

Appendix content

A	Technical appendices and supplementary material	2
A.1	Implementation	2
A.1.1	Bioactivity prediction framework	2
A.1.2	Graph representation	2
A.1.3	Detailed experimental settings	2
A.1.4	Implementation of initial individual pockets	3
A.1.5	Support of using multiple complex conformations	3
A.1.6	Computational environment	3
A.1.7	Computational time of the geometric edge	4
A.2	Baseline models	4
A.2.1	DTIGN model	4
A.2.2	GIGN model	4
A.2.3	GAT model	4
A.2.4	EHIGN model	5
A.2.5	SIGN model	5
A.2.6	MBP model	5
A.3	Detail of datasets	6
A.3.1	DTIGN dataset	6
A.3.2	SIU dataset	6
A.4	Additional experimental results	6
A.4.1	Evaluation on additional baseline models	6
A.4.2	Evaluation on ligand binding affinity prediction task	7
A.4.3	The impact of different aggregation strategies	9
A.5	Explanation of the relationship between geometric edge and empty space	10
A.6	Evaluation metrics	10
A.7	Comparison with Uni-Mol	11
A.8	Compare with the pairwise loss in other works	11
A.9	Analysis of using docking conformations	12
A.10	Limitations	12

A Technical appendices and supplementary material

A.1 Implementation

A.1.1 Bioactivity prediction framework

This section describes how the proposed geometric representation, Union-Pocket strategy, and pairwise loss are incorporated into the GNN models evaluated in this work.

The geometric representation characterizes the local spatial emptiness around each atom in the protein–ligand graph. It is implemented by introducing additional edges between atoms that are geometrically adjacent—specifically, the nearest neighbors located within angular cones defined in spherical coordinates. This augmentation can be seamlessly integrated into any edge-based GNN architecture.

The Union-Pocket approach constructs a unified binding pocket by taking the union of all local pockets sampled from multiple ligand docking poses on the same protein. This expanded pocket is added to each data point by including previously missing atoms, and is compatible with GNNs that operate on either node or edge features. During the feature aggregation stage for generating the final global graph representation, it is recommended to retain only the small pocket region, namely the pocket atoms in proximity to the ligand, as an excessively large Union-Pocket may obscure the features of the core interaction site.

The pairwise loss is formulated to penalize incorrect ordering between sample pairs in the context of real-valued bioactivity prediction. It is designed to complement standard regression loss functions and can be jointly optimized during GNN training to improve the model’s ability to capture the relative ranking of bioactivities across samples.

A.1.2 Graph representation

The chemical node and edge features adopted in our application are taken from RDKit [RDKit Team]. The details are shown in Table 1 and Table 2 below. Note that for the geometric edge, we use Coulomb and London dispersion forces as features based on interatomic distances, as these intermolecular interactions are generally applicable across all molecular pairs.

Table 1: The information of node features

Node features	types	sizes
Atom type	one-hot	10
Degree	integer	1
Implicit valence	one-hot	7
Hybridization	one-hot	5
Aromatic	binary	1
Hydrogens	integer	5

A.1.3 Detailed experimental settings

For geometric calculation, we uniformly partition the surrounding space into S cones for each atom. Here, we set $S = 32$ in our implementation. This is achieved by dividing both the azimuthal and polar angles in spherical coordinates at intervals of $\frac{\pi}{4}$. This angular partitioning allows the surrounding space of each atom to be discretized into uniform sectors for geometric feature extraction.

We train the models on the training set with early stopping based on validation performance, and report the final results on the test set. In the hybrid loss function

$$\mathcal{L}_{\text{hybrid}} = \lambda_1 \cdot \mathcal{L}_2 + \lambda_2 \cdot \mathcal{L}_{\text{pairwise}}, \quad (1)$$

We adopt $\lambda_1 = 0.8$ and $\lambda_2 = 0.2$ in our training.

Table 2: The information of edge features

Intra-molecular edge features	types	sizes
bond type	one-hot	4
conjugation	binary	1
ring	binary	1
stereo	one-hot	4
Intermolecular edge features	types	sizes
Distance (d) in Coulomb force	real-valued vector	64
Distance (d) in London dispersion forces	real-valued vector	64
Geometric edge features	types	sizes
Distance (d) in Coulomb force	real-valued vector	64
Distance (d) in London dispersion forces	real-valued vector	64

A.1.4 Implementation of initial individual pockets

Individual pockets are defined as the set of protein residues with at least one atom located within 5Å of any ligand atom [Yang et al., 2023, Yin et al., 2024, Gagliardi and Rocchia, 2023]. This definition follows a commonly used proximity-based criterion that reflects potential physical interactions.

The Union-Pocket strategy introduces two key types of new information beyond simply aggregating existing docking poses. First, instead of using one large docking box that covers the entire protein surface, we define multiple smaller docking boxes around each cluster of functional sites. This approach forces ligands to explore and generate binding poses specifically within these localized regions, reducing sampling bias and ensuring more thorough and uniform coverage of all potential binding sites. Second, by taking the union of all such individual pockets, the resulting Union-Pocket–ligand graph captures the global spatial context of the ligand, including its relative location and the overall geometry of empty regions around potential binding sites. This holistic view of binding topography is not available in any single pose or local pocket, thereby providing richer structural information for bioactivity prediction.

Additionally, we would like to illustrate the Union-Pocket adopted in SIU dataset. As shown in Figure 2c (Page 4) of SIU paper [Huang et al., 2025], the SIU dataset contains a total of 249,631 pocket-ligand pairs across 9,544 protein pockets, indicating an average of approximately 26 ligands per pocket. Although it is claimed in the "Pocket definition" section (Page 19) [Huang et al., 2025] that "For each PDB ID, a single pocket was extracted, defined as the region centered on the co-crystal ligand within a 15 Å radius", the datasets they released via their Hugging Face repository (linked from their GitHub page, which is referenced in the paper’s abstract) contain only local pockets surrounding the individual ligand poses. In other words, the single pocket they used for docking was not fully utilized when constructing the pocket–ligand graphs in their released dataset. In our implementation, for each protein conformation, we use the predefined global pocket–docked by all its associated ligands—as the Union-Pocket.

A.1.5 Support of using multiple complex conformations

In our current implementation, multiple complex conformations are explicitly modeled by processing each conformation independently through the graph neural network to obtain individual graph-level embeddings. These embeddings are then aggregated (using either mean or attention-based pooling) before being passed to the bioactivity prediction network (typically a multi-layer perceptron). In this way, our method supports multiple complex conformations as input while leveraging shared network weights to process each conformation consistently.

A.1.6 Computational environment

Most experiments are run on an NVIDIA GeForce RTX 4090 GPU with 24GB of memory. For the DTIGN dataset, the experiments on the subsets containing more than 2000 atoms within the Union-Pocket are run on an NVIDIA H100 NVL GPU with 94GB of memory. For the SIU dataset,

the SIU-0.9- K_i version is run on 4 NVIDIA H100 NVL GPUs with 94GB of memory, and other versions are run on 4 NVIDIA GeForce RTX 4090 GPUs with 24GB of memory.

A.1.7 Computational time of the geometric edge

We report the computational time required to calculate the proposed geometric edge using the SIU-0.9- K_d dataset as an example. The results presented are based on single-CPU execution time; in practice, we parallelize the computation using 128 CPUs to significantly accelerate the process. The CPU model is AMD EPYC-Milan Processor. On average, for a pocket-ligand complex containing approximately 1,100 atoms, the geometric edge calculation takes 4.7 seconds per complex on a single CPU. This runtime is considered reasonable and acceptable for practical applications.

A.2 Baseline models

A.2.1 DTIGN model

The Drug-Target Interaction Graph Neural Network (DTIGN) model, originally proposed in Yin et al. [2024], addresses the challenge of multiple possible ligand poses binding to the same protein. It takes as input multiple docking poses corresponding to a single protein-ligand pair and employs a sophisticated GNN architecture combined with a multi-head self-attention mechanism to process and aggregate inter-molecular graph features for predicting bioactivity.

Distinct from earlier models for bioactivity or binding affinity prediction, DTIGN introduces several novel components. These include intramolecular bond features, as well as newly defined inter-molecular interactions based on high-order negative powers of Euclidean distances—modeled after Coulombic and London dispersion forces—to better capture physical force field characteristics indicative of binding stability [Ding et al., 1992].

Furthermore, DTIGN is the first to integrate multi-head self-attention and semi-supervised learning into this context, enabling the discovery of drug-target interaction patterns from software-generated docking poses. This combination also enhances the model’s ability to focus attention on native-like structural features during training.

To assess the performance of the DTIGN model, the authors constructed a unique benchmark dataset (the DTIGN dataset used in this study) comprising protein-ligand complexes derived from molecular docking simulations. This dataset was designed to evaluate bioactivity prediction models in the context of protein-ligand interactions. DTIGN’s performance was compared against nine leading deep learning-based bioactivity prediction methods. The results demonstrated that DTIGN achieved superior performance, with an average improvement of 27.03% over the baseline models.

A.2.2 GIGN model

The Geometric Interaction Graph Neural Network (GIGN) model, introduced in [Yang et al., 2023], is a GNN architecture designed to predict protein-ligand binding affinities directly from 3D molecular structures. GIGN constructs a joint geometric interaction graph that integrates both the atomic structures of the ligand and the protein pocket, capturing intra- and inter-molecular spatial relationships in a unified representation.

A key innovation of GIGN is its explicit use of 3D spatial information through relative position encodings and distance-aware message passing. The model defines two types of graph edges: covalent edges for intra-molecular connectivity and non-covalent interaction edges based on spatial proximity across molecules. These are encoded with distances to inform the network of interaction strength.

The GIGN model was shown to outperform several baselines on PDBbind 2016 (<http://www.pdbbind.org.cn/>) and CSAR-HiQ [Dunbar Jr et al., 2011] datasets, demonstrating its ability to capture subtle geometric cues relevant for binding affinity prediction.

A.2.3 GAT model

The Graph Attention Network (GAT), proposed in Veličković et al. [2017], is a graph neural network architecture that incorporates self-attention mechanisms to learn node representations by attending over their neighbors. Unlike traditional Graph Convolutional Networks (GCNs) that apply fixed

weighting schemes, GAT introduces trainable attention coefficients, allowing the model to adaptively weigh the importance of neighboring nodes during message passing.

In our context, GAT is applied to molecular graphs for the task of bioactivity prediction. Atoms are represented as graph nodes, and covalent bonds as edges. The model computes attention scores between atom pairs within each molecule, enabling it to focus on chemically or functionally relevant substructures. Multiple attention heads are used to capture diverse interaction patterns, and their outputs are either concatenated or averaged to form enriched node embeddings.

While GAT does not explicitly incorporate 3D geometric information or inter-molecular context (as done in DTIGN or GIGN), its ability to assign dynamic importance to different atomic neighbors makes it a strong baseline for structure-based molecular bioactivity prediction tasks. In our experiments, GAT serves as a 2D graph-based baseline model for comparing the effect of incorporating geometric and docking-based features.

A.2.4 EHIGN model

Explainable Heterogeneous Interaction Graph Neural Network (EHIGN) [Yang et al., 2024] is a graph neural network method for predicting protein-ligand binding affinity from 3D structures, based on an interaction-based inductive bias that aligns with biological binding mechanisms. The method represents protein-ligand complexes as heterogeneous graphs with two node types (ligand and protein atoms) and four edge types corresponding to ligand intramolecular, protein intramolecular, ligand-protein intermolecular, and protein-ligand intermolecular interactions. EHIGN employs four specialized graph convolution networks—two Covalent Interaction Graph Convolution networks (CIGConvs) using sum aggregation for covalent bonds and two Non-covalent Interaction Graph Convolution networks (NIGConvs) using mean aggregation for non-covalent interactions—to independently process each interaction type and prevent covalent interaction information from being overwhelmed by the more numerous non-covalent interactions.

The model predicts binding affinity as the sum of pairwise atom-atom affinities rather than using a bottleneck architecture, which preserves 3D structural information and enhances interpretability. To mitigate the interaction hidden bias (where predictions could spuriously correlate with the number of atom pairs), EHIGN incorporates a learnable bias correction term that uses a bottleneck architecture with attention-weighted pooling to capture count-related features. Experimental results demonstrate that EHIGN achieves superior performance across multiple benchmark datasets (PDBbind core sets, CSAR-HiQ, and LIT-PCBA) and provides physically meaningful interpretations consistent with known binding principles, including accurate distance-affinity relationships and the ability to distinguish between different binding poses.

A.2.5 SIGN model

Structure-aware Interactive Graph Neural Networks (SIGN) [Li et al., 2021] is a graph neural network framework for predicting protein-ligand binding affinity that addresses two critical challenges in structure-based drug discovery: incomplete modeling of 3D spatial structure and the loss of long-range molecular interactions. The method consists of two main components: Polar-Inspired Graph Attention Layers (PGAL) and Pairwise Interactive Pooling (PiPool). PGAL captures comprehensive 3D spatial information by establishing a polar coordinate system for each atom and preserving both distance and angle relationships through alternating node-to-edge and edge-to-node interaction layers. This approach models geometric structures more completely than distance-only methods while remaining invariant to rotation and translation, unlike Cartesian coordinate-based approaches. To recover long-range interaction information lost during graph construction (where only nearby protein atoms are retained), SIGN incorporates PiPool, which performs atomic type-aware pooling on intermolecular edges to generate a global interaction matrix representing protein-ligand contacts across all atomic type pairs. This matrix is trained with a reconstruction loss as an auxiliary self-supervised task, enabling the model to simultaneously learn local spatial geometry and global interaction patterns.

A.2.6 MBP model

Multi-task Bioassay Pre-training (MBP) [Yan et al., 2023] is a pre-training framework designed to improve structure-based protein-ligand binding affinity (PLBA) prediction by leveraging large-scale bioassay data. The method addresses inherent challenges in utilizing extensive but noisy databases

such as ChEMBL, including heterogeneous affinity measurement types, experimental variability across different bioassays, and the absence of experimentally determined 3D structures. MBP employs a multi-task learning approach that combines direct affinity regression with bioassay-specific pairwise ranking, enabling the model to learn robust structural representations while mitigating the impact of label noise and inconsistencies.

A.3 Detail of datasets

A.3.1 DTIGN dataset

The DTIGN dataset comprises 8 protein targets (I1, I2, I3, I4, I5, E1, E2, and E3), each considered as an individual subset. The original dataset is non-I.I.D. due to the manually constructed test set. To approximate an I.I.D. setting, we randomly re-split each subset. For each subset, 20% of the data is used as the test set, while the remaining 80% is split into training and validation sets with a 9:1 ratio. The information of each subset is shown in Table 3. The number of atoms varies depending on the size of the original binding site for each protein.

Table 3: The information of the DTIGN dataset

Subset	Task	Size of Union-Pocket	Number of data points
I1	IC_{50}	978	926
I2	IC_{50}	3,353	1,509
I3	IC_{50}	4,338	3,501
I4	IC_{50}	1,301	5,920
I5	IC_{50}	1,974	9,336
E1	EC_{50}	2,140	996
E2	EC_{50}	1,312	1,639
E3	EC_{50}	1,870	3,532

A.3.2 SIU dataset

The SIU dataset is a recent benchmark for bioactivity prediction, featuring multiple types of bioactivity measurements. In this work, we select K_d and K_i as target labels, as they complement the bioactivity types used in the DTIGN dataset. There are two versions of the SIU dataset: SIU-0.9 and SIU-0.6, where the number indicates the sequence similarity threshold used to exclude test set samples that are too similar to those in the training set. For the SIU dataset, we adopt the ‘distinct pocket’ defined in the dataset (identified by a single PDB ID) as the Union-Pocket. The information of each subset we use is shown in Table 4.

Specifically, in this study, SIU (K_d) subsets are used for evaluation of the ligand binding affinity (LBA) prediction task in the additional study. The detail is shown in Table 5. SIU has a much larger scale than the conventional PDBbind dataset [Wang et al., 2004]; the comparison is shown in Table 6.

A.4 Additional experimental results

A.4.1 Evaluation on additional baseline models

We have expanded our experiments to include additional competitive baselines: the explainable heterogeneous interaction graph neural network (EHIGN), the structure-aware interactive graph neural network (SIGN), and the multi-task bioassay pre-training (MBP) approach on some subsets. Results (Table 7 and Table 8) show that the proposed method consistently improves performance over these models across most metrics (39/43) on representative benchmark subsets.

Together with the existing GAT, GIGN, and DTIGN, these additional methods form a diverse and competitive baseline set that includes heterogeneous, attention-based, and multi-task learning approaches. This broader comparison provides a more comprehensive assessment of the proposed model’s effectiveness.

Table 4: Information of the SIU dataset

Dataset	Task	Subset	Avg size of Union-Pocket	Number of data points
SIU-0.6	K_i	test	1,243	3,153
		train	1,161	33,595
		valid	1,164	3,787
	K_d	test	1,148	1,723
		train	1,132	12,333
		valid	1,125	1,422
SIU-0.9	K_i	test	1,243	3,178
		train	1,145	160,480
		valid	1,143	17,754
	K_d	test	1,155	1,779
		train	1,066	43,472
		valid	1,067	4,945

Table 5: Information of SIU-0.9 and SIU-0.6 (K_d) datasets

Dataset	# Unique protein	# Unique ligand	# Protein-ligand complex	# Ligand conformations
SIU0.6 / K_d	1216	1860	17509	70674
SIU0.9 / K_d	4211	5201	54570	216602

Table 6: Comparison of SIU and PDBbind

Dataset	# Pocket-molecule pairs	# Avg. molecules per pocket	# Unique pockets	# Unique molecules
PDBbind	19,443	1	19,443	19,443
SIU	1,312,827	137.6	9,544	214,686

Table 7: Additional model performance on DTIGN subsets

Dataset / Label	Method	RMSE↓	r ↑	τ ↑
DTIGN I1 / IC ₅₀	EHIGN	1.215	0.146	0.052
	EHIGN+LigoSpace	1.239	0.193	0.108
	SIGN	0.911	0.681	0.492
	SIGN+LigoSpace	0.909	0.712	0.539
	MBP	1.280	0.148	0.036
	MBP+LigoSpace	1.152	0.466	0.325
DTIGN I2 / IC ₅₀	EHIGN	1.089	0.042	0.011
	EHIGN+LigoSpace	0.999	0.053	0.045
	SIGN	0.582	0.848	0.632
	SIGN+LigoSpace	0.569	0.848	0.641
	MBP	1.071	0.073	0.081
	MBP+LigoSpace	0.977	0.449	0.258
DTIGN E1 / EC ₅₀	EHIGN	0.994	-0.050	-0.026
	EHIGN+LigoSpace	1.006	0.010	-0.009
	SIGN	0.887	0.464	0.393
	SIGN+LigoSpace	0.861	0.506	0.446
	MBP	0.972	0.028	0.005
	MBP+LigoSpace	0.913	0.334	0.241

A.4.2 Evaluation on ligand binding affinity prediction task

Ligand binding affinity (LBA), typically quantified by the dissociation constant (K_d), is included in the SIU dataset and serves as a relevant benchmark for evaluating virtual screening performance on docked conformations.

Table 8: Additional model performance on SIU subsets

Dataset / Label	Method	RMSE↓	MAE↓	r ↑	Spearman↑
SIU-0.6 / K_d	EHIGN	1.404	1.180	-0.015	0.020
	EHIGN+LigoSpace	1.325	1.110	0.159	0.213
	MBP	1.693	1.418	-0.195	-0.189
	MBP+LigoSpace	1.406	1.194	0.082	0.041
SIU-0.6 / K_i	EHIGN	1.450	1.222	0.118	0.106
	EHIGN+LigoSpace	1.394	1.164	0.256	0.169
	MBP	1.748	1.483	0.304	0.253
	MBP+LigoSpace	1.699	1.434	0.317	0.225

To assess our method in this context, we analyzed pK_d prediction performance on the SIU-0.9 and SIU-0.6 subsets, which are large-scale and structurally diverse datasets, sharing the same test set but differing in training data.

As shown in Table 9, out of 24 reported metrics (RMSE, MAE, r , Spearman across all models and both subsets), 19 showed improvements with our method.

Table 9: Model performance comparison on SIU- K_d

Dataset/Label	Method	RMSE↓	MAE↓	r ↑	Spearman↑
SIU-0.9 / K_d	Vina	2.018	1.564	0.120	0.127
	Unimol	1.364	1.141	-0.033	-0.082
	DTIGN	1.839	1.490	-0.001	-0.042
	DTIGN+LigoSpace	1.304	1.060	0.321	0.326
	GIGN	1.708	1.367	0.070	0.038
	GIGN+LigoSpace	1.455	1.139	0.296	0.261
	GAT	1.545	1.240	0.092	0.082
	GAT+LigoSpace	1.473	1.166	0.261	0.254
SIU-0.6 / K_d	Vina	2.018	1.564	0.120	0.127
	Unimol	1.389	1.192	-0.149	-0.206
	GIGN	1.371	1.115	0.265	0.281
	GIGN+LigoSpace	1.326	1.078	0.280	0.227
	DTIGN	1.349	1.079	0.329	0.329
	DTIGN+LigoSpace	1.332	1.069	0.327	0.248
	GAT	1.521	1.285	0.233	0.157
	GAT+LigoSpace	1.424	1.182	0.107	0.133

In summary, our results show consistent performance gains: RMSE improved by 11.13%, MAE improved by 11.55%, Pearson correlation improved by 135.48%, Spearman correlation improved by 151.83% averaged across models and datasets, as summarized in Table 10 below.

Table 10: Summary of average improvements on SIU- K_d

	RMSE↓	MAE↓	r ↑	Spearman↑
Baseline avg for K_d	1.554	1.272	0.097	0.085
LigoSpace-enhanced avg for K_d	1.381	1.125	0.229	0.213
Percentage improvement	11.13%	11.55%	135.48%	151.83%

Additionally, we evaluated the Vina scoring function on the SIU- K_d test set (same for both SIU-0.9 and SIU-0.6). For each protein-ligand pair, multiple docked conformations were scored using AutoDock Vina. These scores were converted to predicted dissociation constants using the following standard transformation:

$$pK_d = -\frac{\Delta G}{RT \cdot \ln(10)}$$

where ΔG is the binding free energy reported by Vina (in kcal/mol), R is the gas constant (1.987×10^{-3} kcal/mol·K), and T is the temperature (assumed to be the standard state condition, i.e., 298 K). For consistency, we used pK_d as the prediction target and averaged the pK_d values across all conformations for each pair.

We found that the models enhanced by our method significantly outperform the Vina baseline across all error and correlation metrics (Table 9). These findings reinforce the practical value of our approach in affinity prediction and virtual screening workflows, demonstrating superior predictive performance even when relying solely on docked poses.

Furthermore, we would like to emphasize that, unlike traditional docking methods, which score only a single binding conformation, our approach predicts experimental affinity by integrating information from multiple plausible docking conformations. This multi-conformation strategy aligns with practices adopted in recent datasets and better reflects real-world scenarios where various binding poses collectively influence the observed affinity.

Beyond SIU dataset, we also conducted supplementary experiments to validate our method on a redocked PDBbind v2020 dataset (K_d only). For each PDB entry, the top three docking poses were used. The models were trained on the general subset (4,466 PDB entries, excluding all refined-set samples) and evaluated on the refined set (2,783 PDB entries). As a baseline, Vina scores were converted pK_d and averaged over the top three poses. The results, presented in Table 11, show that our LigoSpace-enhanced methods consistently outperform the baselines across all metrics, demonstrating their generalizability on the PDBbind dataset.

Table 11: Performance comparison on PDBbind dataset

Method	RMSE↓	r ↑	τ ↑
AutoDock Vina	2.488	0.261	0.269
GIGN	1.428	0.216	0.105
DTIGN	1.144	0.499	0.316
GIGN+LigoSpace	1.334	0.421	0.276
DTIGN+LigoSpace	1.045	0.608	0.356

A.4.3 The impact of different aggregation strategies

As discussed in Eq. 5 of the main text, we aggregate multiple pocket–ligand pose embeddings to improve the accuracy of bioactivity prediction. In our current experiments, all models use mean aggregation unless otherwise specified. The only exception is DTIGN, which by default employs its own attention-based aggregation strategy.

To analyze the effect of different aggregation strategies, we conducted controlled comparisons where both the baseline DTIGN and DTIGN+LigoSpace were equipped with the same aggregation function—either mean, sum, or attention. As in our other experiments, this comparison ensures that the observed improvements stem from our method rather than differences in aggregation. The results are shown in Table 12.

It can be noticed that on three DTIGN subsets (I1, I2, E1), our method consistently improves performance across all aggregation strategies, though the extent of improvement varies. In all settings, DTIGN+LigoSpace outperforms the original DTIGN with the same aggregation function:

- With mean aggregation, our method achieves an average RMSE reduction of 9.62%, along with improvements of 26.43% in r and 25.42% in τ .
- With sum aggregation, improvements are more pronounced: 14.26% reduction in RMSE, and increases of 32.44% in r and 46.54% in τ .
- With attention-based aggregation, the LigoSpace-enhanced model achieves the best results—12.16% lower RMSE, and substantial gains in correlation metrics: 41.54% in r and 52.46% in τ .

These results demonstrate that our method is robust to different aggregation choices, enhancing the baseline model regardless of the specific strategy used.

Table 12: Performance comparison across different aggregation functions

Dataset/Aggfunction	Method	RMSE↓	Pearson↑	Tau↑
DTIGN I1 / Mean	DTIGN	1.1971	0.4431	0.3307
	DTIGN+LigoSpace	1.0183	0.6352	0.4356
DTIGN I2 / Mean	DTIGN	0.7863	0.6760	0.4340
	DTIGN+LigoSpace	0.6781	0.7786	0.5453
DTIGN E1 / Mean	DTIGN	0.8817	0.4121	0.3166
	DTIGN+LigoSpace	0.8803	0.4976	0.3764
Overall / Mean - Avg. imp		9.62%	26.43%	25.43%
DTIGN I1 / Sum	DTIGN	1.1441	0.4708	0.2947
	DTIGN+LigoSpace	0.9781	0.6744	0.4872
DTIGN I2 / Sum	DTIGN	0.7466	0.7186	0.5044
	DTIGN+LigoSpace	0.6170	0.8164	0.5932
DTIGN E1 / Sum	DTIGN	0.9313	0.3782	0.2620
	DTIGN+LigoSpace	0.8298	0.5312	0.4105
Overall / Sum - Avg. imp		14.26%	32.44%	46.54%
DTIGN I1 / Attn	DTIGN	1.1977	0.3547	0.2445
	DTIGN+LigoSpace	1.0364	0.5895	0.4239
DTIGN I2 / Attn	DTIGN	0.7952	0.7128	0.4922
	DTIGN+LigoSpace	0.6507	0.7888	0.5454
DTIGN E1 / Attn	DTIGN	0.9086	0.3363	0.2139
	DTIGN+LigoSpace	0.8645	0.4969	0.3705
Overall / Attn - Avg. imp		12.16%	41.54%	52.46%

A.5 Explanation of the relationship between geometric edge and empty space

GeoREC captures spatial emptiness implicitly via geometric edge construction. For each atom, the surrounding space is partitioned into K cones of equal solid angle. Each cone connects the central atom to its nearest neighbor, and the distance s to the nearest atom in each cone provides a lower bound on unoccupied volume in that direction.

Consider: A surrounding sphere of radius r is partitioned into K cones. Then the volume per cone can be represented as the following:

$$V_{\text{cone}} = \frac{4}{3} \pi \frac{r^3}{K}.$$

If the nearest atom in a cone lies at a distance s , the empty volume in that direction is at least:

$$V_{\text{empty}} \geq \frac{4}{3} \pi \frac{s^3}{K}.$$

Thus, longer distances imply more emptiness, and this grows with s^3 . GeoREC encodes the distance to the nearest atom in each cone, providing a compact and physically meaningful proxy for local spatial emptiness.

A.6 Evaluation metrics

Here, we provide the detailed definitions of the metrics adopted in this paper. Let the predicted and ground truth label of bioactivity be \hat{y}_i and y_i , respectively; the total number of data points is N .

Root Mean Squared Error (RMSE) is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2} \quad (2)$$

Pearson Correlation Coefficient (r) is defined as:

$$r = \frac{\sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}})^2} \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}} \quad (3)$$

where $\bar{\hat{y}}$ and \bar{y} are the means of the predicted and ground truth labels, respectively. r ranges from -1 to 1 , indicating the degree of linear correlation.

Kendall’s Tau Correlation Coefficient (τ) is defined as:

$$\tau = \frac{C - D}{\frac{1}{2}N(N - 1)} \quad (4)$$

where C and D are the numbers of concordant and discordant pairs among all $\frac{1}{2}N(N - 1)$ possible pairs (i, j) such that $i < j$. A pair is concordant if $(\hat{y}_i - \hat{y}_j)(y_i - y_j) > 0$, and discordant if $(\hat{y}_i - \hat{y}_j)(y_i - y_j) < 0$.

τ measures the ordinal association between two variables, ranging from -1 (complete disagreement) to 1 (complete agreement).

Mean Absolute Error (MAE) is defined as:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i| \quad (5)$$

It measures the average absolute difference between predicted and ground truth values, treating all errors equally regardless of their direction.

Spearman’s Rank Correlation Coefficient is defined as:

$$\rho = 1 - \frac{6 \sum_{i=1}^N d_i^2}{N(N^2 - 1)} \quad (6)$$

where $d_i = \text{rank}(\hat{y}_i) - \text{rank}(y_i)$ is the difference between the ranks of the predicted and ground truth values for the i -th data point.

ρ measures the strength and direction of the monotonic relationship between predicted and ground truth values. It ranges from -1 (perfect inverse rank correlation) to 1 (perfect rank agreement).

A.7 Comparison with Uni-Mol

When compared with the large-scale pretrained model Uni-Mol, which was trained on 209 million 3D molecules and 3 million candidate pockets [Zhou et al., 2023], our LigoSpace-enhanced baseline models achieve comparable performance without pretraining. Notably, the LigoSpace-enhanced models substantially outperform Uni-Mol in terms of r and Spearman correlation for the K_d label, highlighting the effectiveness of our pairwise loss in capturing inter-sample correlations.

Here, we compare Uni-Mol with our method regarding the training speed, training device, and model size, shown in Table 13. The training speed of Uni-Mol is estimated based on the reported pretraining duration and the number of epochs used for molecular and pocket representation learning, as described in Zhou et al. [2023]. Our method with GNN, exemplified by DTIGN+LigoSpace model, exhibits a significantly smaller model size and achieves comparable training speed. This highlights the practicality and efficiency of our method, demonstrating its potential for easy deployment and integration in bioactivity prediction pipelines.

A.8 Compare with the pairwise loss in other works

Several related works have adopted ranking or pairwise loss. In our work, the pairwise loss complements the primary objective by explicitly modeling relative differences between samples. While the

Table 13: Comparison with Uni-Mol on the training speed, training device, and model size

Comparison metric	Uni-Mol	DTIGN+LigoSpace
Approx. training speed (per min)	16K small pocket-ligand pairs	8K UnionPocket-ligand pairs
Training device	8 V100 GPUs (32GB)	4 RTX4090 GPUs (24GB)
Model size	415MB	2MB

core idea of employing pairwise loss is similar, the problem settings differ. Here, we take MBP [Yan et al., 2023] and ActFound [Feng et al., 2024] as two examples. A detailed comparison with MBP and ActFound is provided below.

In MBP, they use BCE as their ranking loss. We summarize the difference between their ranking loss and our pairwise loss in Table 14:

Table 14: Comparison of loss functions with MBP

Aspect	Loss in MBP	Loss in our work
Objective	Learns relative order (which is larger/smaller)	Learns relative difference (how much larger/smaller)
Loss type	Binary classification	Regression
Noise Robustness	More robust to absolute value noise	Focuses on precise difference alignment

In ActFound, the pairwise loss enables meta-learning for few-shot adaptation. E.g., ligand bioactivity data from the same assay are inherently comparable, allowing the model to learn relative differences even with limited data. This loss also guides the model to refine predictions on new assays using only a few samples. In our work, the pairwise loss is a supplementary term in a hybrid loss for a regression task. The detailed comparison is shown in Table 15.

Table 15: Comparison of loss functions with ActFound

Aspect	Loss in ActFound	Loss in our work
Objective	Enable few-shot adaptation via meta-learning	Improve static prediction via relative order
Calculation scope	Task-specific subsets	All samples in a batch
Typical Use Case	Low-data, cross-assay adaptation	General bioactivity prediction
Loss Integration	Core to meta-learning loops	Hybrid loss with MSE

A.9 Analysis of using docking conformations

Docking methods can generate inaccurate complex conformations, which may impair the performance of models trained on docked structures. Users should be aware of this potential issue and are encouraged to use biologically plausible docking poses obtained from well-established software tools. In the DTIGN dataset, ligand poses were generated using AutoDock Vina [Trott and Olson, 2010], and the top-ranked poses were selected for model training (as stated in the ‘‘Docking method’’ section on Page 2 and the footnote of Table 1 in [Yin et al., 2024]). In the SIU dataset, docking was performed using AutoDock Vina, Glide [Friesner et al., 2004], and GOLD [Verdonk et al., 2003], with a voting strategy applied to select representative poses (as described in the ‘‘Structural data construction via multi-software docking’’ section on Page 5 of [Huang et al., 2025]).

Nevertheless, predicting affinity from docking conformations offers key advantages over using co-crystal structures. Co-crystal structures are scarce and costly to obtain, especially for novel targets, whereas docking provides scalable access to approximate protein–ligand complexes. Most of our training data lack co-crystal structures, so using docked conformations enables broader applicability. Moreover, deep learning can be designed to denoise and learn meaningful patterns, making our approach practical and robust for real-world tasks like virtual screening.

A.10 Limitations

We acknowledge that incorporating global pocket information—i.e., using the Union-Pocket—can increase the number of atom nodes in the GNN, thereby raising computational resource requirements. However, this increase is relatively modest and acceptable in practice, as the number of atoms in the

Union-Pocket typically ranges from a few hundred to a few thousand. Furthermore, the additional computational time is negligible when compared to the much longer timescales involved in real-world drug discovery processes.

References

- Hong-Qiang Ding, Naoki Karasawa, and William A Goddard III. Atomic level simulations on a million particles: The cell multipole method for coulomb and london nonbond interactions. *The Journal of chemical physics*, 97(6):4309–4315, 1992.
- James B Dunbar Jr, Richard D Smith, Chao-Yie Yang, Peter Man-Un Ung, Katrina W Lexa, Nickolay A Khazanov, Jeanne A Stuckey, Shaomeng Wang, and Heather A Carlson. Csar benchmark exercise of 2010: selection of the protein–ligand complexes. *Journal of chemical information and modeling*, 51(9):2036–2046, 2011.
- Bin Feng, Zequn Liu, Nanlan Huang, Zhiping Xiao, Haomiao Zhang, Srбуhi Mirzoyan, Hanwen Xu, Jiaran Hao, Yinghui Xu, Ming Zhang, et al. A bioactivity foundation model using pairwise meta-learning. *Nature Machine Intelligence*, 6(8):962–974, 2024.
- Richard A Friesner, Jay L Banks, Robert B Murphy, Thomas A Halgren, Jasna J Klicic, Daniel T Mainz, Matthew P Repasky, Eric H Knoll, Mee Shelley, Jason K Perry, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of medicinal chemistry*, 47(7):1739–1749, 2004.
- Luca Gagliardi and Walter Rocchia. Siteferret: beyond simple pocket identification in proteins. *Journal of Chemical Theory and Computation*, 19(15):5242–5259, 2023.
- Yanwen Huang, Bowen Gao, Yinjun Jia, Hongbo Ma, Wei-Ying Ma, Ya-Qin Zhang, and Yanyan Lan. Redefining the task of bioactivity prediction. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Shuangli Li, Jingbo Zhou, Tong Xu, Liang Huang, Fan Wang, Haoyi Xiong, Weili Huang, Dejing Dou, and Hui Xiong. Structure-aware interactive graph neural networks for the prediction of protein-ligand binding affinity. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 975–985, 2021.
- RDKit Team. Rdkit: Open-source cheminformatics. URL <https://www.rdkit.org>. Accessed: 2025-05-20.
- 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.
- Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.
- Marcel L Verdonk, Jason C Cole, Michael J Hartshorn, Christopher W Murray, and Richard D Taylor. Improved protein–ligand docking using gold. *Proteins: Structure, Function, and Bioinformatics*, 52(4):609–623, 2003.
- Renxiao Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang. The pdbbind database: Collection of binding affinities for protein- ligand complexes with known three-dimensional structures. *Journal of medicinal chemistry*, 47(12):2977–2980, 2004.
- Jiaxian Yan, Zhaofeng Ye, Ziyi Yang, Chengqiang Lu, Shengyu Zhang, Qi Liu, and Jiezhong Qiu. Multi-task bioassay pre-training for protein-ligand binding affinity prediction. *Briefings in Bioinformatics*, 25(1), 2023.
- Ziduo Yang, Weihe Zhong, Qiuji Lv, Tiejun Dong, and Calvin Yu-Chian Chen. Geometric interaction graph neural network for predicting protein–ligand binding affinities from 3d structures (gign). *The journal of physical chemistry letters*, 2023.

- Ziduo Yang, Weihe Zhong, Qiuji Lv, Tiejun Dong, Guanxing Chen, and Calvin Yu-Chian Chen. Interaction-based inductive bias in graph neural networks: enhancing protein-ligand binding affinity predictions from 3d structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):8191–8208, 2024.
- Yueming Yin, Hilbert Yuen In Lam, Yuguang Mu, Hoi-Yeung Li, and Adams Wai-Kin Kong. Advancing bioactivity prediction through molecular docking and self-attention. *IEEE J. Biomed. Health Informatics*, 2024.
- Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023.