.20 (1.18)	0.78 (0.28)	0.75 (0.26)	0.66 (0.34)	0.70 (0.29)	0.82 (0.14)	0.79 (0.12)	0.76 (0.13)	0.78 (0.11)
DGraphDTA	0.56 (0.87)	0.63 (0.82)	0.64 (0.80)	0.50 (0.61)	0.78 (0.28)	**0.78** (0.28)	**0.78** (0.28)	0.78 (0.28)	0.82 (0.14)	**0.81** (0.13)	0.80 (0.13)	0.81 (0.13)
MGraphDTA	0.56 (0.87)	**0.59** (0.79)	**0.55** (0.67)	0.45 (0.39)	0.78 (0.28)	0.77 (0.29)	**0.78** (0.27)	**0.80** (0.24)	0.82 (0.14)	0.80 (0.14)	**0.81** (0.13)	**0.82** (0.12)
FDA	0.56 (0.87)	0.79 (1.08)	1.15 (1.09)	2.27 (1.62)	0.78 (0.28)	0.75 (0.27)	0.67 (0.21)	0.56 (0.31)	0.82 (0.14)	0.80 (0.12)	0.74 (0.10)	0.70 (0.12)

limitation is the truncation of dissociation constants (K_d): roughly 70% of K_d values are capped at $10\mu M$, obscuring weaker interactions and reducing data granularity. This censoring complicates the interpretation of modification-induced affinity changes, where ΔpK_d often represents only a lower bound or becomes entirely untrackable, thereby impairing predictive modeling. These limitations highlight the need for specialized algorithms and larger, more diverse datasets.

Benchmarking docking-based approaches such as Folding-Docking-Affinity (FDA) presents additional challenges. Although FDA demonstrates stronger zero-shot generalization than docking-free models, potential data leakage arises from overlaps between DAVIS proteins and the training data of AlphaFold-Multimer [11] and DiffDock [7], two componens of FDA, even if exact protein–ligand pairs were rarely shared [47]. These overlaps highlight the need for entirely new benchmark datasets that exclude previously seen proteins, ligands, and their combinations.

Furthermore, intuitively, one might expect the docking-based FDA model to outperform docking-free models completely, as it explicitly captures atom-level protein–ligand interactions, potentially reflecting the structural effects of protein modifications. Our structural investigations, however, suggest that the structure prediction models, including those for protein folding and molecular docking, are not yet fully capable of capturing the structural variations introduced by modifications. For example, we observed that AlphaFold3 [4] predicts a phosphorylated state for both the non-phosphorylated and phosphorylated forms of the ABL1 protein, failing to distinguish between the two. The result is consistent with a recent study [33]. This underscores that subtle structural changes from protein modifications are not yet adequately captured by existing models, limiting the effectiveness of downstream binding affinity predictions.

7 Conclusion

Protein modifications significantly impact protein–ligand interactions and binding affinity, yet experimentally homogeneous datasets incorporating these modifications remain scarce. We address this gap by curating a complete version of the DAVIS dataset with previously ignored modified kinase proteins. Using three benchmarks—Augmented Dataset Prediction, Wild-Type to Modification Generalization, and Few-Shot Modification Generalization—we evaluate state-of-the-art models' abilities to distinguish protein modifications. Results indicate docking-based models demonstrate superior generalization in zero-shot scenarios. Conversely, docking-free models frequently overfit to wild-type proteins, encountering difficulty with unseen modifications; however, their performance improves notably after fine-tuning on a limited number of modified examples. This curated dataset and benchmarks offer valuable resources to advance generalizable affinity prediction models and precision medicine.

8 Author contributions

MH.W, Z.X, and D.Z conceived the research project. Z.X, S.J, and D.Z supervised the research project. MH.W developed the computational method, implemented the software, and performed the evaluation analyses. All authors analyzed the results and participated in the interpretation. MH.W wrote the manuscript with support from all other authors.

9 Acknowledgements

This work was not supported by any funding. We thank the creators of the original DAVIS dataset for their work and making their data publicly available.

References

[1] Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1): D609–D617, 2025.

[2] Karim Abbasi, Parvin Razzaghi, Antti Poso, Massoud Amanlou, Jahan B Ghasemi, and Ali Masoudi-Nejad. Deepcda: deep cross-domain compound–protein affinity prediction through lstm and convolutional neural networks. *Bioinformatics*, 36(17):4633–4642, 2020.

[3] Robert Abel, Noeris K Salam, John Shelley, Ramy Farid, Richard A Friesner, and Woody Sherman. Contribution of explicit solvent effects to the binding affinity of small-molecule inhibitors in blood coagulation factor serine proteases. *ChemMedChem*, 6(6):1049–1066, 2011.

[4] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.

[5] Ammar Ammar, Rachel Cavill, Chris Evelo, and Egon Willighagen. Psnpbind: a database of mutated binding site protein–ligand complexes constructed using a multithreaded virtual screening workflow. *Journal of Cheminformatics*, 14(1):8, 2022.

[6] Lieyang Chen, Anthony Cruz, Steven Ramsey, Callum J Dickson, Jose S Duca, Viktor Hornak, David R Koes, and Tom Kurtzman. Hidden bias in the dud-e dataset leads to misleading performance of deep learning in structure-based virtual screening. *PloS one*, 14(8):e0220113, 2019.

[7] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2022.

[8] Gabriele Corso, Arthur Deng, Benjamin Fry, Nicholas Polizzi, Regina Barzilay, and Tommi Jaakkola. Deep confident steps to new pockets: Strategies for docking generalization. *arXiv preprint arXiv:2402.18396*, 2024.

[9] Gilda da Cunha Santos, Frances A Shepherd, and Ming Sound Tsao. Egfr mutations and lung cancer. *Annual Review of Pathology: Mechanisms of Disease*, 6(1):49–69, 2011.

[10] Mindy I. Davis, Jeremy P. Hunt, Sanna Herrgard, Pietro Ciceri, Lisa M. Wodicka, Gabriel Pallares, Michael Hocker, Daniel K. Treiber, and Patrick P. Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature Biotechnology*, 29:1046–1051, 11 2011. ISSN 10870156. doi: 10.1038/nbt.1990.

[11] Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with alphafold-multimer. *biorxiv*, pages 2021–10, 2021.

[12] Xiaoyi Hu, Jing Li, Maorong Fu, Xia Zhao, and Wei Wang. The jak/stat signaling pathway: from bench to clinic. *Signal transduction and targeted therapy*, 6(1):402, 2021.

[13] Kexin Huang, Cao Xiao, Lucas M Glass, and Jimeng Sun. Moltrans: molecular interaction transformer for drug–target interaction prediction. *Bioinformatics*, 37(6):830–836, 2021.

[14] Lei Huang, Jiecong Lin, Rui Liu, Zetian Zheng, Lingkuan Meng, Xingjian Chen, Xiangtao Li, and Ka-Chun Wong. Coadti: multi-modal co-attention based framework for drug–target interaction annotation. *Briefings in Bioinformatics*, 23(6):bbac446, 2022.

[15] Mingjian Jiang, Zhen Li, Shugang Zhang, Shuang Wang, Xiaofeng Wang, Qing Yuan, and Zhiqiang Wei. Drug–target affinity prediction using graph neural network and contact maps. *RSC advances*, 10(35):20701–20712, 2020.

[16] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A.A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589, 8 2021. ISSN 14764687. doi: 10.1038/s41586-021-03819-2.

[17] Masha Karelina, Joseph J Noh, and Ron O Dror. How accurately can one predict drug binding modes using alphafold models? *Elife*, 12:RP89386, 2023.

[18] Hamed Khakzad, Ilia Igashov, Arne Schneuing, Casper Goverde, Michael Bronstein, and Bruno Correia. A new age in protein design empowered by deep learning. *Cell Systems*, 14(11): 925–939, 2023.

[19] Tanja Kortemme. De novo protein design—from new structures to programmable functions. *Cell*, 187(3):526–544, 2024.

[20] Gregory A Landrum and Sereina Riniker. Combining ic50 or k i values from different sources is a source of significant noise. *Journal of chemical information and modeling*, 64(5):1560–1567, 2024.

[21] Ingoo Lee, Jongsoo Keum, and Hojung Nam. Deepconv-dti: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS computational biology*, 15 (6):e1007129, 2019.

[22] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[23] Tiqing Liu, Linda Hwang, Stephen K Burley, Carmen I Nitsche, Christopher Southan, W Patrick Walters, and Michael K Gilson. Bindingdb in 2024: a fair knowledgebase of protein-small molecule binding data. *Nucleic Acids Research*, 53(D1):D1633–D1644, 11 2024. ISSN 1362-4962. doi: 10.1093/nar/gkae1075. URL `https://doi.org/10.1093/nar/gkae1075`.

[24] Tiqing Liu, Linda Hwang, Stephen K Burley, Carmen I Nitsche, Christopher Southan, W Patrick Walters, and Michael K Gilson. Bindingdb in 2024: a fair knowledgebase of protein-small molecule binding data. *Nucleic acids research*, 53(D1):D1633–D1644, 2025.

[25] Yunan Luo, Yang Liu, and Jian Peng. Calibrated geometric deep learning improves kinase–drug binding predictions. *Nature Machine Intelligence*, 5(12):1390–1401, 2023.

[26] Charlotte M Miton and Nobuhiko Tokuriki. Insertions and deletions (indels): a missing piece of the protein engineering jigsaw. *Biochemistry*, 62(2):148–157, 2022.

[27] Thin Nguyen, Hang Le, Thomas P Quinn, Tri Nguyen, Thuc Duy Le, and Svetha Venkatesh. Graphdta: predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8):1140–1147, 2021.

[28] Tri Minh Nguyen, Thin Nguyen, Thao Minh Le, and Truyen Tran. Gefa: Early fusion approach in drug-target affinity prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19:718–728, 2022. ISSN 15579964. doi: 10.1109/TCBB.2021.3094217.

[29] Michael Offin, Hira Rizvi, Megan Tenet, Andy Ni, Francisco Sanchez-Vega, Bob T Li, Alexander Drilon, Mark G Kris, Charles M Rudin, Nikolaus Schultz, et al. Tumor mutation burden and efficacy of egfr-tyrosine kinase inhibitors in patients with egfr-mutant lung cancers. *Clinical Cancer Research*, 25(3):1063–1069, 2019.

[30] Alexey V Onufriev and Emil Alexov. Protonation and pk changes in protein–ligand binding. *Quarterly reviews of biophysics*, 46(2):181–209, 2013.

[31] Tatu Pantsar and Antti Poso. Binding affinity via docking: fact and fiction. *Molecules*, 23(8): 1899, 2018.

[32] Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, pages 2025–06, 2025.

[33] Pathmanaban Ramasamy, Jasper Zuallaert, Lennart Martens, and Wim F Vranken. Assessing the relation between protein phosphorylation, alphafold3 models and conformational variability. *bioRxiv*, pages 2025–04, 2025.

[34] Ahmet Süreyya Rifaioglu, Rengül Cetin Atalay, D Cansen Kahraman, Tunca Doğan, Maria Martin, and Volkan Atalay. Mdeepred: novel multi-channel protein featurization for deep learning-based binding affinity prediction in drug discovery. *Bioinformatics*, 37(5):693–704, 2021.

[35] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pages 9323–9332. PMLR, 2021.

[36] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30, 2017.

[37] Danislav S Spassov. Binding affinity determination in drug design: insights from lock and key, induced fit, conformational selection, and inhibitor trapping models. *International Journal of Molecular Sciences*, 25(13):7124, 2024.

[38] Romain A Studer, Benoit H Dessailly, and Christine A Orengo. Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. *Biochemical journal*, 449(3):581–594, 2013.

[39] Jing Tang, Agnieszka Szwajda, Sushil Shakyawar, Tao Xu, Petteri Hintsanen, Krister Wennerberg, and Tero Aittokallio. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *Journal of Chemical Information and Modeling*, 54(3): 735–743, 2014.

[40] ByteDance AML AI4Science Team, Xinshi Chen, Yuxuan Zhang, Chan Lu, Wenzhi Ma, Jiaqi Guan, Chengyue Gong, Jincai Yang, Hanyu Zhang, Ke Zhang, et al. Protenix-advancing structure prediction through a comprehensive alphafold3 reproduction. *BioRxiv*, pages 2025–01, 2025.

[41] Chai Discovery team, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhonikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life. *BioRxiv*, pages 2024–10, 2024.

[42] Mikhail Volkov, Joseph-André Turk, Nicolas Drizard, Nicolas Martin, Brice Hoffmann, Yann Gaston-Mathé, and Didier Rognan. On the frustration to predict binding affinities from protein–ligand structures with deep neural networks. *Journal of medicinal chemistry*, 65(11):7946–7958, 2022.

[43] Izhar Wallach and Abraham Heifets. Most ligand-based classification benchmarks reward memorization rather than generalization. *Journal of chemical information and modeling*, 58(5): 916–932, 2018.

[44] Kaili Wang, Renyi Zhou, Yaohang Li, and Min Li. Deepdtaf: a deep learning method to predict protein–ligand binding affinity. *Briefings in Bioinformatics*, 22(5):bbab072, 2021.

[45] Lisa M Wodicka, Pietro Ciceri, Mindy I Davis, Jeremy P Hunt, Mark Floyd, Sara Salerno, Xuequn H Hua, Julia M Ford, Robert C Armstrong, Patrick P Zarrinkar, et al. Activation state-dependent binding of small molecule kinase inhibitors: structural insights from biochemistry. *Chemistry & biology*, 17(11):1241–1249, 2010.

[46] Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, et al. Boltz-1: Democratizing biomolecular interaction modeling. *bioRxiv*, pages 2024–11, 2024.

[47] Ming-Hsiu Wu, Ziqian Xie, and Degui Zhi. A folding-docking-affinity framework for protein-ligand binding affinity prediction. *Communications Chemistry*, 8(1):1–9, 2025.

[48] Fuxiao Xin and Predrag Radivojac. Post-translational modifications induce significant yet not extreme changes to protein structure. *Bioinformatics*, 28(22):2905–2913, 2012.

[49] Jincai Yang, Cheng Shen, and Niu Huang. Predicting or pretending: artificial intelligence for protein-ligand interactions lack of sufficiently large and unbiased datasets. *Frontiers in pharmacology*, 11:508760, 2020.

[50] Ziduo Yang, Weihe Zhong, Lu Zhao, and Calvin Yu-Chian Chen. Mgraphdta: deep multiscale graph neural network for explainable drug–target binding affinity prediction. *Chemical science*, 13(3):816–833, 2022.

[51] Ziduo Yang, Weihe Zhong, Qiujie Lv, Tiejun Dong, and Calvin Yu-Chian Chen. Geometric interaction graph neural network for predicting protein-ligand binding affinities from 3d structures (gign). *The Journal of Physical Chemistry Letters*, 14(8):2020–2033, 2023.

[52] Weining Yuan, Guanxing Chen, and Calvin Yu-Chian Chen. Fusiondta: attention-based feature polymerizer and knowledge distillation for drug-target binding affinity prediction. *Briefings in Bioinformatics*, 23(1):bbab506, 2022.

[53] Yuni Zeng, Xiangru Chen, Yujie Luo, Xuedong Li, and Dezhong Peng. Deep drug-target binding affinity prediction with multiple attention blocks. *Briefings in bioinformatics*, 22(5): bbab117, 2021.

[54] Qichang Zhao, Guihua Duan, Mengyun Yang, Zhongjian Cheng, Yaohang Li, and Jianxin Wang. Attentiondta: Drug–target binding affinity prediction by sequence-based deep learning with attention mechanism. *IEEE/ACM transactions on computational biology and bioinformatics*, 20(2):852–863, 2022.

[55] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: Deep drug-target binding affinity prediction. *Bioinformatics*, 34:i821–i829, 9 2018. ISSN 14602059. doi: 10.1093/bioinformatics/bty593.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims mentioned in abstract and introduction are fully supported by the results and discussion in section 5.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please check Section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Please check the Model Training Details section in the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code and data are available at `https://github.com/ZhiGroup/DAVIS-complete`

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Please check the Model Training Details section in the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The all results are accompanied by error bars and please check section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please check `https://github.com/ZhiGroup/DAVIS-complete` for computer resource information.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Please check `https://github.com/ZhiGroup/DAVIS-complete`

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: On the positive side, we emphasizes advancing precision medicine by improving binding affinity prediction for modified proteins, which supports personalized drug discovery (Section 1, Abstract). On the negative side, we highlights dataset limitations—such as kinase bias and truncated binding affinity values—that may limit generalizability and model reliability (Section 6)

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The DAVIS dataset and the benchmark models are appropriately cited in the paper. The license of DAVIS dataset is CC-BY 4.0. For the version and the license of benchmark models, please check the Benchmark Models section in the supplementary information.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The introduced DAVIS-complete dataset is well documented and please check the document in the Details of Protein Modification section in the supplementary material. The benchmark code is documented in `https://github.com/ZhiGroup/DAVIS-complete`.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLM is used solely for rephrasing and language refinement.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.