Pre-training on noncovalent interactions in synthetic protein-ligand structures to better predict binding affinity

Anonymous Authors[1]

Abstract

Accurate prediction of protein-ligand binding affinity is a central challenge in computational drug discovery, yet existing graph neural network approaches typically represent protein-ligand complexes as homogeneous atom-level graphs, neglecting the role of aromatic ring systems and the rich hierarchy of noncovalent interactions that govern binding. In this work, we introduce the **Protein-Ligand Interaction Pre-training** or PLIP approach, a heterogeneous equivariant graph transformer that explicitly encodes four node types - ligand atoms, protein atoms, ligand rings, and protein rings connected by interaction-specific edge relations such as hydrogen bonds, hydrophobic contacts, π-π stacking, and cation-π interactions. We pre-train on interactions found in 5.1 million synthetic protein-ligand structures from the Structural and Interaction Repository (SAIR) using a multi-task self-supervised objective comprising interaction-type classification, interatomic distance regression, and binding affinity prediction. Systematic evaluation on four drug targets - acetylcholinesterase (AChE), SARS-CoV-2 main protease (SARS-M^{pro}), Zika protease, and μ-opioid receptor - demonstrates that multi-task pre-training on all three pre-training objectives (interaction type, distance, and affinity) achieves Pearson correlations of 0.667 on AChE (+26% over training from scratch), 0.490 on Zika (+6%), 0.374 on μ-opioid receptor (+32%), and 0.255 on SARS-M^{pro} (+39%). Comparisons against competing baselines demonstrate the benefits of pre-training on protein-ligand interactions for structure-based binding affinity prediction.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the AI for Science workshop (ICML 2026). Do not distribute.

1. Introduction

Structure-based prediction of protein-ligand binding affinity is fundamental to rational drug design, virtual screening, and lead optimization (Li et al., 2019; Volkov et al., 2022). Classical scoring functions decompose binding into physics-based energy terms (Gohlke et al., 2000) or empirically fitted linear models (Wang et al., 2004), but they struggle to capture the complexity of protein-ligand interactions. Machine learning approaches, particularly graph neural networks (GNNs) and equivariant graph neural networks (EGNNs), have emerged as powerful alternatives by learning representations directly from atomic coordinates (Torng & Altman, 2019; Lim et al., 2019; Jiang et al., 2021). Methods such as Edge-enhanced Interaction Graph Network or EIGN (Yang et al., 2025) use a graph isomorphism network (GIN) to propagate atom-level messages, while Protein-Ligand Binding Affinity Prediction using a Novel Interaction-Based Graph Neural Network Framework or PLAIG (Samudrala et al., 2025) augments GNN architectures with interaction features derived from BINANA analysis (Durrant & McCammon, 2011). There have also been recent efforts in 3D-based atomic neural networks to learn from multiple docked poses (Shim et al., 2022; Kim et al., 2025).

Current graph-based models treat the protein-ligand complex as a *homogeneous* graph of atoms, discarding the chemical hierarchy in which aromatic rings mediate critical noncovalent interactions such as π-π stacking and cation-π contacts (Hunter & Sanders, 1990; Ma & Dougherty, 1997). Second, although self-supervised pre-training has yielded dramatic improvements in natural language processing and computer vision, its application to modelling protein-ligand binding remains limited (Gorantla et al., 2025; Li & Gong, 2025). A potential reason is that there is a lack of a large corpus of protein-ligand complexes that could be used in the pretraining. However, this has started to change with the advent of AI biomolecular structure prediction models such as AlphaFold (Jumper et al., 2021), Boltz (Wohlwend et al., 2024), and OpenFold (Ahdritz et al., 2024). Such tools could be used to generate structures based on sequence data already publicly available. One such effort, which resulted in a large dataset of 5 million Boltz 1x-generated structures, is the Structural and Interaction Repository (SAIR) (Lemos

et al., 2026). Next, existing pretraining strategies for binding affinity prediction often operate on two-dimensional molecular graphs or SMILES strings (Hu et al., 2020; Chithrananda et al., 2020; Gorantla et al., 2025; Yi et al., 2026) rather than on the three-dimensional protein-ligand structures, which may limit the richness of the input representation. It may be more straightforward and interpretable to operate on three-dimensional structures when modelling binding. Thirdly, atom-level featurization in most models is limited to one-hot encodings of element type, ignoring the local three-dimensional chemical environment that modulates reactivity and binding.

In this work, we address these gaps through the following contributions:

1. **Heterogeneous graph representation with explicit ring nodes.** We construct protein-ligand interaction graphs with four node types (ligand atoms, protein atoms, ligand rings, protein rings) connected by interaction-specific edge relations (hydrogen bonds, hydrophobic contacts, π-π stacking, cation-π, π-metal, and ring membership) using the Open Drug Discovery Toolkit (Wójcikowski et al., 2015) or ODDT.

2. **Multi-task self-supervised pretraining on 5.1 million structures.** We build our PLIP approach on the Structural and Interaction Repository (SAIR) with a joint objective comprising interaction-type classification, interatomic distance regression, and binding affinity prediction to learn rich, transferable representations of protein-ligand interactions.

3. **Atomic environment featurization.** We adapt the atomic environment vector (AEV) featurization based on radial symmetry functions over 22 chemical species (Smith et al., 2017a; Valsson et al., 2025), capturing the local three-dimensional neighborhood of each atom within a spherical radius cutoff of 5.0.

4. **Systematic evaluation across four drug targets.** We evaluate PLIP on four therapeutically relevant targets - AChE, SARS-M$^{\text{pro}}$, Zika protease, and μ-opioid receptor -spanning diverse dataset sizes (606 to 7,108 affinity measurements as seen in Table 1) using a novel robust splitting strategy called **PLEC-Butina**, and compare against EIGN and PLAIG baselines.

2. Methodology

Given a protein-ligand complex with three-dimensional atomic coordinates, we aim to predict the binding affinity $y \in \mathbb{R}$ (expressed as pIC_{50}). We represent the complex as a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{T}_V, \mathcal{T}_E)$ where $\mathcal{V}$ is the set of nodes, $\mathcal{E}$ the set of edges, $\mathcal{T}_V$ a set of node types, and $\mathcal{T}_E$ a set of edge (relation) types. Each node $v \in \mathcal{V}$ carries a feature vector $\mathbf{x}_v \in \mathbb{R}^{d_{\tau(v)}}$ (where $\tau(v) \in \mathcal{T}_V$ is the node type and $d_{\tau(v)}$ is the type-specific feature dimension) and a three-dimensional position $\mathbf{p}_v \in \mathbb{R}^3$. Each edge $(u, v, r) \in \mathcal{E}$ has a relation type $r \in \mathcal{T}_E$ and has an attribute vector encoding the interaction type and interatomic distance.

2.1. Heterogeneous Graph Construction

We construct the heterogeneous protein-ligand interaction graphs from a three-dimensional structure (PDB or mmCIF format) through the following pipeline.

Node extraction. We define four node types: *ligand atom*, *protein atom*, *ligand ring*, and *protein ring*. Atom nodes are extracted directly from the parsed structure and carry three-dimensional Cartesian coordinates. Ring nodes are derived from the smallest set of smallest rings (SSSR) computed via Open Babel, (O'Boyle et al., 2011) with positions set to the arithmetic mean of the Cartesian coordinates of their constituent atoms, $\mathbf{p}_{\text{ring}} = \frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} \mathbf{p}_i$, where $\mathcal{R}$ is the set of atoms in the ring.

Edge construction and interaction fingerprinting. We compute noncovalent interactions using the Open Drug Discovery Toolkit (ODDT) (Wójcikowski et al., 2015) with a distance cutoff of 5.0 Å. This yields eight interaction types or classes: hydrogen bond, hydrophobic, salt bridge, π-stack, cation-π, π-metal, metal coordination, halogen bond. We also define a 9th class called *close-contact*, which describes an edge between a protein atom and a ligand atom where no noncovalent interaction exists, but they are within 5.0 Å of each other.

Therefore, each atom-atom edge carries a 10-dimensional attribute vector: a 9-dimensional multi-hot encoding of the interaction type concatenated with the pairwise Euclidean distance. To focus computation on the binding interface, we retain only ligand atoms within 5.0 Å of any protein atom, along with all protein atoms and ring nodes involved in at least one interaction, ensuring every node has degree ≥ 1.

2.2. Atomic Environment Vectors

To capture the local three-dimensional chemical environment, we adapt the (**A**tomic **E**nvironment **V**ector featurisation, a 352-dimensional descriptor based on radial symmetry functions (Behler, 2011; Smith et al., 2017b). For each atom i in the joint protein-ligand complex, the radial AEV component for chemical species s and radial basis index k is:

$$G_{k,s}^{(\text{rad})}(i) = \sum_{\substack{j \neq i \\ d_{ij} < r_c}} f_c(d_{ij}) \cdot \exp\left(-\eta_r \left(d_{ij} - R_{s,k}\right)^2\right) \cdot \delta(s_j, s) \tag{1}$$

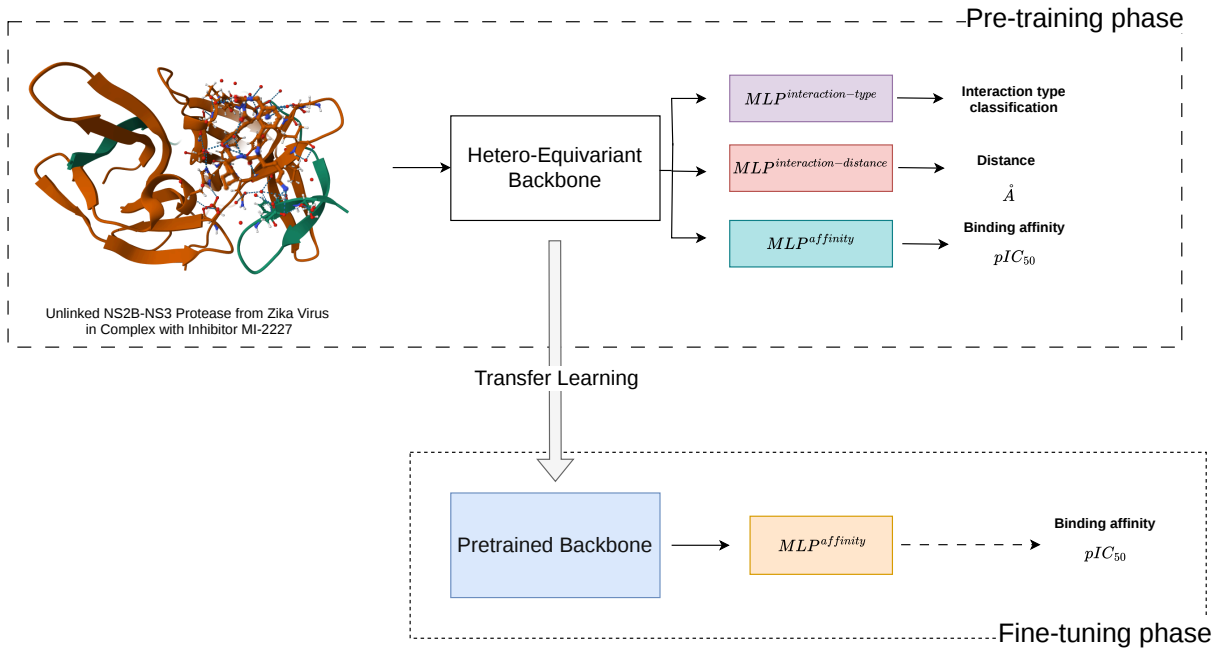

Figure 1. An overview of the PLIP framework. Here, inhibitor MI-2227 is bound to the NS2B-NS3 protease from Zika virus (pdbid: 7ZYS)

where $d_{ij} = \|\mathbf{p}_i - \mathbf{p}_j\|$ is the interatomic distance, f_c is a cosine cutoff function:

$$f_c(r) = \begin{cases} \frac{1}{2}\left[\cos\left(\frac{\pi r}{r_c}\right) + 1\right] & r < r_c \\ 0 & r \geq r_c \end{cases} \tag{2}$$

$r_c = 5.0$ Å is the cutoff radius, $\eta_r = 19.7$ is the radial width parameter, $R_{s,k}$ are 16 radial shift values uniformly spaced from 0.80 Å to 4.83 Å, and $\delta(s_j, s)$ is the Kronecker delta selecting neighbors of species s.

We define 22 chemical species spanning the elements most commonly encountered in protein-ligand complexes: H, B, C, N, O, F, Na, Mg, Si, P, S, Cl, K, Ca, Mn, Fe, Co, Ni, Cu, Zn, Br, and I. The total AEV dimensionality is thus $22 \times 16 = 352$. For computational efficiency, neighbor search is performed using a k-dimensional tree, avoiding $O(N^2)$ pairwise distance computation on large proteins.

2.3. Hetero-Equivariant Backbone

The PLIP backbone processes the heterogeneous graph through three stages: type-specific embedding, node unification, and a stack of transformer layers.

Type-specific embedding. Each node type τ has a dedicated linear projection $\mathbf{h}_v^{(0)} = \mathbf{W}_\tau \mathbf{x}_v + \mathbf{b}_\tau$ mapping the type-specific input dimension d_τ to a shared hidden dimension d_h (default 256), followed by per-type layer normalization. After embedding, all node types are concatenated into a single feature matrix $\mathbf{H}^{(0)} \in \mathbb{R}^{N \times d_h}$ (where N is the total number of nodes) with a unified index space.

2.4. Prediction Heads

PLIP uses the following three prediction heads operating on the backbone's output representations to pre-train on the interaction-type, distance, and affinity.

Interaction-type head. A three-layer multi-layer perceptron or MLP classifies each positive edge into one of nine interaction types. The loss is cross-entropy with the close-contact class (index 8) masked, as close contacts dominate the edge distribution but carry limited physicochemical information. Close-contact edges remain in the graph for message passing.

Distance head. A three-layer MLP regresses the pairwise Euclidean distance of each edge, trained with mean squared error loss.

Affinity head (molecule-level label). A three-layer MLP predicts a per-complex binding affinity from the mean-pooled node embeddings. For complexes with multiple docked poses, an attention-based pooling module computes per-pose attention scores via a two-layer network, normalizes them with softmax over poses of the same complex,

and produces the complex-level prediction as the attention-weighted sum of per-pose affinities:

$$\hat{y}_{\text{complex}} = \sum_{g \in \text{poses}} \frac{\exp(s_g)}{\sum_{g'} \exp(s_{g'})} \cdot \hat{y}_g \tag{3}$$

where s_g is the attention score for pose g. The loss is the mean squared error between the complex-level prediction and the experimental affinity label.

2.5. Interaction-Aware Cross-Validation: PLEC-Butina Split

All downstream experiments use 10-fold cross-validation with our novel **PLEC-Butina** clustering, which is an adaptation of the Butina clustering (Butina, 1999) for protein-ligand extended connectivity fingerprints or PLEC (Wójcikowski et al., 2019) to partition the protein-ligand pairs into separate splits or folds such that there is no leakage in terms of the interaction motifs appearing across splits. A high PLEC similarity between two complexes would mean similar binding contacts, and we want to avoid that across splits.

For each protein-ligand complex, we compute a 65,536-bit PLEC fingerprint using the ODDT toolkit (Wójcikowski et al., 2015) with a distance cutoff of 5.0 Å. The PLEC encodes the local protein-ligand interaction environment at each contacting atom pair, producing a binary vector $\mathbf{f}_i \in \{0,1\}^{65536}$. Then, we measure the overlap of protein-ligand interaction patterns between all pairs of complexes using the Jaccard similarity on their PLEC bit vectors:

$$J(\mathbf{f}_i, \mathbf{f}_j) = \frac{|\mathbf{f}_i \cap \mathbf{f}_j|}{|\mathbf{f}_i \cup \mathbf{f}_j|} = \frac{\sum_b f_{i,b} \cdot f_{j,b}}{\sum_b f_{i,b} + \sum_b f_{j,b} - \sum_b f_{i,b} \cdot f_{j,b}} \tag{4}$$

yielding a symmetric similarity matrix $\mathbf{S} \in [0,1]^{N \times N}$. A high value $J(\mathbf{f}_i, \mathbf{f}_j) \approx 1$ indicates that compounds i and j are in contact with the same protein residue environments through the same types of local contacts.

We set a PLEC similarity threshold θ of 0.25. The Butina algorithm clusters complexes by interaction similarity in a greedy fashion. This produces M clusters $\{\mathcal{C}_1, \ldots, \mathcal{C}_M\}$ of varying sizes, where compounds within each cluster share similar protein-ligand interaction patterns.

3. Related Work

Structure-based scoring functions. Traditional scoring functions for binding affinity prediction fall into three categories: physics-based methods that compute force-field energies (Kollman et al., 2000), knowledge-based potentials derived from crystallographic statistics (Gohlke et al., 2000), and empirical scoring functions that fit weighted energy terms to experimental data (Wang et al., 2004).

Graph neural networks for molecular property prediction. GNNs have become the dominant paradigm for learning from molecular structure (Gilmer et al., 2017). For protein-ligand binding affinity, approaches range from atom-level message passing on contact graphs (Torng & Altman, 2019; Lim et al., 2019) to methods that incorporate interaction fingerprints (Wójcikowski et al., 2019). EIGN (Jiang et al., 2021) employs a GIN backbone (Xu et al., 2019) on protein-ligand contact graphs, while PLAIG integrates BINANA-derived interaction features (Durrant & McCammon, 2011) into a GNN framework. However, these methods represent the complex as a homogeneous atom-level graph and do not distinguish between different types of noncovalent interactions at the architectural level.

Self-supervised pretraining for molecules. Pretraining strategies for molecular representations include masked atom prediction, (Hu et al., 2020) contrastive learning, (Wang et al., 2022) and denoising objectives (Zaidi et al., 2023). Most approaches operate on two-dimensional molecular graphs or SMILES strings (Chithrananda et al., 2020). For protein-ligand interactions, pretraining on three-dimensional structures remains limited, with notable exceptions including EquiBind (Stärk et al., 2022) and Uni-Mol (Zhou et al., 2023). PLIP extends this line of work by pretraining on 5.1 million protein-ligand structures with a multi-task objective that simultaneously learns interaction types, distances, and affinities.

4. Experimental Setup

4.1. Datasets

SAIR pretraining corpus. The Structural and Interaction Repository (SAIR) contains 5.1 million protein-ligand complex structures spanning approximately 1,000 unique protein targets. Each complex is processed into a heterogeneous interaction graph as described in Section 2.1. Binding affinity labels are available for a subset of complexes and are used in the affinity pretraining objective. In SAIR, Boltz-1x is used to generate 5 structures (of varying confidence scores) for each unique protein-ligand pair from BindingDB (Liu et al., 2024) and ChEMBL35 (Zdrazil et al., 2023). About 1 million such unique pairs and their relevant bioactivity data were extracted. We perform pretraining on the following three subsets of SAIR:

- **All**: All structures in SAIR that pass every PoseBusters check. PoseBusters provides a reliable filter for physical validity, ensuring we pretrain only on geometrically plausible structures. This yields 5,177,259 structures.
- **Best**: From the **All** subset, we retain only the highest-confidence structure per protein-ligand complex, using the confidence scores produced by Boltz-1x. This se-

lects the most reliable prediction for each complex and results in 1,030,067 structures.

- **Best Positive**: From the **Best** subset, we further discard structures with $pIC_{50} < 6.0$, retaining only complexes that exhibit moderate binding affinity. This avoids training on weakly-binding complexes that Boltz-1x has nonetheless modelled as bound. The resulting subset contains 699,838 structures.

Downstream target datasets. We evaluate on four therapeutically relevant targets spanning a range of dataset sizes and biological contexts. These downstream target datasets represent a selected subset of our in-house virtual high-throughput screening benchmark effort built around curated experimental pIC_{50} labels, target-specific evaluation settings, and standardized structure-based inputs. Receptor structures were generated from protein sequences using AlphaFold3 (Abramson et al., 2024), and protein-ligand docked poses were subsequently prepared with an AutoDock Vina docking workflow (Eberhardt et al., 2021). The datasets are as follows:

- **Acetylcholinesterase (AChE)** is an enzyme critical for the termination of nerve impulse transmission at cholinergic synapses, and a primary target for Alzheimer's disease therapeutics (Soreq & Seidman, 2001). We process the AChE inhibition data curated by Vignaux et al. (2023).
- **SARS-CoV-2 main protease (SARS-M^{pro})** catalyzes the proteolytic processing of viral polyproteins essential for SARS-CoV-2 replication (Dai et al., 2020). The dataset contains 3,799 affinity data points (Shim et al., 2022). This target has been central to integrated computational-experimental discovery pipelines (Lau et al., 2021).
- **Zika protease** is a flaviviral NS2B-NS3 protease essential for Zika virus replication (Phoo et al., 2016). This dataset is notable for its five distinct protein structures (Table 1), which have slight conformational differences in NS2B positioning, providing some structural diversity (Braun et al., 2020; Ranganath et al., 2026).
- **μ-Opioid receptor** is a G protein-coupled receptor that mediates the analgesic and addictive effects of opioid drugs (Manglik et al., 2012). The dataset contains 606 affinity data points (Shim et al., 2024), making it the smallest of the four datasets.

4.2. Baselines

We compare PLIP with two competing baselines:

Table 1. The four downstream datasets

Dataset	#Affinity datapoints	Target PDB IDs
AChE	7,108	4EY4, 6O4W
μ-Opioid	606	8EF5, 8K9K
SARS-M^{pro}	3,799	6LU7, 6Y84
Zika	1,976	6KK4, 6Y3B, 7M1V, 7ZPD, 7ZYS

EIGN (Yang et al., 2025) uses a Graph Isomorphism Network (GIN) (Xu et al., 2019) backbone operating on protein-ligand contact graphs with atom-level nodes. It does not distinguish interaction types at the architectural level.

PLAIG (Samudrala et al., 2025) augments a GNN architecture with BINANA-derived (Durrant & McCammon, 2011) interaction features, incorporating handcrafted descriptors of hydrogen bonds, hydrophobic contacts, and π interactions as additional edge or graph-level features.

Both baselines are trained and evaluated on each downstream dataset using the same cross-validation protocol as PLIP.

4.3. Implementation Details

PLIP is implemented in PyTorch (Paszke et al., 2019) using PyTorch Geometric (Fey & Lenssen, 2019) for heterogeneous graph operations and PyTorch Lightning for training orchestration with distributed data parallel (DDP) training on a single node of four AMD MI300X GPUs. Pretraining on SAIR uses hidden dimension 256, 6 transformer layers, 8 attention heads, dropout 0.01, batch size 128, learning rate 10^{-4}, and weight decay 10^{-5} for up to 500 epochs. Downstream fine-tuning uses hidden dimension 256, 6 layers, 4 heads, dropout 0.1, batch size 2, and learning rate 10^{-2} with weight decay 10^{-4} for up to 200 epochs. All experiments use the AdamW optimizer.

SAIR pretraining. Pretraining uses hidden dimension 256, 6 transformer layers, 8 attention heads, dropout 0.01, batch size 128, and weight decay 10^{-5} for up to 500 epochs. Each prediction head receives its own learning rate: backbone 3×10^{-5}, gate head 4×10^{-4}, interaction-type head 2×10^{-4}, distance head 1×10^{-5}, and affinity head 2×10^{-4} (hyperparameters were tuned using Optuna (Akiba et al., 2019)). As established in section 4.1, we pretrain on three SAIR subsets of increasing selectivity: *all*, *best*, and *best-positive*. Each subset is pretrained on three target combinations: interaction type; interaction type + distance; interaction type + distance + affinity, yielding nine pretrained checkpoints that are all evaluated downstream. These 9

checkpoints allow us to analyse the effects of varying the quality of the structures used in pretraining and various pretraining targets.

Transfer learning When transferring from a pretrained checkpoint, only the backbone weights (type-specific embeddings, transformer layers, and layer norms) are loaded into the downstream model; the affinity prediction head is always initialized from scratch. In scratch mode, no pretrained checkpoint is loaded. Dataset-specific and mode-specific hyperparameters (number of layers, heads, dropout, batch size, learning rate, weight decay) are additionally tuned with Optuna.

4.4. Verifying data leakage between pretrain and downstream datasets

Given the possibility that pretraining structures could have binding-site contacts similar to downstream complexes, there may be room for implicit information transfer through learned interaction representations. To quantify the risk of data leakage from the pre-training phase to the fine-tuning phase, we performed a cross-dataset similarity analysis of 65,536-bit PLEC fingerprints between SAIR and the four downstream datasets.

We uniformly sampled 5000 protein-ligand complexes from the SAIR pretraining corpus, and computed the PLEC fingerprints using the ODDT toolkit (Wójcikowski et al., 2015) (65,536 bits, 5.0 Å distance cutoff, hydrogen atoms added). For each downstream dataset, we computed PLEC fingerprints for 500 randomly sampled complexes.

We then computed the full pairwise Jaccard similarity matrix $\mathbf{J} \in \mathbb{R}^{500\times500}$ between the SAIR and downstream PLEC vectors using Equation 4. For each SAIR structure, we recorded its nearest-neighbor similarity to any downstream complex: $\mathrm{NN}_i = \max_j J(\mathbf{f}_i^{\mathrm{SAIR}}, \mathbf{f}_j^{\mathrm{downstream}})$. Mean pairwise Jaccard similarities range from 0.024 (μ-opioid) to 0.036 (SARS-M$^{\mathrm{pro}}$). This negligible overlap demonstrates that the SAIR pretraining corpus and downstream datasets occupy essentially disjoint regions of protein-ligand interaction fingerprint space.

5. Results and Discussion

We aim to understand if PLIP performs better than scratch which is the first baseline it should outperform and eventually we compare it to competing methods from previous work, namely, EIGN and PLAIG.

5.1. Pretraining benefits over scratch

Figure 3 compares PLIP scratch and finetuning across three train set sizes. Several patterns emerge. First, pretraining consistently improves performance over-training from scratch across all four datasets, with the largest absolute gain in terms of Pearson correlation (r) on AChE ($\Delta r = +0.082$, a 12% relative improvement from scratch $r = 0.676$ to pretrained $r = 0.758$). A distinctive feature of our experimental design is the evaluation of three downstream train sizes-1/3, 2/3, and Full-allowing us to assess how the volume of training data affects downstream performance (Figure 3). For AChE, increasing train data produces a clear monotonic improvement: Pearson r rises from 0.667 (1/3 SAIR) to 0.724 (2/3) to 0.758 (Full), a total gain of $+0.091$ ($+14\%$). Zika exhibits a similar monotonic pattern ($0.491 \rightarrow 0.514 \rightarrow 0.530$), as does SARS-M$^{\mathrm{pro}}$ frozen ($0.251 \rightarrow 0.271 \rightarrow 0.281$). μ-Opioid is the exception: performance is non-monotonic ($0.374 \rightarrow 0.352 \rightarrow 0.360$), with the 1/3 subset yielding the highest point estimate. The small size of this dataset (606 complexes) likely introduces higher variance. In all cases, finetuning is better than scratch, suggesting that PLIP is effective in boosting the results regardless of the downstream train sizes. Also, the improvements for scratch across different train sizes are larger compared to improvements for finetuning across train sizes, which suggests that pretraining helps reduce the sensitivity to the amount of downstream train set sizes.

Figure 4 provides a per-checkpoint breakdown across all nine checkpoints and three downstream train data scales, showing that the multi-task (interaction type + distance +affinity or I+D+A) checkpoint on **All** consistently ranks among the top configurations. On AChE, the full combination achieves $r = 0.758$ with finetune mode, compared to lower values for partial combinations, demonstrating that including the affinity prediction objective during pretraining provides the most significant boost to downstream binding affinity performance. For frozen fine-tuning, which isolates the quality of the pretrained backbone, the full multi-task checkpoint (All/I+D+A) is optimal on AChE, Zika, and SARS-M$^{\mathrm{pro}}$, while μ-opioid favors the simpler interaction type only checkpoint, possibly because distance and affinity objectives are less informative for the limited receptor conformational diversity of this small dataset.

5.2. Diversity of pretraining set

As noted, we used three pretraining subsets from SAIR: **All**, **Best** and **Best Positive**. In Figure 4, it can be seen that the pretraining label combination matters more than the pretraining subset used. Interaction type, distance, and affinity (I+D+A), regardless of the pretrain subset, result in the best performance except for μ-opioid. μ-opioid seems to benefit the most from pretraining on just the interaction type and significantly worse performance with checkpoints trained on I+D+A. Given that μ-opioid is the smallest dataset out of the four, pretraining on affinities from SAIR, which could have vastly different proteins and ligands from μ-opioid, may have affected that checkpoint's ability from modelling

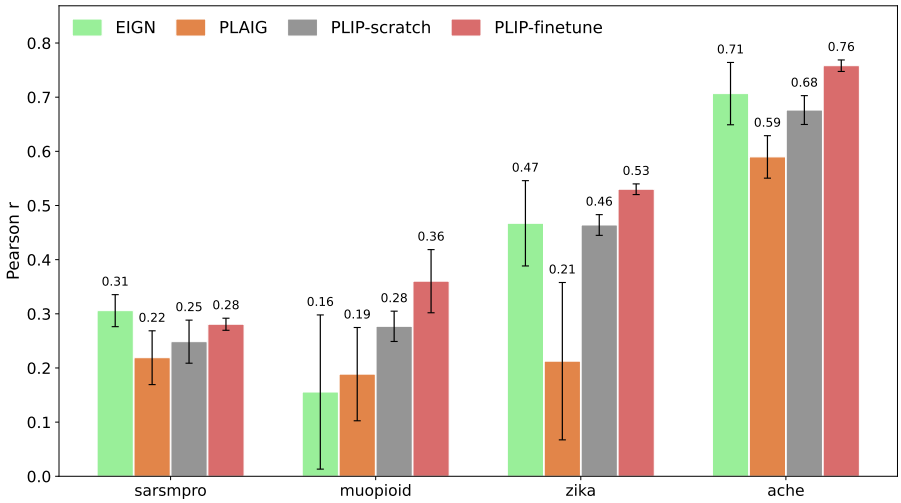

Figure 2. Pearson correlations of EIGN, PLAIG, PLIP-scratch, and PLIP-finetuned. The PLIP-finetune results shown here are from the All (interaction type, distance, affinity) pretrained checkpoint.

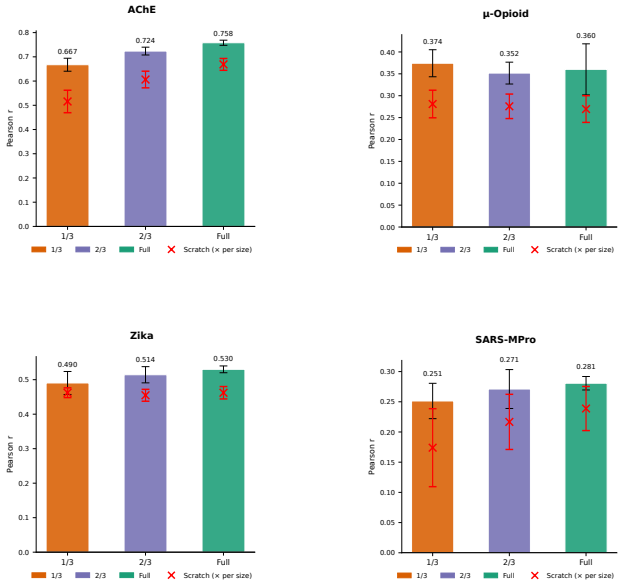

Figure 3. Comparison of PLIP-finetuned (bar plots) and PLIP-scratch (red crosses) on train set sizes scaled from a third, to two-thirds, to the entire train set for the downstream datasets

the affinities in the μ-opioid dataset.

Across, the three pretraining subsets, there is considerably smaller differences in performance for the same pretraining target combination. Thus, a key insight we may gather from this is that the number of structures do not matter as much as what the models learn from them. Therefore, given a number of structures, more emphasis could be placed on extracting physically rich and accurate properties such as interaction type, the interaction energy, and even the bond critical points to pretrain on.

5.3. Comparison with Baselines

Figure 2 compares PLIP against the EIGN and PLAIG baselines across all four downstream datasets. PLIP's explicit modeling of ring nodes and interaction-typed edges provides a biologically motivated inductive bias that homogeneous atom-level graphs lack. Aromatic rings are central to many drug-target interactions: π-π stacking, cation-π, and π-metal contacts are thermodynamically significant yet geometrically complex interactions that cannot be fully captured by pairwise atom-level messages alone (Hunter & Sanders, 1990; Ma & Dougherty, 1997). The advantage of this representation is evident in the comparison with PLAIG, which uses handcrafted BINANA interaction features on a homogeneous graph: PLIP outperforms PLAIG by 28% on AChE and by 149% on Zika, suggesting that end-to-end learning of interaction-typed edges within a heterogeneous architecture is more effective than augmenting a flat graph with static interaction descriptors.

SARS-M$^{\text{pro}}$ is the only dataset on which EIGN ($r = 0.306$) outperforms PLIP ($r = 0.281$). This target has the narrowest pIC_{50} range among the four datasets (4.00-8.57), leaving minimal dynamic range for a correlation-based metric. In this compressed regime, EIGN's simpler GIN backbone may learn a more parsimonious mapping that better exploits the limited affinity variance. Additionally, the rapid pace of SARS-CoV-2 drug discovery efforts may have introduced heterogeneous data quality in the curated experimental measurements. We note, however, that PLIP exhibits substantially lower fold-to-fold variance (standard deviation of 0.011 vs. 0.030), suggesting that its predictive performance are more stable even if slightly lower on average.

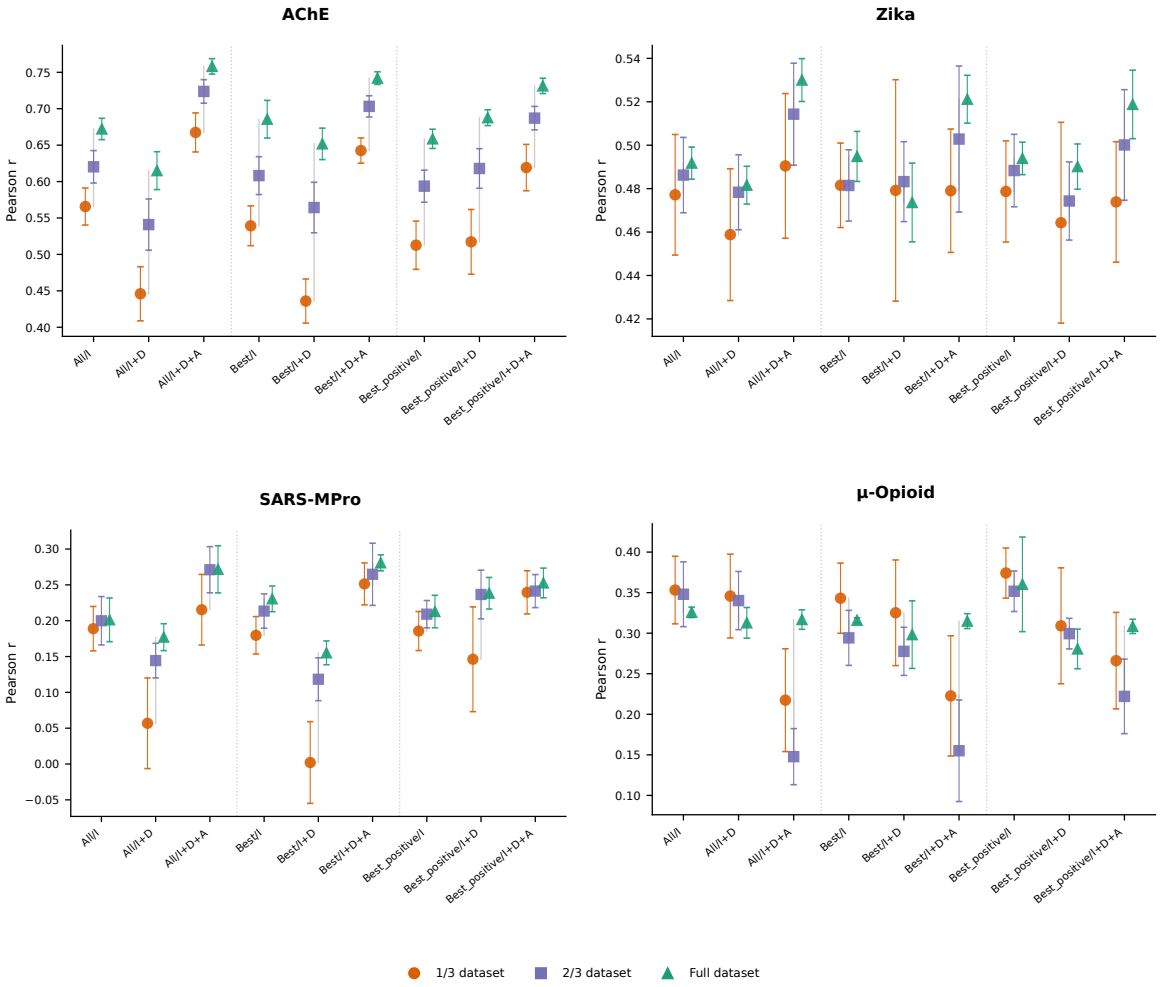

Figure 4. Scaling the train set sizes from a third, to two-thirds, to the entire train set for the downstream datasets and comparing to the scratch performance across all pretraining regimes.

6. Conclusion

We have introduced PLIP, a heterogeneous graph transformer for protein-ligand binding affinity prediction that addresses three key limitations of existing GNN-based approaches: the lack of explicit ring-level representations, the absence of large-scale three-dimensional pretraining, and the use of element-identity-only atom features. Through systematic evaluation on four drug-discovery-relevant targets, we demonstrate that (1) heterogeneous graphs with explicit ring nodes and interaction-typed edges outperform homogeneous atom-level graphs, surpassing EIGN and PLAIG baselines on three of four targets; (2) multi-task self-supervised pretraining on 5.1 million SAIR structures consistently improves downstream binding affinity prediction; and (3) a principled approach to splitting protein-ligand complexes for binding affinity prediction by clustering PLEC fingerprints using Butina clustering. PLIP achieves Pearson correlations of 0.758 on AChE, 0.530 on Zika, 0.360 on μ-opioid, and 0.281 on SARS-M$^{\text{pro}}$, with cross-fold standard deviations as low as 0.010, offering a principled and scalable framework for structure-based drug discovery.

Future work could consist of (1) scaling the pretraining corpus beyond the current 5.1 million structures by ensuring both quality and diversity, (2) applying active learning strategies to efficiently expand small downstream datasets such as μ-opioid, and (3) evaluating out-of-distribution generalization on held-out protein families.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu,

Z., Zemgulyte, A., Arvaniti, E., Beattie, C., Bertolli, O., Bridgland, A., Cherepanov, A., Congreve, M., Cowen-Rivers, A. I., Cowie, A., Figurnov, M., Fuchs, F. B., Gladman, H., Jain, R., Khan, Y. A., Low, C. M. R., Perlin, K., Potapenko, A., Savy, P., Singh, S., Stecula, A., Thillaisundaram, A., Tong, C., Yakneen, S., Zhong, E. D., Zielinski, M., Žídek, A., tor Bapst, V.-., Kohli, P., Jaderberg, M., Hassabis, D., and Jumper, J. M. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630:493 – 500, 2024. URL https://api.semanticscholar.org/CorpusID:269633210.

Ahdritz, G., Bouatta, N., Floristean, C., Kadyan, S., Xia, Q., Gerecke, W., O'Donnell, T., Berenberg, D., Fisk, I., Zanichelli, N., Zhang, B., Nowaczynski, A., Wang, B., Stepniewska-Dziubinska, M., Zhang, S., Ojewole, A., Guney, M., Biderman, S., Watkins, A., and Alquraishi, M. Openfold: retraining alphafold2 yields new insights into its learning mechanisms and capacity for generalization. *Nature Methods*, 21:1–11, 05 2024. doi: 10.1038/s41592-024-02272-z.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2623–2631, 2019.

Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *Journal of Chemical Physics*, 134:074106, 2011.

Braun, N. J., Quek, J. P., Huber, S., Kouretova, J., Rogge, D., Lang-Henkel, H., Cheong, E. Z. K., Chew, B. L. A., Heine, A., Luo, D., and Steinmetzer, T. Structure-based macrocyclization of substrate analogue ns2b-ns3 protease inhibitors of zika, west nile and dengue viruses. *ChemMedChem*, 15(15):1439–1452, 2020. doi: https://doi.org/10.1002/cmdc.202000237. URL https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cmdc.202000237.

Butina, D. Unsupervised data base clustering based on Daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences*, 39(4):747–750, 1999.

Chithrananda, S., Grand, G., and Ramsundar, B. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Dai, W., Zhang, B., Jiang, X.-M., Su, H., Li, J., Zhao, Y., Xie, X., Jin, Z., Peng, J., Liu, F., Li, C., Li, Y., Bai, F., Wang, H., Cheng, X., Cen, X., Hu, S., Yang, X., Wang, J., Liu, X., Xiao, G., Jiang, H., Rao, Z., Zhang, L.-K., Xu, Y., Yang, H., and Liu, H. Structure-based design of antiviral drug candidates targeting the sars-cov-2 main protease. *Science*, 368(6497):1331–1335, 2020. doi: 10.1126/science.abb4489. URL https://www.science.org/doi/abs/10.1126/science.abb4489.

Durrant, J. D. and McCammon, J. A. BINANA: A novel algorithm for ligand-binding characterization. *Journal of Molecular Graphics and Modelling*, 29(6):888–893, 2011.

Eberhardt, J., Santos-Martins, D., Tillack, A. F., and Forli, S. Autodock vina 1.2.0: New docking methods, expanded force field, and python bindings. *Journal of chemical information and modeling*, 2021. URL https://api.semanticscholar.org/CorpusID:236092162.

Fey, M. and Lenssen, J. E. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. pp. 1263–1272, 2017.

Gohlke, H., Hendlich, M., and Klebe, G. Knowledge-based scoring function to predict protein–ligand interactions. *Journal of Molecular Biology*, 295(2):337–356, 2000.

Gorantla, R., Gema, A. P., Yang, I. X., Serrano-Morrás, Á., Suutari, B., Juárez-Jiménez, J., and Mey, A. S. Learning binding affinities via fine-tuning of protein and ligand language models. *Journal of Chemical Information and Modeling*, 65(22):12279–12291, 2025.

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2020.

Hunter, C. A. and Sanders, J. K. The nature of π–π interactions. *Journal of the American Chemical Society*, 112 (14):5525–5534, 1990.

Jiang, D., Hsieh, C.-Y., Wu, Z., Kang, Y., Wang, J., Wang, E., Liao, B., Shen, C., Xu, L., Wu, J., et al. InteractionGraphNet: A novel and efficient deep graph representation learning framework for accurate protein–ligand interaction predictions. *Journal of Medicinal Chemistry*, 64(24):18209–18232, 2021.

Jumper, J. M., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M.,

Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583 – 589, 2021. URL https://api.semanticscholar.org/CorpusID:235959867.

Kim, H., Shim, H., Ranganath, A., He, S., Stevenson, G., and Allen, J. E. Protein-ligand binding affinity prediction using multi-instance learning with docking structures. *Frontiers in Pharmacology*, 15, 2025. doi: 10.3389/fphar.2024.1518875.

Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., et al. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Accounts of Chemical Research*, 33(12):889–897, 2000.

Lau, E. Y., Negrete, O. A., Bennett, W. F. D., Bennion, B. J., Borucki, M., Bourguet, F., Epstein, A., Franco, M., Harmon, B., He, S., Jones, D., Kim, H., Kirshner, D., Lao, V., Lo, J., McLoughlin, K., Mosesso, R., Murugesh, D. K., Saada, E. A., Segelke, B., Stefan, M. A., Stevenson, G. A., Torres, M. W., Weilhammer, D. R., Wong, S., Yang, Y., Zemla, A., Zhang, X., Zhu, F., Allen, J. E., and Lightstone, F. C. Discovery of small-molecule inhibitors of sars-cov-2 proteins using a computational and experimental pipeline. *Frontiers in Molecular Biosciences*, 8, 2021. doi: 10.3389/fmolb.2021.678701.

Lemos, P., Beckwith, Z., Bandi, S., Damme, M. V., Crivelli-Decker, J., Shields, B. J., Merth, T., Jha, P. K., Mitri, N. D., Callahan, T., Nish, A., Abruzzo, P., Salomon-Ferrer, R., and Ganahl, M. SAIR: Enabling deep learning for protein-ligand interactions with a synthetic structural dataset. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=qgk2F6jxH4.

Li, J. and Gong, X. Harnessing pre-trained models for accurate prediction of protein-ligand binding affinity. *BMC Bioinformatics*, 26, 02 2025. doi: 10.1186/s12859-025-06064-w.

Li, Y., Rezaei, M. A., Li, C., Li, X., and Wu, D. O. Deepatom: A framework for protein-ligand binding affinity prediction. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 303–310, 2019. URL https://api.semanticscholar.org/CorpusID:208527531.

Lim, J., Ryu, S., Park, K., Choe, Y. J., Ham, J., and Kim, W. Y. Predicting drug–target interaction using a novel graph neural network with 3D structure-embedded graph representation. *Journal of Chemical Information and Modeling*, 59(9):3981–3988, 2019.

Liu, T., Hwang, L., Burley, S. K., Nitsche, C. I., Southan, C., Walters, W. P., and Gilson, M. K. Bindingdb in 2024: a fair knowledgebase of protein-small molecule binding data. *Nucleic Acids Research*, 53:D1633 – D1644, 2024. URL https://api.semanticscholar.org/CorpusID:274177047.

Ma, J. C. and Dougherty, D. A. The cation–π interaction. *Chemical Reviews*, 97(5):1303–1324, 1997.

Manglik, A., Kruse, A. C., Kobilka, T. S., Thian, F. S., Mathiesen, J. M., Sunahara, R. K., Pardo, L., Weis, W. I., Kobilka, B. K., and Granier, S. Crystal structure of the μ-opioid receptor bound to a morphinan antagonist. *Nature*, 485:321 – 326, 2012. URL https://api.semanticscholar.org/CorpusID:4371729.

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3:33, 2011.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Phoo, W., Li, Y., Zhang, Z., Lee, M., Loh, Y., Tan, Y. B., Ng, E., Lescar, J., Kang, C., and Luo, D. Structure of the ns2b-ns3 protease from zika virus after self-cleavage. *Nature Communications*, 7:13410, 11 2016. doi: 10.1038/ncomms13410.

Ranganath, A., Kim, H., Shim, H., and Allen, J. E. Slab: simultaneous labeling and binding affinity prediction for protein–ligand structures. *Digital Discovery*, 5:375–383, 2026. doi: 10.1039/D5DD00248F. URL http://dx.doi.org/10.1039/D5DD00248F.

Samudrala, M., Dandibhotla, S., Kaneriya, A., and Dakshanamurthy, S. Plaig: Protein–ligand binding affinity prediction using a novel interaction-based graph neural network framework. *ACS Bio Med Chem Au*, 5, 04 2025. doi: 10.1021/acsbiomedchemau.5c00053.

Shim, H., Kim, H., Allen, J. E., and Wulff, H. Pose classification using three-dimensional atomic structure-based neural networks applied to ion channel–ligand docking. *Journal of Chemical Information and Modeling*, 62(10):2301–2315, 2022. doi: 10.1021/acs.jcim.1c01510.

Shim, H., Allen, J. E., and Bennett, W. F. D. Enhancing docking accuracy with pecan2, a 3d atomic neural network trained without co-complex crystal

structures. *Mach. Learn. Knowl. Extr.*, 6:642–657, 2024. URL https://api.semanticscholar.org/CorpusID:268490409.

Smith, J., Isayev, O., and Roitberg, A. Ani-1: An extensible neural network potential with dft accuracy at force field computational cost. *Chem. Sci.*, 8, 02 2017a. doi: 10.1039/C6SC05720A.

Smith, J. S., Isayev, O., and Roitberg, A. E. ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost. *Chemical Science*, 8(4): 3192–3203, 2017b.

Soreq, H. and Seidman, S. Acetylcholinesterase — new roles for an old actor. *Nature Reviews Neuroscience*, 2:294–302, 2001. URL https://api.semanticscholar.org/CorpusID:5947744.

Stärk, H., Ganea, O.-E., Pattanaik, L., Barzilay, R., and Jaakkola, T. EquiBind: Geometric deep learning for drug binding structure prediction. In *International Conference on Machine Learning*, pp. 20503–20521. PMLR, 2022.

Torng, W. and Altman, R. B. Graph convolutional neural networks for predicting drug-target interactions. *Journal of Chemical Information and Modeling*, 59(10):4131–4149, 2019.

Valsson, , Warren, M., Deane, C., Magarkar, A., Morris, G., and Biggin, P. Narrowing the gap between machine learning scoring functions and free energy perturbation using augmented data. *Communications Chemistry*, 8, 02 2025. doi: 10.1038/s42004-025-01428-y.

Vignaux, P. A., Lane, T. R., Urbina, F., Gerlach, J., Puhl, A. C., Snyder, S. H., and Ekins, S. Validation of acetylcholinesterase inhibition machine learning models for multiple species. *Chemical Research in Toxicology*, 36:188 – 201, 2023. URL https://api.semanticscholar.org/CorpusID:256577383.

Volkov, M., Turk, J.-A., Drizard, N., Martin, N., Hoffmann, B., Gaston-Mathé, Y., and Rognan, D. On the frustration to predict binding affinities from protein–ligand structures with deep neural networks. *Journal of Medicinal Chemistry*, 65(11):7946–7958, 2022.

Wang, R., Lai, L., and Wang, S. Further development and validation of empirical scoring functions for structure-based binding free energy calculation. *Journal of Computer-Aided Molecular Design*, 18(2):55–83, 2004.

Wang, Y., Wang, J., Cao, Z., and Farimani, A. B. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4:279–287, 2022.

Wohlwend, J., Corso, G., Passaro, S., Reveiz, M., Leidal, K., Swiderski, W., Portnoi, T., Chinn, I., Silterra, J., Jaakkola, T., and Barzilay, R. Boltz-1 democratizing biomolecular interaction modeling. *bioRxiv*, 2024. URL https://api.semanticscholar.org/CorpusID:274166333.

Wójcikowski, M., Zielenkiewicz, P., and Siedlecki, P. Open Drug Discovery Toolkit (ODDT): A new open-source player in the drug discovery field. *Journal of Cheminformatics*, 7:26, 2015.

Wójcikowski, M., Zielenkiewicz, P., and Siedlecki, P. Open drug discovery toolkit (oddt): A new open-source player in the drug discovery field. *Journal of Cheminformatics*, 7, 12 2015. doi: 10.1186/s13321-015-0078-2.

Wójcikowski, M., Kukiełka, M., Stepniewska-Dziubinska, M. M., and Siedlecki, P. Development of a protein–ligand extended connectivity (plec) fingerprint and its application for binding affinity predictions. *Bioinformatics*, 35(8):1334–1341, 04 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty757. URL https://doi.org/10.1093/bioinformatics/bty757.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Yang, D., Kuang, L., and Hu, A. Edge-enhanced interaction graph network for protein-ligand binding affinity prediction. *PLOS ONE*, 20(4):1–18, 04 2025. doi: 10.1371/journal.pone.0320465. URL https://doi.org/10.1371/journal.pone.0320465.

Yi, D., Zhao, Y., Xu, H., Zhang, Y., Wan, M., Zan, P., He, S., and Bo, X.-C. Compbind: Complex guided pretraining-based structure-free protein-ligand affinity prediction. *Journal of chemical information and modeling*, 66, 01 2026. doi: 10.1021/acs.jcim.5c02451.

Zaidi, S., Schaarschmidt, M., Martens, J., Kim, H., Teh, Y. W., Sanchez-Gonzalez, A., Battaglia, P., Pascanu, R., and Godwin, J. Pre-training via denoising for molecular property prediction. *arXiv preprint arXiv:2206.00133*, 2023.

Zdrazil, B., Felix, E., Hunter, F., Manners, E. J., Blackshaw, J., Corbett, S., de Veij, M., Ioannidis, H., Lopez, D. M., Mosquera, J. F. M., Magariños, M. P., Bosc, N., Arcila, R., Kizilören, T., Gaulton, A., Bento, A. P., Adasme, M. F., Monecke, P., Landrum, G. A., and Leach, A. R. The chembl database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, 52:D1180 – D1192, 2023. URL https://api.semanticscholar.org/CorpusID:265041119.

Zhou, G., Gao, Z., Ding, Q., Zheng, H., Xu, H., Wei, Z., Zhang, L., and Ke, G. Uni-Mol: A universal 3D molecular representation learning framework. *ChemRxiv*, 2023.