A LARGE-SCALE DATASET AND BENCHMARK: DO PROTEIN-LIGAND MODELS LEARN BINDING SITES OR JUST BINDING LIKELIHOOD?

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein-ligand modeling underpins computational drug discovery and molecular design. Existing protein-ligand benchmarks typically evaluate whether a protein and ligand interact and how strongly they bind, through tasks such as binary binding prediction and affinity regression. However, these evaluations provide limited evidence of whether models can localize binding sites or identify the non-covalent interactions underlying molecular recognition. To address this gap, we introduce **InteractBind**, a large-scale protein-ligand dataset comprising approximately 100k protein-ligand pairs, together with a benchmark for fine-grained evaluation. The core fine-grained task is that of binding-site localization, which uses protein-residue and ligand-atom interaction maps spanning six major types of non-covalent interactions to assess whether model-derived interaction maps localize binding sites. InteractBind further includes protein similarity-controlled splits to support realistic generalization assessment. Using InteractBind, we evaluate eight existing sequence-based and interaction-aware models, assessing binary binding prediction and binding-site localization. Results reveal limited binding-site localization despite strong binary binding prediction, with marked variation across non-covalent interaction types. Overall, InteractBind establishes a benchmark paradigm that encourages the development of more interpretable and physically grounded protein-ligand models.

1 INTRODUCTION

Protein-ligand interactions underlie a wide range of biological and chemical processes, including enzymatic catalysis, molecular recognition, cellular signaling, and therapeutic modulation (Barabasi & Oltvai, 2004; Tan et al., 2024; Schwartz & Schramm, 2009). Experimentally characterizing these interactions at scale remains costly and time-consuming, which makes protein-ligand modeling a central component of computational drug discovery: virtual screening prioritizes candidate binders from large compound libraries to support lead optimization before wet-lab validation (Finan et al., 2017; Dara et al., 2022; Wohlwend et al., 2025). Representative advances include structure-based approaches that model three-dimensional binding geometry, such as docking, molecular dynamics, and geometric deep learning (Adams et al., 2025; Morehead et al., 2025; Miñán et al., 2025), as well as sequence-based approaches that predict binding from protein sequences and ligand string or graph representations (Chen et al., 2025; Zhang et al., 2025; Jia et al., 2026). To evaluate these approaches, existing benchmarks for protein-ligand modeling focus primarily on ranking: given a protein and a panel of candidate ligands, a model must place true binders above non-binders or order compounds by binding strength, whether the supervision comes from binary interaction labels (Gilson et al., 2016; Zitnik et al., 2018) or continuous affinity measurements (Liu et al., 2015; Davis et al., 2011). However, the ranking metrics measure only whether a model separates binders from non-binders, rather than identifying specific binding sites that explain molecular recognition (Mattos & Ringe, 1996; Černỳ & Hobza, 2007). Consequently, evidence that existing models capture such fine-grained interaction patterns often comes from a few visualized case studies, motivating a new benchmark that evaluates these capabilities systematically against large-scale ground-truth annotations.

Beyond ranking, a fine-grained evaluation of protein-ligand models should test whether models locate the binding sites that underlie molecular recognition. In experimentally resolved protein-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ligand complex structures, binding is mediated by non-covalent contacts between ligand atoms and protein residues (Sec. A) (Finan et al., 2017; Peluso & Chankvetadze, 2022). These contacts naturally define binding sites: a residue is considered a binding-site residue if it participates in one or more ligand-stabilizing contacts (Mattos & Ringe, 1996; Burley et al., 2017). Thus, binding-site localization evaluates whether model-derived interaction maps identify where the ligand binds. However, evaluating only the overall binding-site map cannot reveal which physicochemical contact patterns are captured, because binding-site residues arise from interaction types with different geometric and chemical constraints. Retaining interaction-type annotations further enables the evaluation of whether models actually identify the non-covalent interactions that stabilize residue-atom contacts. Such information supports targetable residue identification, selectivity analysis, and structure-informed lead optimization (Crunkhorn, 2024; Holehouse & Kragelund, 2024). Existing sequence-based protein-ligand binding benchmarks rarely provide binding-site and interaction-type-specific annotations at scale, limiting systematic evaluation of whether models localize binding sites and identify the non-covalent interaction patterns underlying ligand recognition.

To address this limitation, we introduce **InteractBind**, a new large-scale protein-ligand dataset derived from the Protein Data Bank (PDB) (Berman et al., 2007), and establish the corresponding **InteractBind benchmark** for the fine-grained evaluation of protein-ligand modeling. The dataset comprises approximately 100k protein-ligand pairs, spanning around 11k proteins and 9k ligands. Existing benchmarks typically provide only binary labels or affinity scores: for example, BioSNAP (Zitnik et al., 2018) provides binary binding labels, while KIBA (Tang et al., 2014) provides scalar binding affinity values. In contrast, as shown in Fig. 1, InteractBind also provides sequence-level interaction maps for six major non-covalent interactions, together with an overall interaction map that defines protein binding sites. The new benchmark centers on binding-site localization as the primary fine-grained evaluation task, assessing whether model-derived interaction maps can localize the protein residues involved in ligand binding. The interaction-type-specific maps further support evaluation of whether models identify the non-covalent interactions that stabilize residue-atom contacts.

To improve benchmark realism and generalization assessment, InteractBind provides two types of dataset splits: an in-distribution (ID) split and protein similarity-controlled out-of-distribution (OOD) splits for assessing generalization to less familiar protein targets. In the ID split, we benchmark eight existing representative interaction-aware models across binary binding prediction, binding-site localization, and interaction-type-specific evaluation. The best-performing model achieves 85.2% AUROC for binary binding prediction but only 2.3% Recall@5 for binding-site localization. Interaction-type-specific evaluation further reveals uneven performance across non-covalent interaction categories, indicating that accurate binding prediction does not necessarily imply reliable identification of the physicochemical contacts that stabilize binding. For OOD evaluation, we construct four protein similarity-controlled dataset splits using global sequence alignment (Hamamsy et al., 2024), with maximum train-test protein similarities of 25%, 28%, 31%, and 33%. Performance decreases as the train–test protein similarity constraint becomes more stringent, highlighting the difficulty of localizing binding sites for less familiar protein targets. To our knowledge, no prior sequence-based benchmark provides explicit residue-level interaction supervision together with a systematic evaluation of binding-site localization derived from non-covalent contacts. Our contributions are summarized as follows:

- ❶ We introduce **InteractBind**, a new large-scale protein-ligand dataset comprising approximately 100k pairs with binding-site annotations and non-covalent interaction maps.
- ❷ We formulate binding-site localization as the core fine-grained evaluation task, supported by metrics for residue-level localization and interaction-type-specific contact analysis, and construct both ID and OOD splits for realistic generalization.
- ❸ We benchmark eight representative protein-ligand models and show that interaction-aware architectures provide measurable binding-site localization signals, while highlighting substantial room for improving localization and non-covalent interaction identification.

1.1 Protein-ligand modeling

Protein-ligand modeling has been studied from both structure-based and sequence-based perspectives. Structure-based methods (Adams et al., 2025; Morehead et al., 2025; Miñán et al., 2025) model binding from 3D structures, including molecular dynamics, geometric deep learning, and

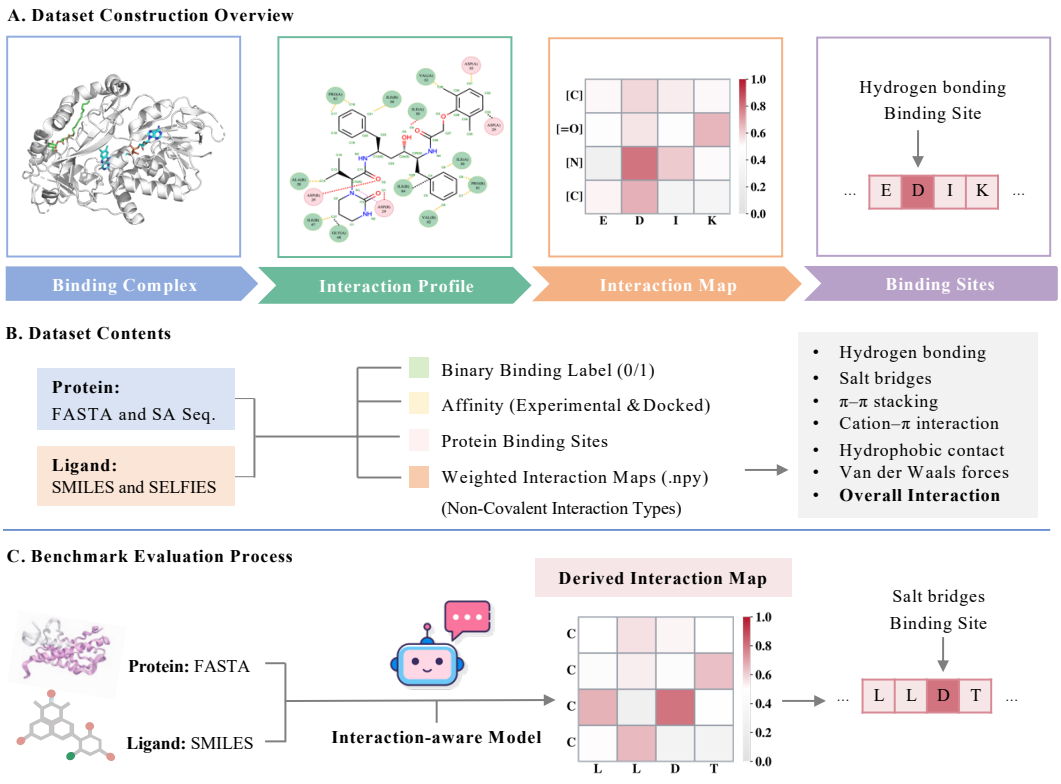

Figure 1: Overview of InteractBind. A. InteractBind is constructed from experimentally resolved protein-ligand complexes by deriving interaction profiles, sequence-level interaction maps, and protein binding-site annotations. **B.** Each sample includes protein and ligand sequence representations, binary binding labels, affinity values, binding-site annotations, and weighted interaction maps for six non-covalent interactions together with an overall interaction map. **C.** Given protein and ligand sequence representations, model-derived interaction maps are evaluated against ground-truth protein binding-site annotations.

recent biomolecular structure prediction models such as AF3 (Abramson et al., 2024) and Boltz-1 (Wohlwend et al., 2025). However, these methods depend on high-quality structures or reliable predicted poses, and geometric accuracy does not necessarily ensure chemically faithful non-covalent interactions (Abramson et al., 2024; Masters et al., 2025). Sequence-based models provide a scalable alternative when complex structures are unavailable, predicting binding from protein sequences and ligand string or graph representations (Huang et al., 2021; Chen et al., 2020). Recent interaction-aware models further use attention or fusion modules to model residue-atom correspondences, with qualitative case studies suggesting that learned interaction maps can highlight plausible binding-site residues (Bai et al., 2023; Meng et al., 2025). However, because existing datasets rarely provide residue-atom interaction annotations, it remains difficult to determine whether these maps systematically localize physically meaningful binding sites or merely correlate with binary binding prediction. Our new InteractBind benchmark is motivated by this gap, providing interaction annotations for evaluating whether sequence-based models localize binding residues and interaction patterns.

1.2 Protein-Ligand Datasets

Protein-ligand datasets serve different purposes depending on the level of information they provide. Structural resources such as the PDB and AlphaFoldDB provide experimentally resolved or predicted protein structures, supporting structural analysis, binding-pose modeling, and structure-based hypothesis generation (Berman et al., 2007; Varadi et al., 2022). Docking-oriented datasets, such as DUD-E, are widely used to evaluate virtual screening methods by providing active compounds and property-matched decoys for target proteins (Mysinger et al., 2012). Pair-level learning datasets support complementary predictive tasks: BindingDB, BioSNAP, and Human are commonly used for binary drug-target interaction prediction (Gilson et al., 2016; Zitnik et al., 2018; Liu et al., 2015; Chen et al., 2020), while Davis and KIBA are widely used for binding affinity prediction (Davis

Table 1: Comparison of sequence-based protein-ligand datasets. "✓" indicates that the information is provided or directly supported by the dataset, whereas "-" indicates that it is not provided. InteractBind provides both available experimental affinity values extracted from PDB records and docking-derived affinity estimates.

Dataset	Binary label	Affinity	Binding sites	Interaction-type
DUD-E (Mysinger et al., 2012)	✓	-	-	-
BindingDB (Gilson et al., 2016)	✓	✓	-	-
BioSNAP (Zitnik et al., 2018)	✓	-	-	-
Human (Liu et al., 2015; Chen et al., 2020)	✓	-	-	-
Davis (Davis et al., 2011)	-	✓	-	-
KIBA (Tang et al., 2014)	-	✓	-	-
InteractBind	✓	✓	✓	✓

et al., 2011; Tang et al., 2014). Together, these resources have driven progress in structure-based analysis, virtual screening, interaction prediction, and affinity estimation. However, they generally do not provide large-scale residue-atom annotations specifying which protein residues contact the ligand and which non-covalent interactions stabilize these contacts. InteractBind fills this gap by extracting non-covalent interactions from experimentally resolved protein-ligand complexes and converting them into sequence-level residue-atom annotations and aggregated binding-site maps, enabling the systematic training and evaluation of binding-site localization beyond pair-level labels.

2 The InteractBind Benchmark

Fig. 1 provides an overview of InteractBind, including dataset construction from experimentally resolved protein-ligand complexes and the resulting annotations for each protein-ligand pair. These annotations enable the evaluation of whether model-derived interaction maps localize the protein residues involved in binding. Tab. 1 further contrasts InteractBind with existing representative protein-ligand datasets, highlighting its support for a broader range of protein-ligand modeling tasks. In the following subsections, we describe the dataset curation procedure (Sec. 2.1), summarize the dataset statistics (Sec. 2.2), and define the tasks and evaluation protocol used in this study (Sec. 3).

2.1 Dataset curation

InteractBind is curated from experimentally resolved protein-ligand complexes in the Protein Data Bank (PDB), which provide atomic coordinates for both proteins and bound ligands (Berman et al., 2007). For each protein, we construct both the canonical FASTA sequence and the SaProt structure-aware (SA) sequence (Su et al., 2024); for each ligand, we convert PDB coordinates to SMILES and SELFIES, using SELFIES as a robust ligand sequence representation for sequence-based interaction map modeling (Krenn et al., 2022). Non-covalent residue-atom interactions are detected from each three-dimensional complex using PLIP (Schake et al., 2025) and GetContacts (GetContacts developers, 2018), and projected onto the corresponding protein residue and ligand atom to construct sequence-level interaction maps. Therefore, the resulting sample contains paired protein and ligand sequence representations together with sequence-aligned interaction annotations, as illustrated in Fig. 1 A-B. Details of structure collection, filtering, representation conversion, and sample construction are provided in Appendix B.1.

The interaction annotations are constructed from six major non-covalent interaction types: hydrogen bonds, salt bridges, van der Waals interactions, hydrophobic contacts, π-π stacking, and cation-π interactions. These interactions represent major physicochemical forces underlying protein-ligand recognition and provide an interpretable basis for describing residue-atom contacts (Černý & Hobza, 2007; Holehouse & Kragelund, 2024). Interaction validity is determined using rule-based geometric and chemical constraints, including distance cutoffs and, where applicable, angular conditions. The six interaction-specific maps are also aggregated into an overall interaction map, from which protein binding-site annotations are derived: a residue is labeled as part of the binding site if it participates in at least one detected ligand-stabilizing non-covalent interaction. The full interaction annotation protocol and interaction-specific cutoffs are provided in Appendix B.2.

2.2 DATASET STATISTICS

InteractBind contains 99,391 protein-ligand pairs, covering 11,473 unique proteins and 9,017 unique ligands. Fig. 2 summarizes the interaction-level and binding-site statistics of InteractBind. Panel a shows that the dataset is dominated by common protein-ligand contact patterns, including van der Waals interactions, hydrogen bonds and hydrophobic contacts, while also retaining less frequent but chemically informative interaction types such as salt bridges, π-π stacking and cation-π interactions. This distribution reflects the natural imbalance of non-covalent contacts in protein-ligand complexes and motivates interaction-type-specific evaluation rather than treating all contacts as a single homogeneous signal. Panel b further shows that annotated binding sites vary substantially in size, with most complexes containing a moderate number of binding-site residues and a long tail of larger sites. This diversity indicates that InteractBind covers binding interfaces of different extents, providing a basis for evaluating whether models can localize binding residues beyond simple or narrowly defined pocket patterns. The detailed statistics for these datasets are provided in Tab. 3.

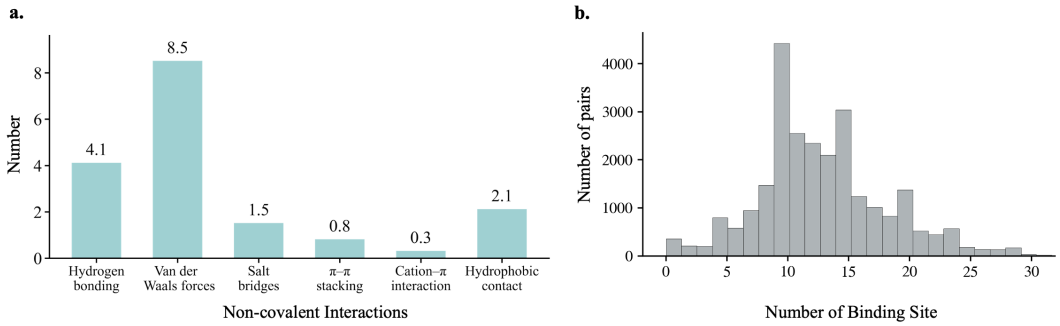

Figure 2: Descriptive statistics of InteractBind. a. Average numbers of the six annotated non-covalent interaction types across InteractBind. **b.** Distribution of annotated binding-site residues across InteractBind.

3 TASKS AND EVALUATION PROTOCOL

Dataset splits. InteractBind supports both in-distribution and out-of-distribution evaluation. For the in-distribution setting, the ID split treats curated experimentally resolved protein-ligand pairs from the PDB as positive samples and constructs negative pairs at a positive-to-negative ratio of 1:10. Following the negative-sampling protocol commonly used in existing protein-ligand datasets (Zitnik et al., 2018; Liu et al., 2015), negatives are generated by randomly re-pairing proteins and ligands from the positive set. Specifically, for each positive pair, either the protein or the ligand is replaced by another partner sampled from the corresponding empirical frequency distribution, alternating between the two sides. This design helps reduce simple frequency-based shortcuts by approximately preserving the protein and ligand marginal distributions of the positive set. The resulting dataset is evaluated using stratified 5-fold cross-validation, preserving the 1:10 class ratio in each fold, and performance is reported as the mean $\pm$ standard deviation across folds.

For out-of-distribution evaluation, we construct protein similarity-controlled splits because this setting more closely reflects the practical scenario of predicting interactions for less familiar protein targets (Jia et al., 2026). In sequence-based protein-ligand modeling, the difficulty of generalization is strongly affected by the similarity between training and test sequences, whereas cold-start partitioning based only on identifiers or names does not necessarily define a more challenging or realistic setting (Chatterjee et al., 2023; Jia et al., 2026). Protein similarity is measured using global sequence alignment based on the Needleman-Wunsch algorithm (Hamamsy et al., 2024). This approach yields four protein-based OOD datasets with maximum train-test protein similarities of 25%, 28%, 31%, and 33%. Each dataset is divided into training, validation, and test subsets with a ratio of 8:1:1. The reported OOD results are obtained over 5 runs with different random seeds and are presented as mean $\pm$ standard deviation. These splits provide a controlled setting for evaluating binary binding prediction and binding-site localization under target-side distribution shift.

Baseline models. We evaluate eight existing representative sequence-based protein-ligand models whose interaction modules have been qualitatively used to support claims of binding-site localization.

Because these models expose or permit the derivation of interaction-aware signals in addition to binary binding predictions, InteractBind enables us to evaluate their residue-level localization systematically. The evaluated models include transformer- or cross-attention-based architectures, including MolTrans (Huang et al., 2021), TransformerCPI (Chen et al., 2020), HyperAttentionDTI (Zhao et al., 2022), PerceiverCPI (Nguyen et al., 2023), CAT-DTI (Zeng et al., 2024), and FusionDTI (Meng et al., 2025), and bilinear-attention-network-based architectures, including DrugBAN (Bai et al., 2023) and GraphBAN (Hadipour et al., 2025). Detailed baseline descriptions and interaction-map extraction details are provided in Appendix C.

Binary binding prediction evaluation. For binary binding prediction, we follow standard metrics used in protein-ligand interaction modeling, including AUROC, AUPRC, Matthews correlation coefficient (MCC), and F1 score (Hadipour et al., 2025; Meng et al., 2025). These metrics quantify whether a model predicts binding correctly, but they do not indicate whether the model localizes the protein residues involved in ligand recognition.

Binding-site localization evaluation. Binding-site localization is evaluated only on positive protein-ligand pairs with annotated binding residues. Precision@K measures the fraction of the Top-K predicted residues that are true binding residues, while Recall@K measures the fraction of annotated binding residues recovered within the Top-K predictions. For each protein-ligand pair, Average Precision (AP) summarizes the complete residue ranking by averaging precision at ranks corresponding to true binding residues, and Mean Average Precision (MAP) is obtained by averaging AP across all evaluated pairs. We additionally report Binding Residue Hit Rate (BRHR@K), which measures the fraction of protein-ligand pairs for which at least one of the Top-K predicted residues overlaps with an annotated binding residue; its formal definition is provided in Appendix D. Together, these metrics capture complementary aspects of binding-site localization, including top-ranked precision, binding-site coverage, overall ranking quality, and successful site-level localization.

4 Results

4.1 From binary binding prediction to binding-site localization

A central goal of InteractBind is to evaluate whether protein-ligand models can move beyond binary binding prediction and localize the binding sites that support ligand recognition. We therefore examine eight representative interaction-aware baselines on both binary binding prediction and binding-site localization under the in-distribution split. As shown in Tab. 2, models with stronger binary binding prediction generally also achieve stronger binding-site localization, suggesting a positive association between the two tasks on the InteractBind benchmark. FusionDTI performs best across both tasks, achieving 85.2% AUROC and 28.3% MCC for binary binding prediction, together with 19.4% BRHR@1, 7.5% Precision@5, 2.3% Recall@5, and 11.1% MAP for binding-site localization. This is consistent with FusionDTI's token-level fusion design, which explicitly models residue-atom correspondences. Bilinear-attention models also show competitive localization performance, with GraphBAN reaching 16.3% BRHR@1 and DrugBAN reaching 7.1% Precision@5. These results indicate that learned protein-ligand interaction maps contain measurable binding-site signals, especially when models explicitly learn residue-atom interactions. Nevertheless, binding-site localization remains substantially underdeveloped across the existing baselines. Indeed, even FusionDTI reaches only 19.4% BRHR@1 despite achieving 85.2% AUROC, showing that strong binary binding prediction does not yet translate into reliable top-ranked binding-site localization. This limitation is expected because these baselines are trained primarily with binary binding labels, without large-scale binding-site annotations to guide their interaction maps. InteractBind therefore provides the supervision and evaluation setting needed to assess whether models localize binding sites, rather than only whether they predict binding correctly.

4.2 Effect of binding-site size on localization

We further investigate how binding-site size affects residue-level localization performance by grouping experimentally validated protein-ligand pairs according to the number of annotated binding residues: 1-5, 6-10, 11-20, and > 20. As shown in Fig. 3, MAP generally increases with binding-site size across all models. FusionDTI consistently achieves the best performance, improving from 5.4% MAP for sites with 1-5 residues to 10.8%, 21.9%, and 22.8% for the 6-10, 11-20, and > 20 groups, respectively.

Table 2: Performance (%) of binary binding prediction and binding-site localization under the ID setting. The best results are shown in **bold**, and the second-best results are underlined.

Binary binding prediction				
Model	**AUROC**	**AUPRC**	**MCC**	**F1**
MolTrans	79.9 ± 1.4	35.5 ± 1.3	23.5 ± 0.8	19.4 ± 0.8
TransformerCPI	78.7 ± 1.1	35.0 ± 1.2	23.9 ± 0.8	18.5 ± 0.7
HyperAttentionDTI	81.3 ± 1.1	37.7 ± 1.4	22.3 ± 0.7	19.8 ± 0.6
PerceiverCPI	80.2 ± 0.9	39.6 ± 0.9	24.2 ± 0.5	20.5 ± 0.7
CAT-DTI	81.8 ± 1.9	39.7 ± 0.9	25.3 ± 1.1	21.1 ± 1.3
DrugBAN	83.6 ± 0.9	40.3 ± 0.7	27.1 ± 0.5	21.8 ± 1.0
GraphBAN	83.9 ± 1.1	40.5 ± 0.9	26.8 ± 0.6	22.2 ± 0.7
FusionDTI	**85.2 ± 1.2**	**42.4 ± 0.8**	**28.3 ± 0.6**	**24.6 ± 0.9**
Random	50.0	9.1	0.0	9.1
Binding-site localization				
Model	**BRHR@1**	**Precision@5**	**Recall@5**	**MAP**
MolTrans	10.5 ± 0.5	4.6 ± 0.5	1.4 ± 0.1	9.7 ± 0.4
TransformerCPI	9.7 ± 0.6	4.8 ± 0.2	1.3 ± 0.3	9.0 ± 0.2
HyperAttentionDTI	11.7 ± 0.5	4.6 ± 0.6	1.9 ± 0.2	9.6 ± 0.4
PerceiverCPI	9.6 ± 0.3	4.0 ± 0.3	1.5 ± 0.2	9.2 ± 0.3
CAT-DTI	12.6 ± 0.7	4.8 ± 0.3	1.3 ± 0.2	8.5 ± 0.4
DrugBAN	15.8 ± 0.9	7.1 ± 0.6	2.0 ± 0.4	10.3 ± 0.7
GraphBAN	16.3 ± 0.6	6.7 ± 0.4	1.8 ± 0.2	9.8 ± 0.3
FusionDTI	**19.4 ± 0.4**	**7.5 ± 0.6**	**2.3 ± 0.5**	**11.1 ± 0.5**
Random	1.5	1.5	1.0	4.7

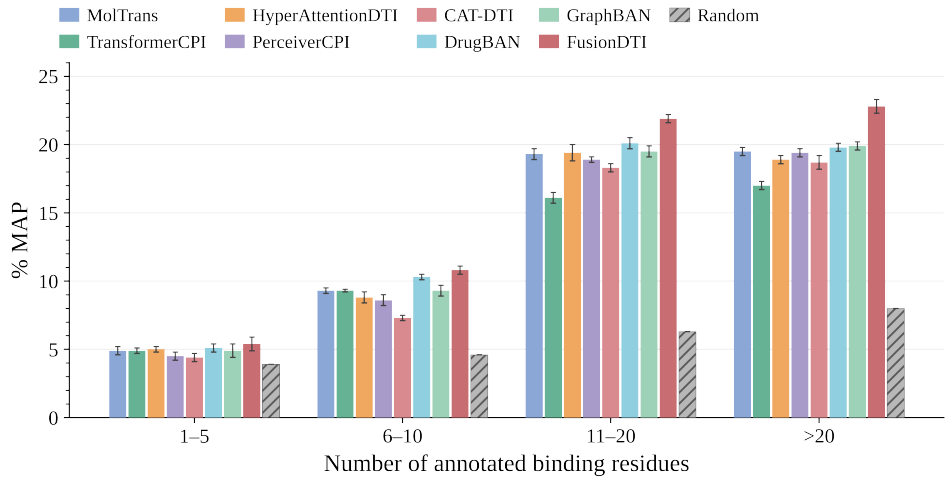

Figure 3: Binding-site localization performance across different binding-site sizes under the ID setting.

DrugBAN achieves the second-best performance in the first three groups, while GraphBAN is second-best for sites with more than 20 residues. Notably, the random baseline also increases from 3.9% to 8.0% as site size grows, indicating that part of the apparent improvement for larger sites arises from the increased chance of ranking a true binding residue highly. Nevertheless, the gap between learned models and the random baseline becomes substantially larger for medium and large binding sites. In contrast, performance on the 1-5 group remains relatively close to random, suggesting that small binding sites are particularly challenging to localize precisely.

4.3 Per-interaction-type analysis of binding-site localization

A distinctive feature of InteractBind is that binding-site annotations are decomposed into six non-covalent interaction types, enabling a fine-grained evaluation beyond an aggregate binding-site map.



Figure 4: Interaction-type-specific binding-site localization in the in-distribution setting. Heatmap of BRHR@5 (%) across eight representative models and six non-covalent interaction types.

We therefore evaluate binding-site localization separately for each interaction category. As shown in Fig. 4, BRHR@5 varies substantially across interaction types under exact-hit matching. Van der Waals interactions are the easiest to localize, with FusionDTI reaching 33.2% BRHR@5, followed by GraphBAN at 30.7% and DrugBAN at 29.8%. This trend is consistent with the relatively common and spatially broad nature of van der Waals contacts, which provide more diffuse residue-level localization signals than highly specific interaction modes. Hydrophobic contacts and hydrogen bonds also show relatively strong localization performance, with FusionDTI achieving 28.2% and 28.9% BRHR@5, respectively. Salt bridges follow a similar pattern, with FusionDTI reaching 27.9% and GraphBAN 26.1%.

By contrast, π-π stacking and cation-π interactions remain more challenging: the best BRHR@5 values are 18.2% for π-π stacking and 16.1% for cation-π, both achieved by FusionDTI. This is expected because aromatic and electrostatic interactions require more specific geometric arrangements and residue-atom alignment, making exact-hit localization more difficult. These results indicate that binding-site localization is not uniform across physicochemical regimes. Common and spatially broader contacts, such as van der Waals, hydrophobic, and hydrogen-bond interactions, are more readily localized, whereas aromatic, electrostatic, and geometry-specific interactions remain harder to recover. Although models with stronger overall binding-site localization also tend to perform better across interaction types, the low scores for π-π stacking and cation-π show that fine-grained non-covalent interaction patterns remain difficult to identify reliably. Thus, the interaction-type-specific annotations in InteractBind provide a diagnostic benchmark signal that aggregate binding-site localization alone cannot capture, revealing which physicochemical contact patterns are localized well and which remain poorly resolved.

4.4 Generalization towards novel proteins

A key goal of InteractBind is to support realistic generalization assessment beyond random or in-distribution evaluation. To this end, we construct four protein similarity-controlled datasets, enabling controlled evaluation of model inference on novel protein targets. As shown in Fig. 5 (left), binary binding prediction improves as the protein similarity constraint becomes less stringent. FusionDTI increases from 81.8% AUROC at 25% similarity to 91.8% at 33%, while GraphBAN improves from 81.7% to 90.8% and DrugBAN from 79.1% to 89.0%. A similar trend is observed for other baselines. This trend is expected for sequence-based models, as test proteins that are more similar to training proteins provide more familiar sequence patterns and reduce target-side distribution shift. InteractBind makes this effect explicit by quantifying performance across controlled protein similarity thresholds. Binding-site localization shows a similar but more moderate trend. As shown in Fig. 5 (right), FusionDTI 's performance increases from 28.4% BRHR@5 at 25% similarity to 30.1% at 33%, while it ranges from 27.8% to 29.1% for DrugBAN. In contrast, other baselines generally

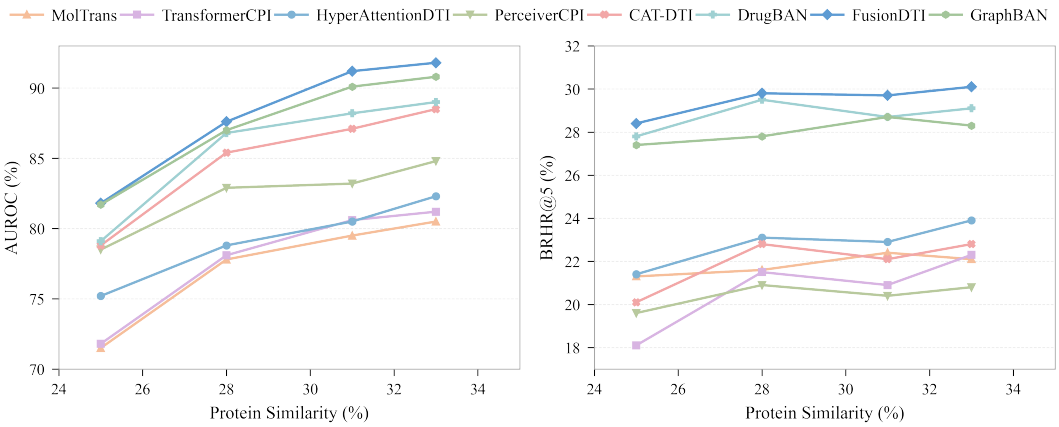

Figure 5: Generalisation across protein similarity-controlled OOD datasets. The four splits constrain maximum train-test protein sequence similarity to 25%, 28%, 31%, and 33%. **Left:** binary binding prediction performance measured by AUROC. **Right:** binding-site localisation performance measured by BRHR@5.

remain below 24% BRHR@5 across the four datasets. These findings indicate that train-test protein similarity affects not only binary binding prediction but also binding-site localization. Thus, the protein similarity-controlled splits in InteractBind provide a more informative stress test for evaluating whether models can localize binding sites on less familiar protein targets. We also construct ligand similarity-controlled splits and observe a smaller effect on model performance, with detailed results provided in Appendix E.

5 Discussion

The results on InteractBind reveal a **clear gap between binary binding prediction and residue-level localization**. Current interaction-aware models exhibit measurable binding-site localization ability, suggesting that their learned protein-ligand interaction maps capture useful residue-level signals, and models with stronger binary binding prediction generally also perform better in localization. However, even the strongest baseline, FusionDTI, achieves 85.2% AUROC for binary binding prediction but only 19.4% BRHR@1, 2.3% Recall@5, and 11.1% MAP for binding-site localization. This indicates that accurately predicting whether binding occurs does not necessarily imply reliable identification of the residues responsible for binding, highlighting the need for large-scale residue-level supervision. The size of the binding site further affects the difficulty of localization: FusionDTI improves from 5.4% MAP for sites containing 1-5 residues to 22.8% for sites containing more than 20 residues. Although the random baseline also increases with site size, learned models show a substantially larger advantage for medium and large sites, whereas performance on small sites remains relatively close to random. These results highlight that **small and sparse binding sites remain particularly challenging to localize**.

The fine-grained analysis further indicates that **localization difficulty depends strongly on physicochemical interaction type**. Common and spatially broader contacts, such as van der Waals, hydrophobic, and hydrogen-bond interactions, are generally easier to localize, whereas π-π stacking and cation-π interactions remain more challenging, indicating that aggregate localization metrics can obscure important differences across interaction patterns. Furthermore, protein similarity-controlled OOD evaluation shows that train-test similarity affects both binary binding prediction and binding-site localization: performance improves as the similarity constraint becomes less stringent, with a stronger effect on binary prediction and a more moderate effect on localization. This highlights the importance of **similarity-controlled evaluation for assessing generalization to less familiar protein targets**. Overall, by providing large-scale residue-level and interaction-type-specific supervision together with similarity-controlled evaluation, InteractBind supports both the development and systematic assessment of protein-ligand models beyond pair-level binding prediction, while also highlighting a new paradigm for future model design toward more fine-grained interaction understanding.

AI USE STATEMENT

In this work, we used generative AI tools for assisting the construction of the InteractBind dataset, including writing and debugging code for structure collection and the docking affinity. We have not used generative AI tools for generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating or proving mathematical claims, proposing or refining hypotheses, designing research methodology or experiments, or interpreting results, and the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for polishing the manuscript (grammar and readability edits). We have reviewed all AI-assisted work: all AI-generated code was manually reviewed, tested, and verified for correctness by the authors, and all resulting dataset statistics and annotations were validated against the underlying PDB structures; all AI-suggested text edits were manually inspected by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide an anonymous code repository containing the implementation and scripts used in our experiments at https://anonymous.4open.science/r/benchmark-B666/README.md. The InteractBind dataset has also been included in the supplementary materials submitted through the review system. Details of the dataset construction, preprocessing procedures, evaluation splits, negative sampling, model configurations, and experimental settings are described in the paper and code repository. Together, these resources provide the necessary information and materials to reproduce the benchmark construction and the experimental results reported in this work.

REFERENCES

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, 630(8016):493–500, 2024.

Keir Adams, Kento Abeywardane, Jenna C Fromer, and Connor W Coley. Shepherd: Diffusing shape, electrostatics, and pharmacophores for bioisosteric drug design. In ICLR, 2025.

Peizhen Bai, Filip Miljković, Bino John, and Haiping Lu. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. Nature Machine Intelligence, 5(2):126–136, 2023.

Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. Nature reviews genetics, 5(2):101–113, 2004.

Helen Berman, Kim Henrick, Haruki Nakamura, and John L Markley. The worldwide protein data bank (wwpdb): ensuring a single, uniform archive of pdb data. Nucleic acids research, 35 (suppl_1):D301–D303, 2007.

Stephen K Burley, Helen M Berman, Gerard J Kleywegt, John L Markley, Haruki Nakamura, and Sameer Velankar. Protein data bank (pdb): the single global macromolecular structure archive. Protein crystallography: methods and protocols, pp. 627–641, 2017.

Kevin Carter-Fenk and John M Herbert. Reinterpreting π-stacking. Physical Chemistry Chemical Physics, 22(43):24870–24886, 2020.

Jiří Černỳ and Pavel Hobza. Non-covalent interactions in biomacromolecules. Physical Chemistry Chemical Physics, 9(39):5291–5303, 2007.

Ayan Chatterjee, Robin Walters, Zohair Shafi, Omair Shafi Ahmed, Michael Sebek, Deisy Gysi, Rose Yu, Tina Eliassi-Rad, Albert-László Barabási, and Giulia Menichetti. Improving the generalizability of protein-ligand binding predictions with ai-bind. Nature communications, 14(1):1989, 2023.

Lifan Chen, Xiaoqin Tan, Dingyan Wang, Feisheng Zhong, Xiaohong Liu, Tianbiao Yang, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Transformercpi: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. Bioinformatics, 36(16):4406–4414, 2020.

Yigang Chen, Xiang Ji, Ziyue Zhang, Zihao Zhu, Yuming Zhou, Chang Su, Yang-Chi-Dung Lin, Hsi-Yuan Huang, Kangping Wei, Yi Lai, et al. Semi-inductive dataset construction and framework optimization for practical drug target interaction prediction with scopedti. Nature Communications, 2025.

Sarah Crunkhorn. Predicting ligand-protein interactions. Nature reviews. Drug Discovery, 2024.

Suresh Dara, Swetha Dhamercherla, Surender Singh Jadav, CH Madhu Babu, and Mohamed Jawed Ahsan. Machine learning in drug discovery: a review. Artificial Intelligence Review, 55(3): 1947–1999, 2022.

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(11):1046–1051, 2011.

Dennis A Dougherty. The cation- π interaction. Accounts of chemical research, 46(4):885–893, 2013.

Chris Finan, Anna Gaulton, Felix A Kruger, R Thomas Lumbers, Tina Shah, Jorgen Engmann, Luana Galver, Ryan Kelley, Anneli Karlsson, Rita Santos, et al. The druggable genome and support for target identification and validation in drug development. Science translational medicine, 9(383): eaag1166, 2017.

GetContacts developers. GetContacts: Interaction analysis for molecular structures and ensembles. https://getcontacts.github.io/, 2018. Accessed: 2026-05-06.

Michael K Gilson, Tiqing Liu, Michael Baitaluk, George Nicola, Linda Hwang, and Jenny Chong. Bindingdb in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic acids research, 44(D1):D1045–D1053, 2016.

Hamid Hadipour, Yan Yi Li, Yan Sun, Chutong Deng, Leann Lac, Rebecca Davis, Silvia T Cardona, and Pingzhao Hu. Graphban: An inductive graph-based approach for enhanced prediction of compound-protein interactions. Nature Communications, 16(1):2541, 2025.

Tymor Hamamsy, James T Morton, Robert Blackwell, Daniel Berenberg, Nicholas Carriero, Vladimir Gligorijevic, Charlie EM Strauss, Julia Koehler Leman, Kyunghyun Cho, and Richard Bonneau. Protein remote homology detection and structural alignment using deep learning. Nature biotechnology, 42(6):975–985, 2024.

Alex S Holehouse and Birthe B Kragelund. The molecular basis for cellular function of intrinsically disordered protein regions. Nature Reviews Molecular Cell Biology, 25(3):187–211, 2024.

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.

Yinjun Jia, Bowen Gao, Jiaxin Tan, Jiqing Zheng, Xin Hong, Wenyu Zhu, Haichuan Tan, Yuan Xiao, Liping Tan, Hongyi Cai, et al. Deep contrastive learning enables genome-wide virtual screening. Science, 391(6781):eads9530, 2026.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. Advances in neural information processing systems, 31, 2018.

Mario Krenn, Qianxiang Ai, Senja Barthel, Nessa Carson, Angelo Frei, Nathan C Frey, Pascal Friederich, Théophile Gaudin, Alberto Alexander Gayle, Kevin Maik Jablonka, et al. Selfies and the future of molecular string representations. Patterns, 3(10), 2022.

Jie Liang, Clare Woodward, and Herbert Edelsbrunner. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein science, 7(9): 1884–1897, 1998.

Hui Liu, Jianjiang Sun, Jihong Guan, Jie Zheng, and Shuigeng Zhou. Improving compound–protein interaction prediction by building up highly credible negative samples. Bioinformatics, 31(12): i221–i229, 2015.

Matthew R Masters, Amr H Mahmoud, and Markus A Lill. Investigating whether deep learning models for co-folding learn the physics of protein-ligand interactions. Nature Communications, 16(1):8854, 2025.

Carla Mattos and Dagmar Ringe. Locating and characterizing binding sites on proteins. Nature biotechnology, 14(5):595–599, 1996.

Zhaohan Meng, Zaiqiao Meng, Ke Yuan, and Iadh Ounis. FusionDTI: Fine-grained binding discovery with token-level fusion for drug-target interaction. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 2025. Association for Computational Linguistics.

Raúl Miñán, Javier Gallardo, Álvaro Ciudad, and Alexis Molina. Informed protein–ligand docking via geodesic guidance in translational, rotational and torsional spaces. Nature Machine Intelligence, 7 (9):1555–1560, 2025.

Alex Morehead, Nabin Giri, Jian Liu, Pawan Neupane, and Jianlin Cheng. Assessing the potential of deep learning for protein–ligand docking. Nature Machine Intelligence, pp. 1–10, 2025.

Shinji Murai, Fumitoshi Kakiuchi, Shinya Sekine, Yasuo Tanaka, Asayuki Kamatani, Motohiro Sonoda, and Naoto Chatani. Efficient catalytic addition of aromatic carbon-hydrogen bonds to olefins. Nature, 366(6455):529–531, 1993.

Michael M Mysinger, Michael Carchia, John J Irwin, and Brian K Shoichet. Directory of useful decoys, enhanced (dud-e): better ligands and decoys for better benchmarking. Journal of medicinal chemistry, 55(14):6582–6594, 2012.

Ngoc-Quang Nguyen, Gwanghoon Jang, Hajung Kim, and Jaewoo Kang. Perceiver cpi: a nested cross-attention network for compound–protein interaction prediction. Bioinformatics, 39(1): btac731, 2023.

Paola Peluso and Bezhan Chankvetadze. Recognition in the domain of molecular chirality: from noncovalent interactions to separation of enantiomers. Chemical Reviews, 122(16):13235–13400, 2022.

Svetlana Pylaeva, Martin Brehm, and Daniel Sebastiani. Salt bridge in aqueous solution: Strong structural motifs but weak enthalpic effect. Scientific reports, 8(1):13626, 2018.

Nicholas Rego and David Koes. 3dmol. js: molecular visualization with webgl. Bioinformatics, 31 (8):1322–1324, 2015.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. Journal of chemical information and modeling, 50(5):742–754, 2010.

Philipp Schake, Sarah Naomi Bolz, Katja Linnemann, and Michael Schroeder. Plip 2025: introducing protein–protein interactions to the protein–ligand interaction profiler. Nucleic acids research, 53 (W1):W463–W465, 2025.

Steven D Schwartz and Vern L Schramm. Enzymatic transition states and dynamic motion in barrier crossing. Nature chemical biology, 5(8):551–558, 2009.

Ian Sitarik, Quyen V Vu, Justin Petucci, Paulina Frutos, Hyebin Song, and Edward P O'Brien. A widespread protein misfolding mechanism is differentially rescued in vitro by chaperones based on gene essentiality. Nature Communications, 16(1):10870, 2025.

Antonia Stank, Daria B Kokh, Jonathan C Fuller, and Rebecca C Wade. Protein binding pocket dynamics. Accounts of chemical research, 49(5):809–815, 2016.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. In International Conference on Learning Representations, volume 2024, pp. 6987–7009, 2024.

Jiaxin Tan, Yuan Xiao, Fang Kong, Xiaochun Zhang, Hanwen Xu, Angqi Zhu, Yiming Liu, Jianlin Lei, Boxue Tian, Yafei Yuan, et al. Molecular basis of human noradrenaline transporter reuptake and inhibition. Nature, 632(8026):921–929, 2024.

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.

Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of computational chemistry, 31(2):455–461, 2010.

Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. Nature Biotechnology, 42(2):243–246, 2024.

Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic acids research, 50(D1):D439–D444, 2022.

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, et al. Boltz-1 democratizing biomolecular interaction modeling. BioRxiv, pp. 2024–11, 2025.

Lixu Yang, Catherine Adam, Gary S Nichol, and Scott L Cockroft. How much do van der waals dispersion forces contribute to molecular recognition in solution? Nature chemistry, 5(12): 1006–1010, 2013.

Xiaoting Zeng, Weilin Chen, and Baiying Lei. Cat-dti: cross-attention and transformer network with domain adaptation for drug-target interaction prediction. BMC bioinformatics, 25(1):141, 2024.

Zhijun Zhang, Lijun Quan, Junkai Wang, Liangchen Peng, Qiufeng Chen, Bei Zhang, Lexin Cao, Yelu Jiang, Geng Li, Liangpeng Nie, et al. Labind: Identifying protein binding ligand-aware sites via learning interactions between ligand and protein. Nature Communications, 16(1):7712, 2025.

Qichang Zhao, Haochen Zhao, Kai Zheng, and Jianxin Wang. Hyperattentiondti: improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism. Bioinformatics, 38(3):655–662, 2022.

Marinka Zitnik, Rok Sosič, Sagar Maheshwari, and Jure Leskovec. Biosnap datasets: Stanford biomedical network dataset collection. Note: http://snap. stanford. edu/biodata Cited by, 5(1), 2018.

A PRELIMINARIES

A.1 COVALENT BONDS AND NON-COVALENT INTERACTIONS.

Molecular systems involve two fundamental types of interactions: *covalent bonds* and *non-covalent interactions* (Černỳ & Hobza, 2007). Covalent bonds govern intra-molecular organization and are typically treated as fixed in modeling. In contrast, non-covalent, generally reversible interactions, including hydrogen bonds (Murai et al., 1993), salt bridges (Pylaeva et al., 2018), van der Waals contacts (Yang et al., 2013), hydrophobic contacts (Sitarik et al., 2025), π-π stacking (Carter-Fenk & Herbert, 2020), and cation-π interactions (Dougherty, 2013), do not involve bond formation but arise from physicochemical compatibility between molecules. These interactions are predominantly inter-molecular and collectively determine protein-ligand binding, including affinity, specificity, and binding-site localization. However, such local mechanistic details are typically absent from existing datasets. InteractBind addresses this limitation by explicitly representing such local interaction patterns through sequence-level interaction maps.

A.2 BINDING-SITE AND BINDING POCKET.

In structural biology, a *binding pocket* generally refers to a broader three-dimensional region on the protein surface that can accommodate a ligand, often characterized by geometric cavities or surface concavities (Liang et al., 1998). In contrast, a *binding-site* is more specific and denotes the subset of protein residues that directly participate in ligand recognition through non-covalent interactions (Mattos & Ringe, 1996). While binding pockets describe a spatial region that may potentially host a ligand, binding sites capture the actual interacting components that determine molecular recognition and binding specificity. Although the two concepts are closely related and sometimes used interchangeably, they differ in granularity and interpretability. Binding pocket identification is often based on structural geometry and provides a coarse description of possible ligand-accessible regions, whereas binding-site annotation reflects the underlying specific physicochemical interactions that stabilize binding (Stank et al., 2016). In many sequence-based settings where high-resolution structural information is limited, binding pocket definitions can be ambiguous, making it difficult to evaluate whether a model has learned the true interaction determinants (Masters et al., 2025). In this work, we therefore focus on binding sites at the residue level, defined by experimentally-derived interaction annotations. This formulation provides a more precise and mechanistically grounded target for evaluation, enabling a fine-grained assessment of whether models localize the residues involved in protein-ligand recognition.

Broader impact. Protein-ligand binding prediction underpins computational drug discovery and molecular design, yet strong binary binding prediction does not necessarily indicate that models can localize binding sites or identify the interaction patterns governing molecular recognition. Models may instead exploit dataset-specific correlations, leading to unreliable predictions under distribution shift. By supporting a systematic evaluation of binding-site localization and interaction-type-specific contact identification, InteractBind encourages a shift from outcome-only evaluation towards fine-grained assessment of protein-ligand models. This may promote the development of more physically interpretable methods and improve the reliability of computational drug discovery pipelines, where incorrect binding-site or interaction-pattern inference can lead to misleading biological conclusions or suboptimal compound prioritization.

Limitations. The current version of InteractBind focuses on sequence-level evaluation and does not yet include a structure-level benchmark. Extending the evaluation to three-dimensional structures is non-trivial, since protein-ligand binding often admits multiple plausible poses and involves conformational flexibility that is not captured by a single static structure. This ambiguity makes it difficult to define a consistent and reliable ground truth for structure-level comparison. As a result, structure-based evaluation may introduce additional noise and reduce interpretability compared to residue-level sequence-based assessment. Developing a robust and standardized protocol for structure-level benchmarking remains an important and challenging direction for future work.

Table 3: Statistics of the ID and protein similarity-controlled OOD benchmark datasets.

Dataset	Positive:Negative	Split criterion	Pairs	Ligands	Proteins
InteractBind-ID	1:10	Stratified 5-fold	1,093,301	9,017	11,473
InteractBind-P-25% OOD	1:1	Protein similarity	13,126	3,278	9,527
InteractBind-P-28% OOD	1:1	Protein similarity	13,126	2,966	8,957
InteractBind-P-31% OOD	1:1	Protein similarity	13,126	3,202	10,083
InteractBind-P-33% OOD	1:1	Protein similarity	13,126	4,807	11,084

B Curation of the InteractBind Dataset

B.1 Structure Collection and Sample Construction.

We construct **InteractBind**, a large-scale non-covalent interaction supervised protein-ligand binding dataset, by collecting protein-ligand complex structures from the Protein Data Bank (PDB) format (version 3.2) using py3Dmol (Rego & Koes, 2015). Structures were filtered to retain only those that contain at least one organic ligand that binds to the protein. Each retained structure underwent a standardized pre-processing pipeline. All inorganic ligands and crystallographic water molecules were removed. Moreover, complexes containing metal ions in either the ligand or the protein were excluded. If a PDB entry contained multiple organic ligands, distance-based filtering was applied to extract individual ligand-protein complexes: for each ligand, only protein chains within $0.5\,\mathrm{nm}$ of the ligand were retained, while others were discarded. As a result, PDB files containing multiple ligands were split into multiple complexes, each consisting of a single ligand and its interacting protein chain(s). For each complex, the input data were then separated into two files while preserving the original atomic coordinates. Protein structures were converted to structure-aware (SA)/FASTA format. We derive structure-aware (SA) protein representations via Foldseek (Van Kempen et al., 2024), providing an explicit local structural context (Su et al., 2024). Ligands were converted from PDB to SMILES and subsequently to SELFIES representations.

B.2 Interaction Annotation and Strength Modeling

Protein-ligand binding is governed by a diverse set of non-covalent interactions, including hydrogen bonds (Murai et al., 1993), salt bridges (Pylaeva et al., 2018), van der Waals contacts (Yang et al., 2013), hydrophobic contacts (Sitarik et al., 2025), π-π stacking (Carter-Fenk & Herbert, 2020), and cation-π interactions (Dougherty, 2013). These interactions jointly determine binding affinity, specificity, and structural complementarity, forming the physicochemical basis of molecular recognition (Holehouse & Kragelund, 2024). Protein-ligand interactions in **InteractBind** were detected using rule-based criteria grounded in geometric and chemical constraints. Hydrogen bonds, salt bridges, π-π stacking, cation-π interactions, and hydrophobic contacts were identified using PLIP (Schake et al., 2025). Van der Waals interactions, which are not captured by PLIP, were computed using GetContacts (GetContacts developers, 2018), and only attractive Van der Waals interactions were retained. To quantify distance-based interaction weight, we adopt a distance-dependent piecewise linear decay function that maps interatomic or geometric distances to continuous values in the range $[10^{-6}, 1]$ within interaction-specific cutoff windows. Distance thresholds and geometric constraints for each non-covalent interaction are summarized in Tab. 4. Angular and geometric criteria, when applicable, are used as binary filters, while the distance-based decay controls the distance-based interaction weight.

Based on the detected interactions and their distance-dependent weights, weighted interaction maps were constructed at the sequence level. For each protein-ligand complex in InteractBind, an $m \times n$ matrix was generated, where m and n denote the numbers of ligand atoms and protein FASTA residues, respectively. Each matrix entry encodes the distance-based interaction weight between a ligand atom and a protein residue. For each complex, six interaction-specific maps were generated, together with an additional unified map obtained by summing all interaction-specific maps. All interaction maps were stored using unique identifiers for efficient retrieval during training.

Table 4: Geometric criteria and strength definitions for key non-covalent interactions in **InteractBind**.

Non-covalent interaction	Parameter	Geometric criterion	Distance-based weight $s(d)$
Hydrogen bond	Donor-acceptor distance	$2.2\,\text{Å} < d \leq 4.1\,\text{Å}$	$s(d) = \begin{cases} \max\left(10^{-6}, 1 - \frac{d - 2.2\,\text{Å}}{4.1\,\text{Å} - 2.2\,\text{Å}}\right), & 2.2\,\text{Å} < d \leq 4.1\,\text{Å} \\ 10^{-6}, & \text{otherwise} \end{cases}$
	Donor angle	$\geq 100°$	(angle used as filter; distance controls s)
Salt bridge	Distance between charge centres	$2.8\,\text{Å} < d \leq 5.5\,\text{Å}$	$s(d) = \begin{cases} \max\left(10^{-6}, 1 - \frac{d - 2.8\,\text{Å}}{5.5\,\text{Å} - 2.8\,\text{Å}}\right), & 2.8\,\text{Å} < d \leq 5.5\,\text{Å} \\ 10^{-6}, & \text{otherwise} \end{cases}$
π-π stacking	Ring centroid distance	$3.4\,\text{Å} < d \leq 5.5\,\text{Å}$	$s(d) = \begin{cases} \max\left(10^{-6}, 1 - \frac{d - 3.4\,\text{Å}}{5.5\,\text{Å} - 3.4\,\text{Å}}\right), & 3.4\,\text{Å} < d \leq 5.5\,\text{Å} \\ 10^{-6}, & \text{otherwise} \end{cases}$
	Angle deviation	$\leq 30°$	(geometry filter)
	Ring offset	$\leq 2.0\,\text{Å}$	(geometry filter)
Cation-π interaction	Cation-ring centroid distance	$3.0\,\text{Å} < d \leq 6.0\,\text{Å}$	$s(d) = \begin{cases} \max\left(10^{-6}, 1 - \frac{d - 3.0\,\text{Å}}{6.0\,\text{Å} - 3.0\,\text{Å}}\right), & 3.0\,\text{Å} < d \leq 6.0\,\text{Å} \\ 10^{-6}, & \text{otherwise} \end{cases}$
Hydrophobic contact	Distance between apolar atoms	$3.0\,\text{Å} < d \leq 5.0\,\text{Å}$	$s(d) = \begin{cases} \max\left(10^{-6}, 1 - \frac{d - 3.0\,\text{Å}}{5.0\,\text{Å} - 3.0\,\text{Å}}\right), & 3.0\,\text{Å} < d \leq 5.0\,\text{Å} \\ 10^{-6}, & \text{otherwise} \end{cases}$
Van der Waals forces	Distance between non-hydrogen atoms	$d_0 < d \leq 4.5\,\text{Å}$, where $d_0 = R_{\text{vdW}}(A) + R_{\text{vdW}}(B)$	$s(d) = \begin{cases} \max\left(10^{-6}, 1 - \frac{d - d_0}{4.5\,\text{Å} - d_0}\right), & d_0 < d \leq 4.5\,\text{Å} \\ 10^{-6}, & \text{otherwise} \end{cases}$

C BASELINES

We compare the following eight representative baselines:

1. MolTrans (Huang et al., 2021), which tokenizes protein sequences and ligands into substructures, encodes them with transformer-based encoders, and applies an interaction module to model cross-modal binding patterns.
2. TransformerCPI (Chen et al., 2020), which represents proteins and compounds as sequences and employs transformer architectures to model their interactions for compound-protein interaction prediction.
3. HyperAttentionDTI (Zhao et al., 2022), which introduces a hyper-attention mechanism to capture fine-grained interactions between protein residues and ligand atoms for drug-target interaction prediction.
4. PerceiverCPI (Nguyen et al., 2023), which adopts a Perceiver-based architecture to encode long protein and ligand sequences and model compound-protein interactions through latent cross-attention.
5. CAT-DTI (Zeng et al., 2024), which uses cross-attention mechanisms to jointly model protein and ligand representations and improve interaction prediction.
6. DrugBAN (Bai et al., 2023), which encodes ligands with a graph neural network and proteins with a 1D CNN, and employs a bilinear attention network (Kim et al., 2018) to capture fine-grained residue-atom interactions before final prediction.
7. GraphBAN (Hadipour et al., 2025), which extends bilinear attention-based interaction modeling with graph neural representations for ligands and protein encoders for interaction prediction.
8. FusionDTI (Meng et al., 2025), which uses token-level fusion to model residue-atom interactions between protein and ligand representations for drug-target interaction prediction.

C.1 DERIVATION AND ALIGNMENT OF MODEL INTERACTION MAPS

To ensure a consistent binding-site localization evaluation across models, we explicitly identify the source of each model-derived interaction map $\mathbf{P}$ and align model-specific interaction units to protein residues. Because the evaluated architectures expose interaction information at different granularities, $\mathbf{P}$ is not assumed to correspond to native attention for every model. The extraction and alignment procedures are summarized below:

- **HyperAttentionDTI.** $\mathbf{P}$ is obtained from the returned ligand–protein cross-attention map. The native pairwise interaction map is retained, and convolution outputs are aligned to the center residues of their corresponding receptive fields.

- **DrugBAN. P** is obtained from the bilinear attention map. Attention scores are averaged over BAN glimpses, and convolution positions on the protein side are mapped to their corresponding residue centers.
- **CAT-DTI. P** is taken directly from the returned cross-attention map. The model provides a native pairwise interaction map whose ligand-atom and protein-residue positions can be aligned directly.
- **PerceiverCPI.** Since the model does not expose a directly comparable pairwise attention map, we use input-gradient saliency as a post-hoc attribution signal. The interaction map is constructed from the outer product of ligand-atom hidden-state gradient magnitudes and protein-residue embedding gradient magnitudes.
- **GraphBAN. P** is obtained from the bilinear attention map. Attention scores are averaged over BAN glimpses, and protein convolution positions are mapped to the corresponding residue centers.
- **MolTrans. P** is obtained from the learned ligand-token/protein-token interaction matrix. Protein BPE tokens are mapped back to the center residues covered by the corresponding tokens.
- **TransformerCPI. P** is obtained from the returned decoder cross-attention. Attention scores are averaged over heads, and protein k-mer outputs are mapped to the center residues of the corresponding k-mers.
- **FusionDTI. P** is obtained from the token-fusion/BAN attention module. BAN glimpses are averaged when applicable, and the resulting map is transposed when necessary to obtain the ligand-atom $\times$ protein-residue orientation.

After model-specific extraction and alignment, all interaction maps are projected onto the original protein-residue axis. The ligand dimension is then aggregated using the same maximum operation described in Appendix D before computing localization metrics.

D EVALUATION METRIC: BRHR

We define a Top-K metric termed the *Binding Residue Hit Rate* (BRHR). For a given protein-ligand pair, let $\mathbf{P} \in \mathbb{R}^{m\times n}$ denote the model-derived interaction map, where m is the number of ligand tokens or atoms, n is the number of protein residues, and each entry P_{ij} indicates the predicted interaction score between ligand unit i and protein residue j. Because BRHR focuses on protein-side binding-site localization, we aggregate the interaction map along the ligand dimension to obtain a residue-level score:

$$s_j = \max_{i\in\{1,\dots,m\}} P_{ij}, \tag{1}$$

where s_j denotes the predicted binding-site score for protein residue j. We then rank all protein residues by s_j and define:

$$R_K = \text{Top-}K(\{s_j\}_{j=1}^{n}) \tag{2}$$

as the set of Top-K predicted protein residues. Let T denote the set of ground-truth binding-site residues. We define BRHR for a single sample as follows:

$$\text{BRHR}(K) = \begin{cases} 1, & \text{if } R_K \cap T \neq \varnothing, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

In other words, a prediction is counted as correct if at least one of the Top-K predicted protein residues matches a ground-truth binding-site residue. The dataset-level score is computed as:

$$\text{BRHR}_{\text{avg}}(K) = \frac{1}{N}\sum_{i=1}^{N} \text{BRHR}_i(K), \quad \text{BRHR}_{\text{avg}}(K) \in [0,1], \tag{4}$$

where N is the number of evaluated protein-ligand pairs. Larger values indicate better binding-site localization performance.

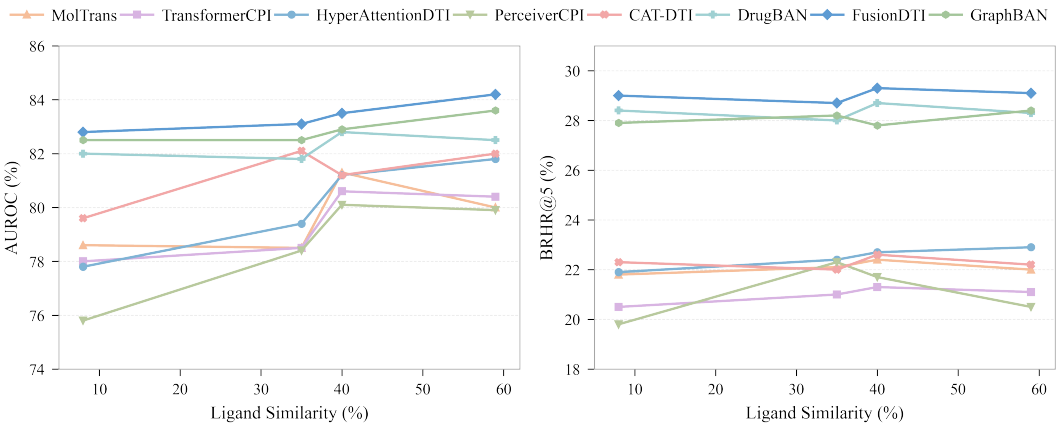

Figure 6: Performance across four ligand similarity-controlled OOD datasets. The four datasets have maximum train-test ligand similarities of 8%, 35%, 40%, and 59%, with a lower similarity indicating a stronger ligand-side distribution shift. **a.** Binary binding prediction performance measured by AUROC on the four datasets. **b.** Binding-site localization performance measured by BRHR@5 on the four datasets.

E LIGAND SIMILARITY-CONTROLLED DATASET SPLITS

For the OOD setting, we also construct ligand similarity-controlled evaluation scenarios to assess whether chemical dissimilarity between training and test ligands affects model generalization. Ligand chemical similarity is quantified using Tanimoto similarity over extended connectivity fingerprints (ECFP) (Rogers & Hahn, 2010), yielding four splits with maximum train-test ligand similarities of 8%, 35%, 40%, and 59%. To isolate the effect of ligand-side distribution shift, protein similarity is fixed at 33% across all splits. Each dataset is divided into training, validation, and test subsets with a ratio of 8:1:1. As shown in Fig. 6a, binary binding prediction remains relatively stable across the four ligand similarity-controlled splits. FusionDTI shows the strongest performance, increasing only slightly from 82.8% AUROC at 8% ligand similarity to 84.2% at 59%. GraphBAN similarly varies within a narrow range, from 82.5% to 83.6%, while DrugBAN remains stable between 81.8% and 82.8%. Other baselines also show only modest variation, for example CAT-DTI ranges from 79.6% to 82.1%, HyperAttentionDTI from 77.8% to 81.8%, and MolTrans from 78.5% to 81.3%. Overall, these results indicate that ligand similarity has a comparatively limited effect on binary binding prediction under the current InteractBind setting. Binding-site localization shows an even smaller change across ligand similarity-controlled splits (Fig. 6b). FusionDTI remains the strongest model, with BRHR@5 varying only slightly from 28.7% to 29.3%, while DrugBAN ranges from 28.0% to 28.7% and GraphBAN from 27.8% to 28.4%. The remaining baselines also fluctuate within narrow ranges, including MolTrans from 21.8% to 22.4%, HyperAttentionDTI from 21.9% to 22.9%, and CAT-DTI from 22.0% to 22.6%. Although PerceiverCPI shows somewhat less consistent behavior, its BRHR@5 still remains within a limited range of 19.8% to 22.3%. Compared with the protein similarity-controlled splits, these changes are substantially smaller, indicating that protein-side similarity exerts a stronger influence on model inference than ligand-side chemical similarity. One possible explanation is that binding-site localization is evaluated on the protein axis, making protein sequence diversity more directly related to the difficulty of identifying ligand-binding residues. These results support the use of protein similarity-controlled splits as the primary OOD evaluation setting, while ligand-controlled splits provide a complementary assessment of chemical generalization.

F LICENSES AND TERMS OF USE FOR EXISTING RESOURCES

In the following, we summarize the licenses and terms of use of the main existing resources used in this work.

- **Protein Data Bank (Burley et al., 2017)** Source: wwPDB / RCSB PDB archive. License: CC0 1.0 Universal (public domain dedication). We use experimentally resolved protein-ligand complex structures from the PDB for dataset construction, and attribute original depositors where appropriate.
- **PLIP (Schake et al., 2025)** Source: `pharmai/plip`. License: GNU GPL v2.0. We use PLIP to identify hydrogen bonds, salt bridges, π-π stacking, cation-π interactions, and hydrophobic contacts during interaction annotation.
- **GetContacts (GetContacts developers, 2018)** Source: `getcontacts/getcontacts`. License: Apache License 2.0. We use GetContacts to compute van der Waals interactions not covered by PLIP.
- **AutoDock Vina (Trott & Olson, 2010)** Source: official AutoDock Vina release / `ccsb-scripps/AutoDock-Vina`. License: Apache License 2.0. We use AutoDock Vina for docking-based affinity estimation.