

GADTI: A Geometry-Aware Dual-Tower Framework for Drug-Target Interaction Prediction

Pan Zeng^a, Tian Liu^a, Yajie Meng^b, Xianfang Tang^b, Feifei Cui^c, Zilong Zhang^c, Junlin Xu^{d,*}, Taoyang Wu^{e,*}

^a*School of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China*

^b*School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan, Hubei 430200, China*

^c*School of Computer Science and Technology, Hainan University, Haikou, Hainan 570228, China*

^d*School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, 430065, Hubei, China*

^e*School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, U.K.*

Abstract

Accurate prediction of drug–target interactions (DTI) is essential for virtual screening and early-stage drug discovery, as it helps identify potential therapeutic compounds and narrow the experimental search space. Existing geometric structure-based DTI models have attempted to incorporate three-dimensional coordinates or spatial relationships to better characterize molecular interactions. However, they still face challenges in fully characterizing the spatial constraints involved in real binding processes. First, the alignment between full-length protein semantics and local binding-pocket information remains insufficient, which may either introduce noise from non-binding regions or discard useful global evolutionary context. Second, chemical semantic features and three-dimensional geometric features lie in heterogeneous feature spaces, making simple feature concatenation inadequate for precise cross-modal fusion. Third, current geometric structure models may still lack rotational and translational consistency, limiting their ability to generalize to realistic binding patterns and novel scaffold samples. To address the above issues, this study proposes GADTI, a geometry-aware dual-tower framework for DTI prediction. First, GADTI introduces a pocket-centric global–local alignment strategy, which maps full-length protein semantics to predicted binding pocket residues, thereby preserving global evolutionary information while focusing on local structural regions. Second, GADTI adopts a semantic–geometric dual-tower architecture to represent the 1D/3D features of drugs and targets, and uses hierarchical bidirectional cross-attention to achieve modality alignment and interaction, rather than relying on simple concatenation. Meanwhile, a PaiNN-based $E(3)$ -equivariant geometric encoder is introduced to learn spatial features from three-dimensional coordinates that are consistent under rotational and translational transformations. The experimental results show that GADTI achieves competitive predictive performance on the structurally filtered Davis and BindingDB benchmark samples and maintains relatively stable generalization under the scaffold-split setting. These findings indicate that, for DTI samples with complete structural and semantic inputs, pocket alignment, semantic–geometric fusion, and equivariant geometric encoding contribute to enhancing the model’s ability to learn structure-aware representations of drug–target interactions.

Keywords: Drug-target interaction, Geometric deep learning, $E(3)$ -equivariant geometric encoder, Hierarchical cross-attention, 3D spatial information

1. Introduction

Accurate prediction of drug–target interactions (DTI) is a key task in early-stage drug discovery and virtual screening (Gao et al., 2024; Bagherian et al., 2021; Coley, 2023; Lin et al., 2020). Although traditional high-throughput experimental screening can directly validate the activity of candidate compounds, it is often limited by long experimental cycles, substantial resource consumption, and restricted coverage of the chemical space (DiMasi et al., 2016). Therefore, machine learning- and deep learning-based DTI prediction methods have increasingly become useful tools for narrowing the search space of candidate compounds and accelerating lead compound discovery. Early approaches mainly relied on similarity measures, feature engineering, and conventional machine learning algorithms, whereas recent deep learning models are able to automatically learn representations of drugs and proteins. Previous reviews have systematically summarized machine learning-based DTI prediction methods and highlighted the importance of integrating heterogeneous drug and target information (Chen et al., 2018). Nevertheless, how to effectively integrate drug chemical information, protein sequence information, and three-dimensional structural information within a unified model remains an important challenge in this field.

Existing DTI prediction methods mainly include sequence-based, graph-based, and three-dimensional structure-based models. Sequence-based models usually treat drug SMILES strings and protein amino acid sequences as one-dimensional textual inputs. For example, DeepDTA employs convolutional neural networks to extract local sequential features from drugs and targets (Öztürk et al., 2018). Graph neural network-based models represent drugs as molecular graphs and capture molecular structural information through topological connections between atoms and chemical bonds. For instance, GraphDTA enhances drug representation by using molecular graph structures (Nguyen et al., 2021). These 1D/2D methods generally offer favorable computational efficiency and often achieve relatively high performance under random data splits. However, they lack explicit three-dimensional spatial constraints and are therefore less capable of modeling physical factors involved in real binding processes, such as interatomic distances, directionality, steric effects, and pocket complementarity. As a result, when test samples contain novel chemical scaffolds that are not observed during training, models may rely on statistical correlations or frequent substructures in the training data, which can lead to reduced scaffold-level generalization (Ji et al., 2023; Chen et al., 2021).

To address the limitations of low-dimensional representations, recent studies have started to introduce three-dimensional structure-based DTI modeling approaches, such as 3D CNNs, structure-aware graph neural networks, and 3D GNNs (Wallach et al., 2015; Li et al., 2019, 2022). These methods directly use the three-dimensional coordinate information of proteins or ligands and, in principle, can better capture spatial dependencies between atoms, local conformations, and stereochemical characteristics. Recent molecular generation studies have also emphasized the importance of structural constraints and causal relationships in molecular space modeling, although such methods are mainly designed for molecule generation rather than direct DTI prediction (Ahamed et al., 2025). However,

*Corresponding author

Email addresses: xjl@hnu.edu.cn (Junlin Xu), Taoyang.wu@uea.ac.uk (Taoyang Wu)

existing 3D structure-based models still have several limitations. On the one hand, molecular and protein structures have rotational and translational degrees of freedom in three-dimensional space. If a model does not explicitly guarantee rotational and translational invariance or equivariance, the learned spatial features may depend on the specific coordinate system, leading to inconsistent representations of the same molecule under different spatial orientations. On the other hand, conventional 3D CNNs or complex 3D GNNs often need to process a large number of spatial relationships among atoms or residues, resulting in considerable computational and memory costs in DTI prediction scenarios, which may restrict their application to large DTI datasets. In addition, although some 3D methods incorporate structural inputs, they still rely on simple concatenation or standard attention mechanisms for cross-modal fusion, which may be insufficient for aligning one-dimensional chemical semantics, global protein evolutionary information, and three-dimensional geometric constraints. Consequently, these methods may still face difficulties in consistently capturing the spatial complementarity and physical feasibility involved in real binding processes.

In addition to the insufficient exploitation of structural information, existing multimodal DTI methods still have limitations in feature fusion. Early multimodal approaches commonly concatenate drug representations, protein representations, and structural features directly (Jiang et al., 2020; Zhao et al., 2022). However, one-dimensional sequence semantics and three-dimensional geometric features reside in different feature spaces, and simple linear fusion may be insufficient for accurate alignment and may even introduce additional noise (Liu et al., 2022). In recent years, methods such as MolTrans and DrugBAN have modeled interactions between drug substructures and protein fragments through cross-attention or bilinear attention, thereby improving feature interaction to some extent (Huang et al., 2021b; Bai et al., 2023). Associative learning-based DTI models further indicate that explicitly modeling drug-target association patterns can improve representation learning and prediction performance (Zhu et al., 2023). Nevertheless, the attention weights in these methods are mainly derived from statistical correlations among sequence, substructure, or topological features, rather than being explicitly constrained by true three-dimensional spatial proximity and geometric complementarity. As a result, the model may assign high weights to atoms or residues that are spatially distant and physically unlikely to interact, leading to potential mispredictions for samples with similar topology but mismatched spatial conformations. As also noted in the original manuscript, existing attention mechanisms lack explicit geometric guidance and may be unable to distinguish complex binding patterns that are topologically similar but spatially inconsistent.

To address these issues, we propose GADTI, a geometry-aware dual-tower framework for DTI prediction, aiming to improve both the utilization of three-dimensional spatial information and the fusion of cross-modal features. First, GADTI constructs a pocket-centric global-local alignment strategy. This strategy preserves the evolutionary semantics of the full-length protein sequence while focusing structural modeling on the potential binding pocket region, thereby reducing spatial noise introduced by non-binding regions. Second, the model incorporates a PaiNN-based E(3)-equivariant geometric encoder, which enables atom-level three-dimensional representations to remain physically consistent under rotational and translational transformations and thus helps capture distances, directions, and local spatial dependencies. Finally, GADTI designs a hierarchical bidirectional cross-attention mechanism to align one-

dimensional semantic features and three-dimensional geometric features within the drug tower and target tower, and further models cross-molecular interactions between drugs and targets. This design helps mitigate heterogeneous feature mismatches caused by simple concatenation. As described in the Methodology section, GADTI uses ChemBERTa and ESM-2 to extract one-dimensional semantic features, incorporates PaiNN to process three-dimensional spatial structures, and subsequently performs modality alignment and intermolecular interaction modeling through hierarchical bidirectional cross-attention.

The experimental results show that GADTI achieves competitive predictive performance on the structure-complete Davis and BindingDB benchmark samples and shows a smaller performance decrease than the compared baselines under the scaffold split setting. These observations suggest that explicit geometric constraints may provide useful spatial information for DTI prediction in novel-scaffold scenarios.

2. Methods

Geometry-Semantic Dual-Tower Architecture. To accurately capture drug-target interaction patterns at the atomic level, we propose the GADTI framework. As illustrated in Figure 1, the model employs a dual-tower architecture that processes proteins and drug molecules in parallel. During feature construction, ChemBERTa and ESM-2 are used only for offline extraction of 1D semantic embeddings from drug SMILES and protein sequences, respectively. In the main GADTI model, each tower contains two parallel branches: a semantic projection branch operating on the cached embeddings and an E(3)-equivariant geometric encoder processing 3D spatial structures (Schütt et al., 2021). Instead of simple concatenation, the model employs a hierarchical bidirectional cross-attention mechanism to sequentially achieve intra-tower modality alignment and inter-tower molecular interaction. Finally, the interacting drug and protein features are fed into a classification head composed of a multilayer perceptron (Ba et al., 2016; Hendrycks and Gimpel, 2016) and layer normalization. This module serves as a readout function, mapping the high-dimensional interactive representations to a binding probability score to perform binary classification of the drug-target interaction state.

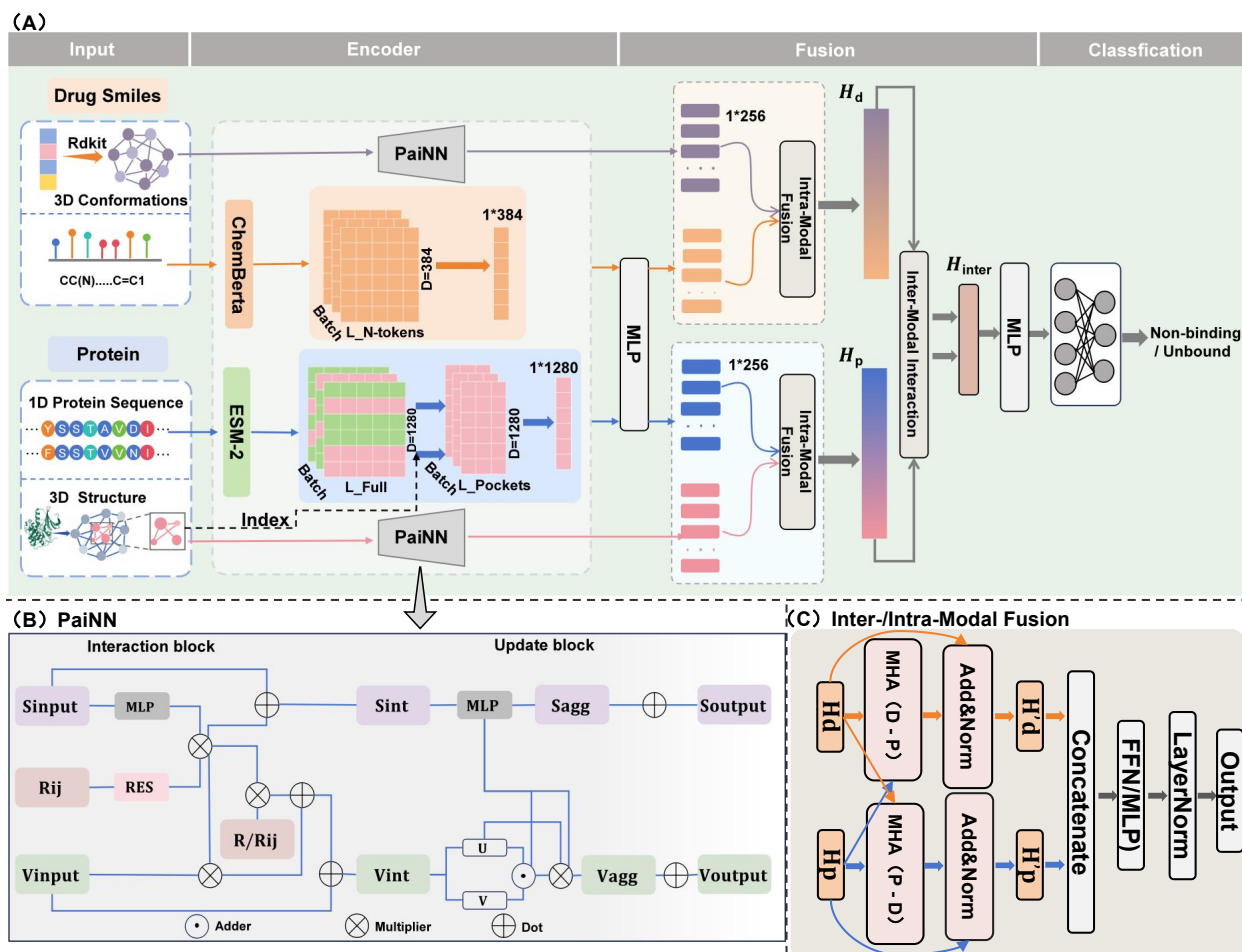


Figure 1. Overview of the GADTI architecture. (A) The overall framework utilizes a dual-tower structure, employing pre-trained models (ChemBERTa/ESM-2) and PaiNN to extract 1D semantic and 3D geometric features of drugs and targets, respectively. (B) Detailed internal structure of the PaiNN encoder, which captures geometric dependencies in 3D space through interaction and update modules. (C) The inter-modality fusion mechanism, which utilizes bidirectional cross-attention to deeply aggregate drug-target features and output the final prediction.

To accurately capture directional interactions between atoms in continuous 3D space and resolve the alignment of heterogeneous features, we incorporate PaiNN as the geometric encoder and design a hierarchical dual-tower attention mechanism to enhance feature fusion:

Layer-wise Implementation of GADTI. The GADTI dual-tower framework is implemented following the sequence: semantic projection $\rightarrow$ geometric encoding $\rightarrow$ intra-tower fusion $\rightarrow$ inter-tower interaction $\rightarrow$ classification prediction. For both the drug and target towers, cached 1D semantic features are first mapped to a unified hidden representation space via separate MLP projection layers, yielding preliminary semantic embeddings for the drug and protein. Concurrently, the geometric branch performs PaiNN message passing on the drug atom graph and the protein pocket graph to obtain node-level 3D geometric features, which are aggregated into graph-level geometric embeddings via global average pooling. Within each tower, the 1D semantic and 3D geometric embeddings are fused through bidi-

rectional cross-attention to produce the drug-level and protein-level embeddings. Finally, the outputs of the drug and protein towers interact via inter-tower bidirectional cross-attention, resulting in a joint drug–target representation that is fed into an MLP classifier to generate prediction logits for assessing drug–target interactions.

PaiNN-based Geometric Feature Enhancement. To explicitly model the anisotropy of the local geometric environment, we employ PaiNN instead of traditional graph networks that reduce coordinates to scalar distances. In this study, E(3)-equivariant geometric modeling means that the learned spatial representations remain physically consistent when the molecular system is translated or rotated in three-dimensional space. In other words, the model does not learn binding patterns from a fixed coordinate orientation, but from relative geometric relationships such as interatomic distances, directions, and local spatial arrangements. This property is important for protein–ligand modeling because the biological meaning of a molecular structure should not change when the same structure is viewed from a different coordinate frame. This network enhances the model’s perception of spatial structure by maintaining rotation-equivariant vector features $\vec{v}_i \in \mathbb{R}^{F \times 3}$ and rotation-invariant scalar features $s_i \in \mathbb{R}^F$. During message passing, the spatial relationship between neighboring nodes (i, j) is decomposed into radial basis expansion and directional propagation:

$$\Delta \vec{v}_i = \sum_{j \in \mathcal{N}(i)} \left(\phi_s(s_j) \cdot W_r(e_{ij}) \cdot \frac{\vec{r}_{ij}}{|\vec{r}_{ij}|} + \beta_{ij} \odot \vec{v}_j \right) \quad (1)$$

Here, the radial component describes how strongly two atoms interact as a function of their distance, while the directional component preserves the orientation of the interaction in 3D space. The model utilizes directional information between atoms to guide feature updates and determines the strength of this spatial information flow through element-wise ($\odot$) multiplication. Subsequently, in the node update phase, the model introduces a geometric gating mechanism (Dauphin et al., 2017) to fuse scalar and vector information (Jing et al., 2021). The input vector $\vec{v}_i$ is linearly mapped into two independent feature spaces, $\mathbf{U}$ and $\mathbf{V}$, and their inner product $\langle \mathbf{U}, \mathbf{V} \rangle$ serves as a gating factor to modulate the evolution of the scalar feature s_i . This mechanism ensures that microscopic geometric orientations directly regulate the update flow of atomic physicochemical properties. Therefore, the PaiNN branch provides a physically consistent geometric representation that complements the 1D semantic embeddings used in the drug and target towers.

Hierarchical Dual-Tower Attention Fusion. Given that 1D evolutionary/chemical semantics and 3D geometric features exist in heterogeneous feature spaces, simple concatenation may lead to information misalignment. Therefore, we construct a two-stage hierarchical feature fusion strategy. The first stage is intra-molecular modality alignment. For either the drug or protein branch, let $\mathbf{h}_{1D}$ be the sequence projection feature and $\mathbf{h}_{3D}$ be the globally pooled geometric feature. Bidirectional cross-attention is used to establish the mapping between semantics and geometry:

$$\mathbf{h}_{mol} = \text{FFN}([\text{Attn}(Q = \mathbf{h}_{1D}, KV = \mathbf{h}_{3D}) \parallel \text{Attn}(Q = \mathbf{h}_{3D}, KV = \mathbf{h}_{1D})]) \quad (2)$$

This process forces semantic features to attend to specific spatial conformations while enriching geometric representations with chemical context, resulting in unified representations $\mathbf{H}_{drug}$ and $\mathbf{H}_{prot}$ for the drug and protein, respectively.

The second stage is global interactive reasoning, where the model further captures intermolecular interactions. Using the drug representation H_{drug} as the Query and the protein representation H_{prot} as the Key/Value, the model

calculates the attention of the drug molecule to the pooled protein representation, dynamically updating drug features by integrating target information. Symmetrically, the parallel path uses the protein representation H_{prot} as the Query and the drug representation H_{drug} as the Key/Value to extract the protein representation’s accommodation tendency for the ligand representation, thereby capturing critical geometric complementarity. The final fusion vector is input into the classifier to complete the reasoning process from atomic geometry perception to macroscopic binding activity prediction.

Datasets. In order to comprehensively evaluate the performance of the model, we chose the Davis (Davis et al., 2011) dataset and the BindingDB (Liu et al., 2007) dataset for drug-target interaction prediction (Huang et al., 2021a). The Davis dataset serves as the primary benchmark for model performance, containing 442 kinases, 68 drug molecules, and 25,772 drug-target interaction records. BindingDB is used to evaluate model scalability and generalization on a large-scale heterogeneous dataset, comprising approximately 2.8 million manually curated interaction records involving a large number of compounds and protein targets.

For the Davis dataset, we implemented the following data cleaning pipeline: Drug molecules were processed using RDKit combined with the MMFF94 force field to generate and optimize stable 3D conformations (Riniker and Landrum, 2015). We removed 12,128 records that failed due to the presence of rare elements (e.g., selenium), valence errors, or convergence failures. Subsequently, we used complete full-length protein structures provided by AlphaFold2 (Jumper et al., 2021) to replace crystal structures with missing residues, and removed 1,020 entries where protein names and indices did not match. This yielded a final dataset of 12,456 unique drug-target pairs.

A similar pipeline was applied to the BindingDB dataset. We retained only protein molecules with PDB ID indices (Liu et al., 2015). Following drug 3D conformation generation and quality control, we removed 746 drug molecules that failed to generate valid 3D conformations due to non-standard elements or valence errors. In total, we obtained 171,346 high-quality drug-target interaction pairs with complete 3D coordinates, which were used as the structure-complete BindingDB benchmark samples for training and evaluating geometry-aware DTI models. Detailed statistics for each dataset, including the number of drug–target pairs, unique proteins, and unique molecules, are presented in Table 1.

Table 1. Dataset Statistics

Dataset	Total Pairs	Unique Drugs	Unique Proteins	Positive	Negative
BindingDB	171,346	115,569	2,592	128,375 (74.92%)	42,971 (25.08%)
Davis	12,456	36	346	3,091 (24.82%)	9,365 (75.18%)

This cleaning strategy ensures that each training sample achieves strict physical alignment across four dimensions: 3D pocket geometry, ESM-2 evolutionary features (Lin et al., 2023), drug 3D conformation, and ChemBERTa semantic features (Chithrananda et al., 2020). For label processing, following previous work (Huang et al., 2020), we normalized all binding affinities to pK_x values. Binarization was implemented using a threshold of $pK \geq 6.0$

for BindingDB and $pK \geq 5.0$ for Davis (Zhao et al., 2025) to define the positive and negative sample distributions required for training.

To investigate the differences in physicochemical properties between the two datasets, we visualized the distribution of drug heavy atoms and protein sequence lengths for the cleaned datasets (see Figure 2).

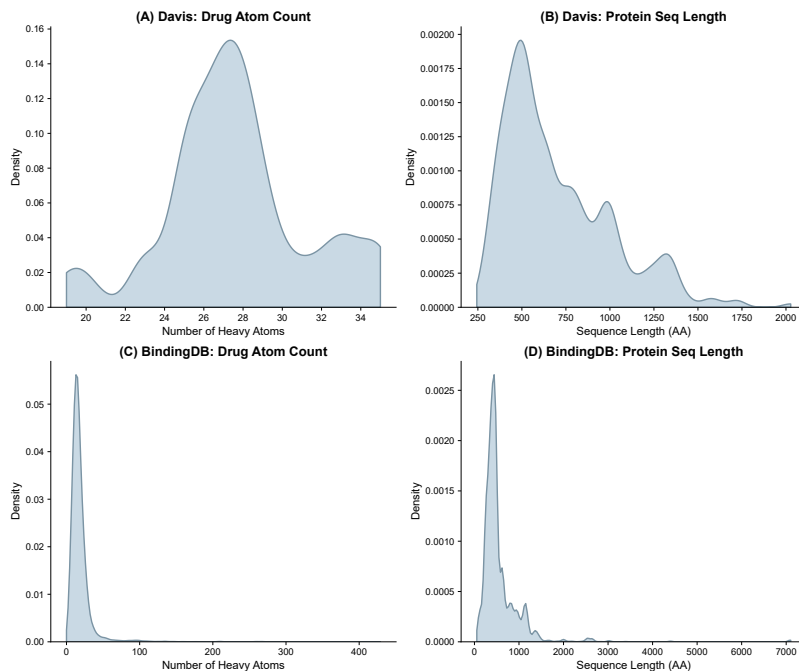


Figure 2. Distribution of proteins and drugs in the datasets.

As illustrated, the Davis dataset (Figure 2A, B) exhibits a relatively smooth and broad quasi-normal distribution. This occurs because the Davis dataset focuses primarily on the kinase family, where protein sequence lengths are concentrated in the 250–1000 amino acid range, and the chemical space of the drug molecules is relatively constrained. Conversely, the BindingDB dataset (Figure 2C, D) displays significant peaks and long-tail distributions. The majority of samples are concentrated in the lower molecular weight and shorter protein sequence regions, but the dataset also includes extremely long protein sequences and complex macromolecular drugs. Consequently, the Davis dataset is suited for evaluating the model’s fine-grained discrimination capability within specific protein families, whereas the BindingDB dataset is used to assess model scalability and applicability on large-scale heterogeneous drug-target pairs.

Data Splitting Protocols. To evaluate the model’s predictive ability across different scenarios, this study employs two data partitioning strategies: random partitioning and Bemis-Murcko scaffold partitioning. Random partitioning is used to evaluate the model’s overall predictive performance under standard benchmark settings. The Davis dataset uses five-fold cross-validation, with 80% of the samples in each fold used for training and 20% for testing. The BindingDB dataset uses an 80:20 random partitioning strategy to evaluate the model’s predictive ability on large-scale datasets.

To further evaluate the model’s generalization ability to unseen chemical scaffolds, this study constructs a Bemis-Murcko scaffold split. Specifically, Bemis-Murcko scaffolds are first extracted from each drug SMILES using RDKit, and all drug–target pairs with the same scaffold are grouped into the same scaffold group. Subsequently, data are partitioned using scaffold groups as the basic unit, ensuring that the same chemical scaffold does not appear simultaneously in both the training and test sets. After partitioning, we further examine the scaffold overlap between the training and test sets, confirming it to be zero. Compared with random partitioning, this strategy can more rigorously evaluate the model’s out-of-distribution generalization ability on novel ligand scaffolds.

Molecular Representation. The design of the protein representation aims to balance the local geometric accuracy of the binding site with the long-range evolutionary dependencies of the sequence. Directly inputting the full-length sequence introduces redundant noise from non-binding regions, while solely encoding isolated pocket fragments results in the loss of the global context required for protein folding. Therefore, this study constructs a pocket-centric geometric graph $G_p = (V_p, E_p)$ and employs a geometric index-based feature alignment strategy. The geometric subspace structure uses P2Rank (Krivák and Hoksza, 2018) to locate the potential binding pocket and extract the C_α coordinates of the residue in the high confidence area. This forms a simplified 3D point cloud, filtering out spatial noise from inactive regions. Specifically, we extract the top-ranked (Top-1) high-confidence pocket residues predicted by P2Rank as the binding site. For the minority of samples lacking distinct pockets where P2Rank fails to predict a site, the model automatically triggers a full-length fallback mechanism. This mechanism utilizes the global attention weights of ESM-2 for adaptive feature compensation, improving inference continuity during large-scale screening.

To achieve the local mapping of global semantics, we encode the full-length amino acid sequence using the pre-trained protein language model ESM-2 (Lin et al., 2023; Elnaggar et al., 2021). This embeds comprehensive evolutionary semantics into the local geometric nodes, generating a full-length embedding matrix $E_{\text{full}} \in \mathbb{R}^{L \times 1280}$. Subsequently, feature extraction is executed based on the pocket residue index set S_{id} determined by the geometric coordinates:

$$V_{\text{pocket}} = \text{IndexSelect}(E_{\text{full}}, S_{id}) \quad (3)$$

This mechanism ensures that although the vectors in V_{pocket} correspond to local residues, their values are computed via the self-attention mechanism of the full-length sequence, thus incorporating global topological and evolutionary information. Decoupling data processing from geometric learning in the architecture allows large-scale inference to be executed as offline preprocessing. This avoids the VRAM overhead of repeatedly loading large models during the training loop and significantly improves iteration efficiency on large datasets. Concurrently, for samples with missing P2Rank predictions, the model employs the full-length fallback mechanism to maintain feature availability.

The drug representation is designed to address the inability of 2D topological graphs to distinguish stereoisomers (e.g., chirality, cis-trans isomerism). We construct an atomic graph $G_d = (V_d, E_d)$ that integrates explicit 3D coordinates with implicit chemical semantics. The 3D conformations of drug molecules are generated and optimized using RDKit (Landrum et al., 2006) with the MMFF94 force field (Halgren, 1996), thereby imposing explicit spatial con-

213 straints and providing the resulting atomic coordinates. This enables the model to bypass the limitations of 2D graph
 214 neural networks and directly perceive spatial positioning and shape complementarity between molecules. Pre-trained
 215 1D semantics are obtained using ChemBERTa-77M-MLM as a semantic encoder to supplement the atomic chemical
 216 environment information. Through pre-computed inference on standardized SMILES sequences, the hidden states
 217 from the final layer of the Transformer (Chen et al., 2020) are extracted:

$$H_{\text{SMILES}} = \text{Transformer}(\text{SMILES}) \quad (4)$$

218 After removing special tokens, the chemical context-rich embedding vector $v_{\text{sem}} \in \mathbb{R}^{384}$ is strictly aligned with the 3D
 219 atomic nodes. This multimodal input ensures the model possesses both a semantic understanding of chemical bond
 220 properties and a perception of atomic spatial distributions.

221 Notably, because acquiring true co-crystallized (Holo) structures in ultra-large-scale virtual screening is computa-
 222 tionally prohibitive and frequently unfeasible, this model uses AlphaFold2-predicted unbound protein structures and
 223 RDKit-generated ligand conformers as input structural proxies. Although the induced-fit effect during ligand binding
 224 is not explicitly modeled, the additional structural-noise analysis suggests that these static conformational proxies can
 225 still provide useful geometric information under the evaluated benchmark settings. Together with the local geometric
 226 perception of the PaiNN module and the multimodal attention mechanism, these proxy inputs support the learning of
 227 spatially informative representations from large-scale structure-complete DTI samples.

228 **Inference and Training.** To address class imbalance, this study employs a binary focal loss instead of the standard
 229 binary cross-entropy loss (Lin et al., 2017). This loss function reweights hard examples by introducing a dynamic
 230 scaling factor. It is defined as follows:

$$L = -\frac{1}{N} \sum_{i=1}^N [\alpha y_i (1 - p_i)^\gamma \log(p_i) + (1 - \alpha)(1 - y_i) p_i^\gamma \log(1 - p_i)] \quad (5)$$

231 where $y_i \in \{0, 1\}$ is the true label and $p_i \in [0, 1]$ is the predicted probability output by the model. Two key hyper-
 232 parameters are introduced to adjust the learning weights: the focusing parameter γ is used to mine hard examples
 233 by down-weighting easily classified samples, while the balancing parameter α is used to offset the class imbalance
 234 between positive and negative samples by adjusting their relative loss contributions, thereby mitigating the impact of
 235 class imbalance.

236 In the inference stage, given a drug-target input pair of (d, p) , the model extracts and merges features through the
 237 dual-tower architecture, and outputs the final scalar logit z . We use the sigmoid activation function to map it to the
 238 binding probability:

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}} \quad (6)$$

239 The probability value $\hat{y}$ reflects the predicted likelihood of drug-target interaction. In practical applications, we
 240 set the threshold value $\tau = 0.5$ and classify the sample of $\hat{y} \geq 0.5$ as potential active drug-target pairs.

3. Results and Discussion

Experimental Setup and Baseline Methods. To systematically evaluate the predictive performance of GADTI, we used AUROC (Davis and Goadrich, 2006; Harrell et al., 1982), AUPRC, accuracy, F1-score, and MCC as the main evaluation metrics. AUROC and AUPRC assess the overall ranking ability of the model across different classification thresholds, whereas accuracy, F1-score, and MCC evaluate classification performance under a fixed decision threshold. For the Davis dataset, all models were evaluated using five-fold cross-validation under random splits. Specifically, the dataset was randomly divided into five mutually exclusive subsets. In each validation round, four subsets were used for training and the remaining subset was used for testing. The final results are reported as the mean and standard deviation over the five independent runs. For the structure-complete BindingDB benchmark samples used in this study, we adopted an 80:20 random split while following the same preprocessing workflow and label construction protocol as used for Davis, thereby evaluating the applicability of the model to large-scale DTI prediction. Affinity values were uniformly converted into the pK representation, and $\text{pK} \geq 5.0$ and $\text{pK} \geq 6.0$ were used as the positive-label thresholds for the Davis and BindingDB datasets, respectively. For scaffold split evaluation, Bemis-Murcko scaffolds were extracted using RDKit, and samples with the same scaffold were assigned to the same subset to prevent scaffold overlap between the training and test sets. This setting was used to assess the generalization ability of the model to unseen chemical scaffolds.

The main hyperparameter settings of GADTI were kept consistent on the Davis and BindingDB datasets. The model used a 4-layer PaiNN geometric encoder with a hidden dimension of 256 and an interatomic cutoff radius of 6.0 Å. The protein semantic branch used ESM-2 embeddings with a dimension of 1280, whereas the drug semantic branch used ChemBERTa embeddings with a dimension of 384. The model was implemented in PyTorch and optimized using AdamW (Kingma and Ba, 2014). The initial learning rate was set to 1×10^{-5} , with a weight decay of 1×10^{-4} . The training process lasted for 50 epochs, and model checkpoints were saved every 5 epochs.

To comprehensively evaluate the relative performance of GADTI, we compared it with representative DTI/DTA prediction methods, including sequence-based models, graph-based models, and substructure- or attention-based interaction models. The sequence-based baselines included DeepDTA (Öztürk et al., 2018), TransformerCPI (Chen et al., 2020), and MPNN-CNN (Tsubaki et al., 2019). The graph-based baselines included GraphDTA (Nguyen et al., 2021) and PSICHIC (Koh et al., 2024), and the substructure- or attention-based baselines included DrugBAN (Bai et al., 2023), MolTrans (Huang et al., 2021b), and TankBind (Lu et al., 2022). For a fair comparison, all reproduced baselines were evaluated on the processed datasets used in this study and followed the same label construction and data-splitting protocols as GADTI, with five-fold cross-validation on Davis and a random split on BindingDB when applicable.

Performance Comparison on Benchmark Datasets. Table 2 summarizes the primary performance metrics for each model under the standard random split of the Davis dataset.

Table 2. Performance comparison of GADTI with baseline models on the Davis dataset.

Model	AUROC	AUPRC	Accuracy	F1-score	MCC
MPNN-CNN(Tsubaki et al., 2019)	0.8678 $\pm$ 0.004	0.7472 $\pm$ 0.005	0.8431 $\pm$ 0.011	0.6116 $\pm$ 0.030	0.5004 $\pm$ 0.051
DeepDTA(Öztürk et al., 2018)	0.8977 $\pm$ 0.002	0.7861 $\pm$ 0.013	0.8558 $\pm$ 0.008	0.6700 $\pm$ 0.015	0.5907 $\pm$ 0.026
GraphDTA(Nguyen et al., 2021)	0.8938 $\pm$ 0.007	0.7814 $\pm$ 0.019	0.8544 $\pm$ 0.013	0.6628 $\pm$ 0.035	0.5845 $\pm$ 0.036
MolTrans(Huang et al., 2021b)	0.8923 $\pm$ 0.003	0.7789 $\pm$ 0.012	0.8514 $\pm$ 0.002	0.6781 $\pm$ 0.015	0.5861 $\pm$ 0.012
TransformerCPI(Chen et al., 2020)	0.9007 $\pm$ 0.002	0.7956 $\pm$ 0.010	0.8624 $\pm$ 0.007	0.7023 $\pm$ 0.006	0.6541 $\pm$ 0.020
DrugBAN(Bai et al., 2023)	0.8777 $\pm$ 0.000	0.7505 $\pm$ 0.000	0.7592 $\pm$ 0.000	0.7289 $\pm$ 0.000	-
PSICHIC(Koh et al., 2024)	0.9003 $\pm$ 0.0086	0.7916 $\pm$ 0.0110	0.8699 $\pm$ 0.0089	0.7281 $\pm$ 0.0262	0.6449 $\pm$ 0.0299
TankBind(Lu et al., 2022)	0.8864 $\pm$ 0.0087	0.7610 $\pm$ 0.0245	0.7962 $\pm$ 0.0086	0.6597 $\pm$ 0.0101	0.5364 $\pm$ 0.0149
GADTI(Ours)	0.9318 $\pm$ 0.001	0.8463 $\pm$ 0.007	0.8840 $\pm$ 0.004	0.7587 $\pm$ 0.011	0.6833 $\pm$ 0.014

Note: Values are presented as “mean $\pm$ standard deviation”. The best performing result in each column is **bolded**.

Table 3. Performance comparison of GADTI and baseline models on the BindingDB dataset.

Model	AUROC	AUPRC	F1-Score	Accuracy	Recall
GraphDTA (Nguyen et al., 2021)	0.8796	0.9529	0.8968	0.8397	0.9303
MolTrans(Huang et al., 2021b)	0.8535	0.9399	0.8857	0.8239	0.9115
DeepDTA(Öztürk et al., 2018)	0.8738	0.9488	0.8939	0.8376	0.9135
TransformerCPI(Chen et al., 2020)	0.8512	0.9395	0.8842	0.8218	0.9087
DrugBAN (Bai et al., 2023)	0.8938	0.9581	-	0.8357	0.7546
PSICHIC(Koh et al., 2024)	0.8828	0.9521	0.9031	0.8512	0.9252
GADTI (Ours)	0.9210	0.9663	0.9016	0.8596	0.8803

Note: The best performing result in each column is **bolded**.

In comparative experiments on the Davis dataset, GADTI demonstrates advantages across all core evaluation metrics. To reduce evaluation bias caused by random data partitioning and improve the reliability of the comparison, we conducted five-fold cross-validation for GADTI and all deep learning baselines, reporting their average performance. PSICHIC and TankBind were additionally included to provide graph-based and structure-aware comparisons, respectively. Experimental results show that our model achieves a mean AUROC of 0.9318 and an AUPRC of 0.8463, outperforming the compared baseline models on the Davis dataset. Compared with PSICHIC and TankBind, GADTI achieves higher AUROC, AUPRC, accuracy, F1-score, and MCC.

Beyond fine-grained evaluation, we conducted extended experiments on the BindingDB benchmark samples retained after structural-completeness filtering to evaluate the model’s scalability and applicability on large-scale heterogeneous data. PSICHIC was included as a scalable graph neural network baseline for the full-dataset evaluation. As shown in Table 3, GADTI achieved competitive performance, surpassing DrugBAN and PSICHIC with an AUROC of 0.9210 and an AUPRC of 0.9663. Although PSICHIC achieves a slightly higher F1-score and GraphDTA obtains the highest recall, GADTI achieves the best AUROC, AUPRC, and accuracy.

TankBind was attempted on all retained BindingDB samples used in this study. Due to the high computational and memory cost of constructing explicit pairwise representations between pocket residues and ligand atoms, several large-pocket or large-ligand samples caused GPU out-of-memory errors even with a batch size of 1. To enable evaluation, we constructed a stratified TankBind-processable subset containing 50,000 samples, including approximately 36,000 positive and 14,000 negative samples, while preserving the original data splits and the positive-to-negative ratio as closely as possible. Both TankBind and GADTI were evaluated on this subset. Considering the additional computational cost of this subset experiment and the observation that validation performance had largely converged within 20 epochs in preliminary runs, both models were trained for 20 epochs under the same subset setting. As summarized in Table 4, GADTI achieved competitive performance compared with TankBind on the TankBind-processable BindingDB subset. TankBind obtained slightly higher AUROC, AUPRC, and recall, whereas GADTI achieved slightly higher accuracy and F1-score. In addition, GADTI required substantially less training time under the same subset setting, completing the 20-epoch experiment within approximately 3 hours, whereas TankBind required more than 10 hours. These results suggest that GADTI achieves a favorable balance between competitive predictive performance and computational efficiency on the large-scale processable subset.

Table 4. Performance comparison of TankBind and GADTI on the TankBind-processable BindingDB subset.

Model	AUROC	AUPRC	Accuracy	F1-score	Recall	Training time
TankBind	0.8706	0.9464	0.8058	0.8644	0.8526	>10 h
GADTI	0.8634	0.9451	0.8070	0.8661	0.8348	<3 h

Note: The table reports results on the stratified BindingDB subset constructed to be processable by TankBind. Both TankBind and GADTI were trained for 20 epochs under the same subset setting. Training time denotes the approximate wall-clock time observed in our reproduction.

Robustness Analysis under Scaffold Split. To more rigorously evaluate the generalization ability of the model in real drug discovery scenarios, we changed the experimental setting from random split to more challenging out-of-distribution scaffold split. Specifically, we utilized RDKit to extract the Bemis-Murcko scaffolds (Bemis and Murcko, 1996) of the ligand molecules in the dataset and performed a greedy split using scaffold clusters as the minimum unit. This ensured that the ligand scaffolds in the test set never appeared in the training set. This strategy prevents topological feature leakage between the training and test sets, testing whether the model has learned transferable spatial patterns rather than relying mainly on high-frequency substructures. As shown in Figure 3, we compared the performance degradation of GADTI and baseline models (DeepDTA, GraphDTA) under both splitting strategies (Yang et al., 2019).

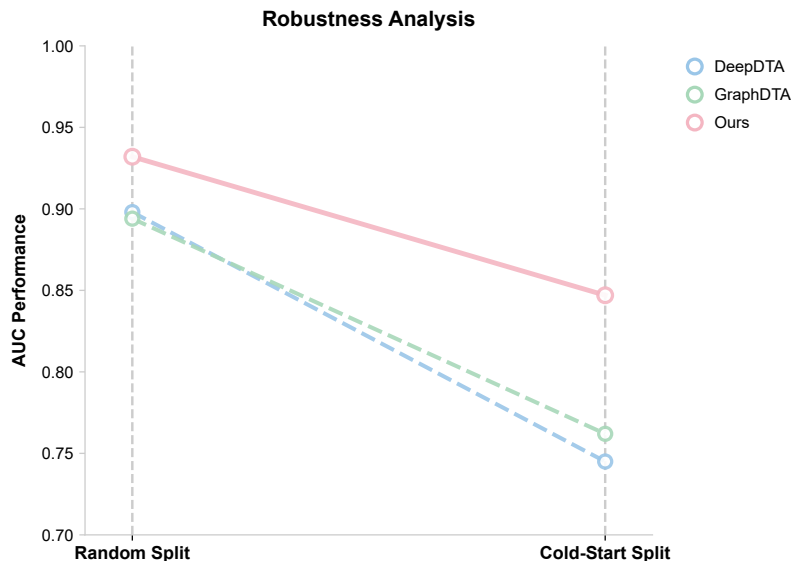


Figure 3. Performance comparison of DeepDTA, GraphDTA, and GADTI under scaffold split conditions.

Compared with the baselines, sequence- and graph-based models typically exhibit an AUROC drop (generally >15%) when evaluated on novel scaffolds, falling to approximately 0.75. This indicates that 1D/2D models rely more on memorizing the distribution of training scaffolds and struggle to transfer learned features to unknown structures. In contrast, GADTI showed a smaller performance decrease than the compared baselines. Across three independent repeated experiments, our model maintained an average AUROC of 0.8469, a decrease of about 9.1% relative to the random split. This observation suggests that the 3D module provides useful spatial constraints under the scaffold-split setting. Although the ligand scaffold is unseen during training, the geometric branch can introduce complementary spatial information for evaluating drug–target pairs with novel scaffolds.

Ablation Study. To investigate the specific roles of 1D semantic features and 3D geometric features in DTI prediction, we constructed two model variants: (1) w/o 1D (retaining only the 3D branch) and (2) w/o 3D (retaining only the sequence branch). Experimental results are summarized in Table 5.

Table 5. Ablation study results for the contributions of sequence (1D) and geometric (3D) features under random split and scaffold split settings.

Split Strategy	Model Variant	AUROC	AUPRC	MCC	Precision	Accuracy
Random Split	w/o 1D	0.8328 ± 0.004	0.6662 ± 0.013	0.4675 ± 0.026	0.7457 ± 0.050	0.8206 ± 0.007
	w/o 3D	0.9313 ± 0.002	0.8453 ± 0.008	0.6933 ± 0.011	0.7632 ± 0.021	0.8847 ± 0.005
	full GADTI	0.9318 ± 0.001	0.8463 ± 0.007	0.6833 ± 0.014	0.7837 ± 0.012	0.8840 ± 0.004
Scaffold Split (Cold Start)	w/o 1D	0.7846 ± 0.000	0.6915 ± 0.000	0.5024 ± 0.000	0.9220 ± 0.000	0.7973 ± 0.000
	w/o 3D	0.8014 ± 0.000	0.7131 ± 0.000	0.4644 ± 0.000	0.7032 ± 0.000	0.7843 ± 0.000
	full GADTI	0.8469 ± 0.000	0.7715 ± 0.000	0.5269 ± 0.000	0.9162 ± 0.000	0.8067 ± 0.000

Experimental results indicate that 1D semantic features provide an important chemical and biological semantic foundation for drug-target interaction prediction. In both random split and scaffold split settings, removing the 1D branch (w/o 1D) causes the most severe performance degradation. Particularly under the scaffold split, the AUROC drops to 0.7846. This indicates that topological information—such as atom types and functional group connections embedded in SMILES strings and amino acid sequences—provides an important basis for the model to construct molecular representations. Completely abandoning one-dimensional features will make the architecture lose basic semantic support, making it difficult to obtain competitive predictive results.

The contribution of the 3D module varies across different scenarios, and this adaptability helps explain the model’s good generalization ability. In the random split experiment with scaffold overlap between the training set and the test set, the AUROC difference between the complete model and the variant without 3D module (w/o 3D) is very small. This suggests that under the random-split setting, one-dimensional sequence patterns account for most of the predictive performance. However, in the more challenging scaffold split, the importance of 3D geometric features becomes more evident; removing this module causes AUROC to drop from 0.8469 to 0.8014. This performance gap suggests that sequence information alone is less effective when encountering unseen molecular scaffolds. The invariants extracted by PaiNN provide complementary information for cross-scaffold generalization.

Although the pure sequence model (w/o 3D) achieved competitive performance under the random split, its lower performance under the scaffold split indicates that sequence information alone is insufficient to distinguish samples with similar semantics but different spatial compatibility. In contrast, the complete model with 3D constraints achieved higher precision and overall classification performance. This suggests that the 3D module can provide spatial constraints that help reduce physically implausible false-positive predictions. In the practical drug screening workflow, this ability to improve prediction reliability is valuable for reducing the cost of ineffective downstream experimental validation.

Robustness, Scalability, and Applicability Analysis.

To examine the reliability of GADTI beyond the main benchmark comparisons, we performed additional analyses on structural input quality, affinity-threshold selection, hyperparameter perturbation, computational cost, statistical reliability, dataset filtering, and failure cases. These analyses involve different matched subsets, filtering criteria, profiling settings, and retained-versus-removed sample comparisons. The detailed experimental settings and results are therefore provided in the corresponding main-text tables and the Supplementary Material, while the main text reports the principal observations. As summarized in Table 6, these analyses were used to evaluate the influence of structural input quality, threshold selection, hyperparameter settings, computational cost, statistical reliability, and dataset filtering on the model results.

Table 6. Summary of additional robustness, scalability, and applicability analyses.

Analysis	Evaluation focus	Main observation	Full results
Static-structure proxy deviation	Difference between static structural proxies and holo structures	Static structural proxies provided useful geometric inputs, but single RDKit conformers should not be regarded as true bound poses.	Table S1
Pocket-level pLDDT filtering	Influence of AF2 pocket confidence	Filtering by pocket-level pLDDT caused only minor changes in the main performance metrics on Davis.	Table S2
Threshold sensitivity	Dependence on affinity cutoff	AUROC remained relatively stable within the tested pK cutoff range, whereas AUPRC, F1-score, and MCC were more affected by class ratio changes.	Table S3
Hyperparameter sensitivity	Dependence on model and training settings	The default 4-layer PaiNN setting provided a practical balance between predictive performance and computational cost.	Table S4
Computational profiling	Model size, inference cost and scalability	The default 4-layer GADTI model contained 9.32M trainable parameters. The main online computational cost came from the PaiNN-based 3D geometric branch, while semantic features were precomputed and cached.	Table S5
Repeated-seed stability	Influence of random data partitioning and model initialization	The main random-split metrics showed relatively small variations across five independent random seeds.	Tables 7–8
Paired statistical analysis	Statistical reliability of model comparison	Using the existing PSICHIC five-fold partitions as fixed matched splits, GADTI obtained higher AUROC and AUPRC than PSICHIC.	Table S9
Dataset filtering and failure cases	Applicability and error sources	The results should be interpreted within structure-complete DTI benchmark settings, and some errors may be related to conserved pockets, boundary labels, or static conformer limitations.	Table S6

For the static-structure proxy analysis, RDKit-generated ligand conformers and AF2- or PDB-indexed protein structures were compared with matched experimental holo structures. In the Davis matched-holo subset, both ligand conformers and protein pockets showed relatively small spatial deviations from the corresponding holo structures. In

BindingDB, the deviation between isolated RDKit ligand conformers and holo ligand poses was larger, indicating that a single RDKit conformer should not be regarded as a bound pose. Correlation analysis did not provide consistent evidence that ligand or pocket deviation was associated with prediction error across the evaluated subsets. The results suggest the feasibility of static structural proxies as geometric inputs in large-scale DTI benchmarks, but also show that they cannot completely replace real holo complexes.

We also examined whether AF2 pocket confidence affected model performance. This analysis was conducted on Davis, where protein structures were derived from AF2 predictions; BindingDB was not included in this filtering analysis because its protein structures were PDB-indexed. Most Top-1 P2Rank pockets in Davis were located in medium- or high-confidence pLDDT regions. After filtering by pocket-level pLDDT, the main performance metrics changed only slightly relative to the all-sample setting. This indicates that low-confidence pockets were uncommon in Davis and had limited influence on the overall evaluation.

The threshold and hyperparameter analyses provided complementary observations. Changing the affinity cutoff within the tested range did not lead to a marked decline in AUROC on Davis or BindingDB, suggesting that the ranking performance was not tied to a single pK threshold. Metrics such as AUPRC, F1-score, and MCC varied more clearly with the positive sample ratio, as expected for class-imbalance-sensitive and fixed-threshold metrics. For hyperparameters, moderate changes in learning rate, PaiNN hidden dimension, and PaiNN depth did not lead to unstable performance. The default 4-layer PaiNN setting provided a practical balance between prediction performance and computational cost. The 6-layer setting was used only as a perturbation in the sensitivity analysis and was not adopted as the main configuration.

The computational efficiency profiling results provide information on the scalability of GADTI in large-scale DTI prediction scenarios. Profiling on Davis and BindingDB showed the model’s pair-wise inference performance and GPU memory usage under the tested settings. Module-level profiling indicated that the main online computational overhead of Full GADTI came from the PaiNN-based 3D geometric branch rather than hierarchical bidirectional cross-attention, ESM-2, or ChemBERTa. Because ESM-2 protein features and ChemBERTa drug features were pre-computed and cached offline, they were not repeatedly executed during each training epoch or online inference stage, and therefore were not the main bottleneck for high-throughput screening. The default 4-layer GADTI model contains 9,319,169 total parameters, all of which are trainable, corresponding to 9.32M trainable parameters; the corresponding parameter counts for different PaiNN depths are provided in Supplementary Table S5.

To assess the statistical stability of the random-split results, we conducted repeated experiments using five independent random seeds on the Davis and BindingDB datasets. For each seed, the data were repartitioned and the model was retrained under the same random-split setting. The main evaluation metrics were summarized as mean $\pm$ standard deviation, and 95% confidence intervals were calculated as $\text{mean} \pm t_{0.975,4} \times \text{std} / \sqrt{5}$. The repeated random-seed experiments showed relatively small variations in the main random-split metrics, especially for AUROC and AUPRC on both Davis and BindingDB. These results indicate that the reported random-split performance was not solely dependent on a single data partition or random initialization, and provide a more complete statistical descrip-

tion of the uncertainty associated with the reported results. The repeated-seed results are reported in Table 7, and the corresponding confidence intervals are reported in Table 8.

Table 7. Repeated random-seed results on the Davis and BindingDB datasets.

Dataset	AUROC	AUPRC	F1-score	Accuracy	MCC	Recall
Davis	0.9235 ± 0.0082	0.8304 ± 0.0138	0.7467 ± 0.0302	0.8652 ± 0.0272	0.6598 ± 0.0423	–
BindingDB	0.9182 ± 0.0014	0.9647 ± 0.0012	0.9015 ± 0.0030	0.8588 ± 0.0035	0.6539 ± 0.0048	0.8872 ± 0.0077

Note: Results are reported as mean $\pm$ standard deviation across five independent random seeds under the random-split setting. “–” indicates that the corresponding metric was not recorded in the repeated-seed experiment.

Table 8. 95% confidence intervals of repeated random-seed results.

Dataset	AUROC	AUPRC	F1-score	Accuracy	MCC	Recall
Davis	[0.9133, 0.9337]	[0.8133, 0.8475]	[0.7092, 0.7842]	[0.8314, 0.8990]	[0.6073, 0.7123]	–
BindingDB	[0.9165, 0.9199]	[0.9632, 0.9662]	[0.8978, 0.9052]	[0.8545, 0.8631]	[0.6479, 0.6599]	[0.8776, 0.8968]

Note: The 95% confidence interval was calculated as mean $\pm t_{0.975,4} \times \text{std} / \sqrt{5}$. “–” indicates that the corresponding metric was not recorded in the repeated-seed experiment.

In addition, to complement the statistical reliability of the model comparison, we conducted a matched five-fold paired analysis between GADTI and PSICHIC on the Davis dataset. Specifically, the existing PSICHIC five-fold partitions were used as fixed matched splits, and GADTI was re-evaluated on these same folds. GADTI obtained higher AUROC and AUPRC than PSICHIC, with paired t-test p-values of 0.0004 and 0.0014, respectively. The complete fold-wise results and statistical test results are provided in Supplementary Table S9. Given the available computational resources and time constraints, the current paired statistical analysis remains limited in its coverage of recent strong baselines and validation on the BindingDB dataset.

Finally, we examined dataset filtering and representative failure cases to clarify the applicability and limitations of the current benchmark results. Because GADTI requires complete multimodal inputs, including drug 3D conformers, protein pocket structures, drug semantic embeddings, and protein semantic embeddings, both Davis and BindingDB were filtered for structural completeness. This improves input consistency, but also restricts the interpretation of the results to structure-complete DTI benchmark settings. The retained-versus-removed sample analysis was used to assess changes in sample composition after filtering. In addition, we screened representative failure cases on the Davis test predictions, including false positives, false negatives, high-confidence wrong predictions, low-pLDDT-associated errors, and large-error samples. As reported in Supplementary Table S6, the screening identified 933 false positives, 676 false negatives, and 110 high-confidence wrong predictions. Among the 110 high-confidence wrong predictions, 63 occurred in high-pLDDT pockets, 45 in medium-pLDDT pockets, and only 2 in low-pLDDT pockets.

These results indicate that most high-confidence errors did not occur in low-confidence pockets. The remaining incorrect predictions may therefore be associated with factors such as conserved kinase pockets, boundary affinity labels, experimental noise, the use of a single ligand conformer, or conformational changes that cannot be represented by static structural inputs.

Visualization. To qualitatively and visually examine the restructuring effect of features on the model’s latent space, we extracted the final interactive feature vectors of the test set samples prior to the classification layer and projected them onto a 2D plane using the t-distributed stochastic neighbor embedding (t-SNE) algorithm. As shown in Figure 4, red represents active samples and blue represents inactive samples.

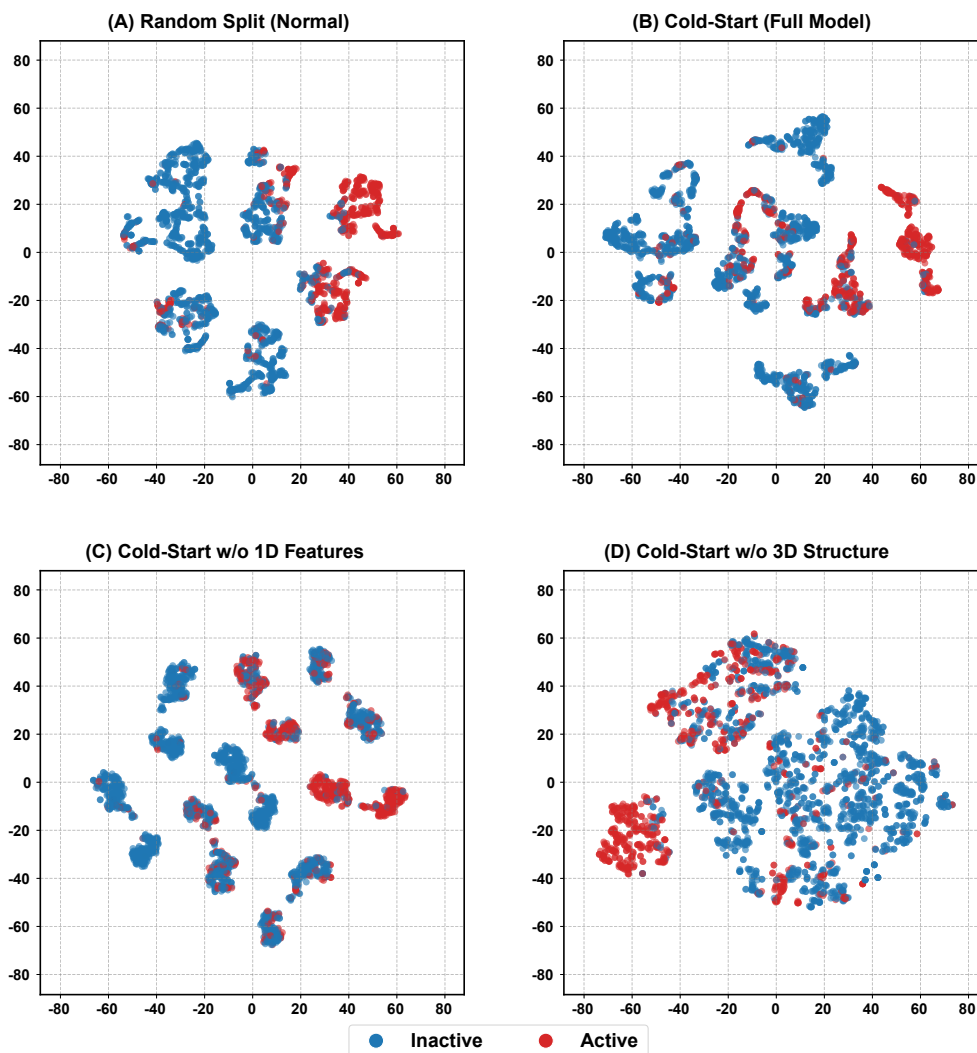


Figure 4. t-SNE visualization results. (A) Random Split (Normal), (B) Cold-Start (Full Model), (C) Cold-Start w/o 1D Features, (D) Cold-Start w/o 3D Structure.

Figure 4A displays the feature manifold of the complete model under the random split. Active and inactive

samples form relatively distinct clusters with limited overlap. This suggests that under the random-split setting, where the scaffold distributions of the training and test sets are similar, the model learns relatively clear decision boundaries, which is consistent with its AUROC of >0.93 . Figure 4C and Figure 4D visually reveal the behavior of the ablated models under scaffold-level generalization to novel scaffolds. Figure 4C shows that removing 1D features (w/o 1D) causes the feature space to fragment into isolated clusters, as the model loses global chemical semantic guidance. Figure 4D shows that relying solely on 1D semantic information (w/o 3D) results in greater class entanglement. Active samples (red points) are widely scattered throughout the blue inactive clusters, lacking a clear clustering structure. The model shows reduced ability to distinguish samples with similar sequences but different activities, resulting in less distinct decision boundaries and high false positive rates.

Compared with the model without 3D features in Figure 4D, the complete model with the 3D module in Figure 4B shows a redistributed feature space under the scaffold split setting. Despite encountering unseen scaffolds, the model reduces the overlapping area between the red and blue samples, and the active samples form relatively compact clusters. This transformation is consistent with our spatial filtering hypothesis: The 3D constraint characteristics extracted by PaiNN provide additional discrimination information, which can help distinguish between active binders and structurally mismatched baits, thus improving the prediction accuracy while maintaining the generalization ability.

To avoid relying solely on two-dimensional t-SNE projections to interpret the feature-space structure, we further quantified the high-dimensional final fusion embeddings used as the input to t-SNE. Specifically, we calculated the Silhouette score, Davies–Bouldin index, Calinski–Harabasz index, inter-/intra-class distance ratio, Fisher discriminant ratio, and nearest-neighbor label consistency. A concise summary of representative metrics is shown in Table 9, and the complete set of embedding-level metrics is provided in Supplementary Table S7. These metrics do not directly evaluate the two-dimensional projection itself; instead, they quantify class separation, intra-class compactness, and local neighborhood consistency in the original high-dimensional embedding space. This analysis was therefore used to examine whether the class-distribution patterns observed in the t-SNE visualization were also preserved before dimensionality reduction.

The quantitative results are summarized in Table 9. Under the random split, the final fusion embeddings showed a relatively stable class structure, with limited variation across the five runs. Under the scaffold split, the Silhouette score and inter-/intra-class distance ratio were slightly higher than those under the random split, suggesting that the model retained a certain degree of global class separation under the out-of-distribution scaffold setting. However, the Davies–Bouldin index, Calinski–Harabasz index, and nearest-neighbor consistency did not show consistent improvement, and their standard deviations were larger, indicating greater variability of the embedding structure under scaffold split. This is consistent with the greater difficulty of scaffold-level generalization.

These embedding-level metrics should not be interpreted as direct substitutes for classification metrics such as AUROC, AUPRC, F1-score, or MCC. Rather, they provide a quantitative complement to the t-SNE visualization from the perspective of representation learning. Taken together, the results indicate that the t-SNE visualization is consistent with the high-dimensional embedding analysis in showing a discernible class structure, while the interpretation of

latent-space separation under scaffold split should remain cautious.

Table 9. Summary of embedding-level quantitative metrics on the Davis dataset.

Dataset	Split	Silhouette $\uparrow$	Davies–Bouldin $\downarrow$	Inter/Intra $\uparrow$	NN@5 $\uparrow$
Davis	Random	0.0745 ± 0.0085	3.6924 ± 0.1853	1.0860 ± 0.0123	0.8302 ± 0.0081
Davis	Scaffold	0.0891 ± 0.0838	5.2824 ± 3.2223	1.1147 ± 0.1118	0.8103 ± 0.0601

Note: Values are reported as mean $\pm$ standard deviation. Metrics were calculated using standardized high-dimensional final fusion embeddings before t-SNE projection. Complete embedding metrics, including Calinski–Harabasz index, intra-/inter-class distance, Fisher ratio, and NN@10, are provided in Supplementary Table S7.

Case Study and Interpretability. Before presenting the specific case analysis, we first clarify the interpretability strategy used in this study. Because GADTI applies graph-level pooling to drug and protein representations before the cross-attention module, the inputs to cross-attention are not atom- or residue-level sequential representations, but pooled global representation vectors. In this setting, the attention weights mainly reflect vector-level interactions between the pooled drug representation and the pooled protein representation, and cannot be rigorously interpreted as one-to-one ligand atom–protein residue interaction weights. Therefore, directly using these attention weights as atom–residue interaction maps may lead to over-interpretation.

To avoid this issue, we did not use pooled attention weights as atom-level interpretability evidence. Instead, we supplemented the analysis with ligand atom-level Gradient $\times$ Input attribution and atom occlusion sensitivity analysis to evaluate the contribution of individual ligand atoms to the model output. The former estimates atom-level contributions by back-propagating from the prediction output, whereas the latter measures the sensitivity of the prediction by occluding ligand atoms one at a time and observing the resulting change in the output. Based on these analyses, we further examined the spatial distances between high-contribution atoms and the predicted pocket-neighbor region to assess whether the ligand atoms emphasized by the model were concentrated in biochemically meaningful pocket-proximal regions. This analysis complements the structural case study by evaluating whether GADTI captures protein–ligand binding-related structural features from both prediction-contribution and spatial-proximity perspectives. The detailed settings and complete results of the attribution and occlusion sensitivity analyses are provided in Supplementary Table S8.

In order to further examine whether GADTI captures binding-relevant structural patterns under the scaffold-split setting, we selected a representative serine/threonine kinase LOK (STK10)–ligand pair as the core case study. This sample has the highest predictive confidence in the test set (ranked 1, probability: 0.896). It is worth noting that the ligand scaffold has never appeared in the training set, providing a useful evaluation case for examining the generalization ability of the model.

For this case, we designed and implemented a multi-dimensional structure verification experiment. First, we used

PyMOL (Schrödinger, LLC, 2015) to render the 3D complex structure used in the case analysis to evaluate its spatial fit. Second, we calculated the structural interaction fingerprint (SIFt) heat map based on the complex coordinates used in the case analysis to characterize the atomic contact intensity. Finally, we utilized RDKit to generate 2D pharmacophore topology diagrams to extract key chemical structural features. Experimental results suggest that the case analysis reflects the classic binding mode of the kinase.

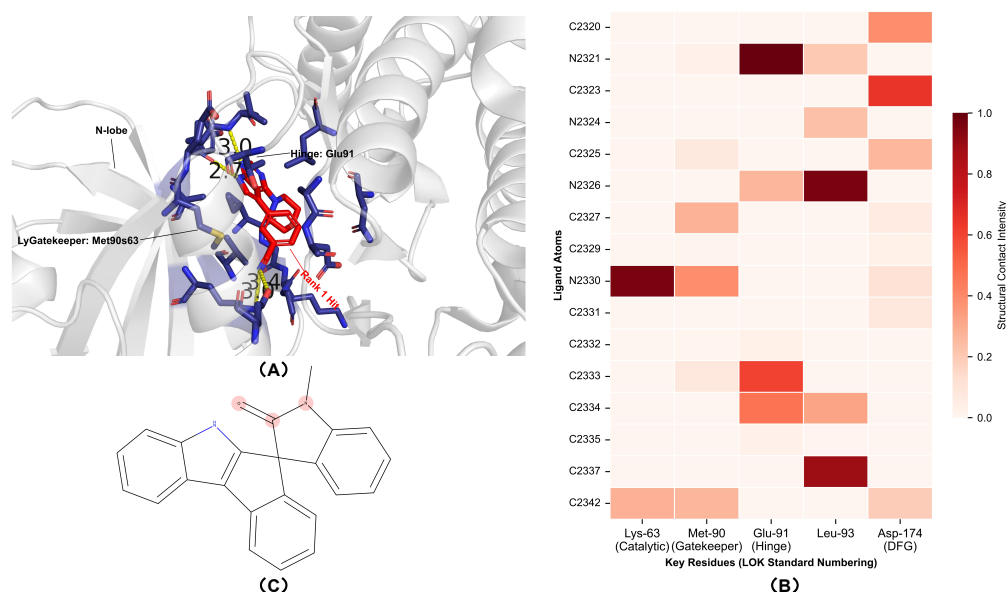


Figure 5. GADTI prediction and multi-dimensional validation of the STK10 kinase binding mode. (A) PyMOL rendering of the 3D complex binding mode, displaying hinge region hydrogen bonds and shape complementarity. (B) SIFt interaction heatmap. (C) RDKit-generated 2D pharmacophore topology diagram.

As shown in Figure 5A, in the complex structure used for the case analysis, the lactam ring of the ligand is oriented toward the hinge region, suggesting the formation of a potential bidentate hydrogen bond with the main chain of Glu-91. Concurrently, the visualized structure shows shape complementarity between the hydrophobic core of the ligand and the gatekeeper residue Met-90, which is consistent with reduced spatial steric hindrance. This physical binding visualization is further validated at the chemical semantic level in the 2D pharmacophore topology diagram (Figure 5C). The light red highlighted lactam group is explicitly identified as the classic hinge-binding module anchoring to Glu-91, while the light blue highlighted planar aromatic core serves as the hydrophobic scaffold, rationally explaining its tight hydrophobic contact with Met-90. As shown in Figure 5B, quantitative analysis via the SIFt heatmap further reveals that the model's geometric perception is not limited to the hinge region. The heatmap data indicate high-intensity structural contacts at Glu-91 and Met-90 (Intensity >0.8), as well as characteristic interactions involving the catalytic site Lys-63 and Asp-174 in the DFG motif. These observations provide qualitative evidence that GADTI, through E(3)-equivariant geometric constraints, may capture binding-relevant biophysical patterns underlying drug-target interactions. Consequently, the case analysis provides interpretable clues for activity prediction

even in the absence of homologous scaffold references.

4. Conclusion

In this study, we proposed GADTI, a geometry-aware dual-tower framework for drug–target interaction prediction. By adopting a pocket-centric global–local alignment strategy, GADTI maps full-length protein evolutionary semantics to local binding-pocket regions, thereby preserving global contextual information while reducing structural noise from non-binding regions. In parallel, GADTI integrates E(3)-equivariant geometric encoding with hierarchical bidirectional cross-attention, enabling the coordinated modeling of one-dimensional chemical semantics and three-dimensional spatial features.

The experimental results show that GADTI achieves competitive predictive performance on the Davis and BindingDB benchmark samples filtered for structural completeness. Ablation studies further indicate that sequence-level semantic features provide the fundamental discriminative information for the model, while the three-dimensional geometric branch offers additional spatial constraints in more challenging generalization scenarios such as scaffold-split evaluation. These findings suggest that E(3)-equivariant geometric encoding helps the model distinguish drug–target pairs with similar topologies but differing spatial compatibilities.

In summary, the significance of GADTI lies not only in its integration of pocket-centric alignment, dual-tower semantic–geometric fusion, and E(3)-equivariant geometric modeling within a unified DTI prediction framework, but also in its ability to provide a more structure-aware perspective for interaction prediction involving novel scaffold molecules. The experimental results indicate that sequence-level semantic information provides an essential basis for discriminating drug–target interactions, while three-dimensional geometric constraints offer additional spatial guidance when the model is evaluated on unseen ligand scaffolds. By focusing geometric modeling on predicted binding-pocket regions while preserving full-length protein semantics, GADTI reduces noise from non-binding regions and improves the alignment between chemical semantics and spatial compatibility. These observations suggest that physically consistent geometric information can help DTI models reduce scaffold-dependent memorization and better capture binding-relevant spatial patterns.

Nevertheless, the current method still has certain limitations. Because some data lack directly matched experimental complex structures, the model still relies on static structural proxies for the corresponding samples, such as AF2-predicted protein structures and RDKit-generated ligand conformers. As a result, it does not yet fully account for protein–ligand conformational changes, induced-fit effects, or ligand conformational diversity during the actual binding process. Future work will further incorporate protein–ligand flexibility modeling, including ligand-induced conformational adaptation and pocket side-chain flexibility, together with diffusion-based conformational sampling and structure-aware pretraining, to improve the applicability of the model in real-world drug discovery scenarios.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62402349 and 62302156), the Natural Science Foundation of Hubei Province (Grant No. 2024AFB127), and the Wuhan Textile University Foundation (Grant Nos. 20230612 and 2024309).

Data and Software Availability

The raw Davis and BindingDB datasets used in this study are publicly accessible. The curated datasets, including the optimized 3D drug conformations and the corresponding AF2 unbound protein structures, along with the source code for the GADTI framework, are freely available on GitHub at <https://github.com/sharpuser1122/GADTI>. Detailed documentation, including data preprocessing scripts, model training procedures, and inference guidelines, is provided within the repository to ensure full reproducibility.

References

- Ahamed, M.A., Ye, Q., Cheng, Q., 2025. Mol-cadiff: Causality-aware autoregressive diffusion for molecule generation. arXiv preprint arXiv:2503.05499 .
- Ba, J.L., Kiros, J.R., Hinton, G.E., 2016. Layer normalization. arXiv preprint arXiv:1607.06450 .
- Bagherian, M., Sabeti, E., Wang, K., Sartor, M.A., Nikolovska-Coleska, Z., Najarian, K., 2021. Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. *Briefings Bioinf.* 22, 247–269.
- Bai, P., Miljković, F., John, B., Lu, H., 2023. Interpretable bilinear attention network with domain adaptation improves drug-target prediction. *Nature Machine Intelligence* 5, 126–136. doi:10.1038/s42256-022-00605-1.
- Bemis, G.W., Murcko, M.A., 1996. The properties of known drugs. 1. molecular frameworks. *J. Med. Chem.* 39, 2887–2893.
- Chen, L., Tan, X., Ding, D., Zhong, F., Wang, X., He, X., Qi, X., Hou, T., 2021. Hidden bias in drug–target interaction prediction. *Bioinformatics* 37, 949–956.
- Chen, L., Tan, X., Wang, D., Zhong, F., Liu, X., Yang, T., Luo, X., Chen, K., Jiang, H., Hou, T., 2020. Transformer-CPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics* 36, 4406–4414.
- Chen, X., Yan, C.C., Zhang, X., Zhang, X., Dai, F., Yin, J., Zhang, Y., 2018. Machine learning for drug-target interaction prediction. *Molecules* 23, 2208. doi:10.3390/molecules23092208.

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

Coley, C.W., 2023. Defining the role of AI in drug discovery. *Cell Syst.* 14, 116–120.

Dauphin, Y.N., Fan, A., Auli, M., Grangier, D., 2017. Language modeling with gated convolutional networks, in: International Conference on Machine Learning (ICML), PMLR. pp. 933–941.

Davis, J., Goadrich, M., 2006. The relationship between precision-recall and ROC curves, in: Proceedings of the 23rd International Conference on Machine Learning, pp. 233–240.

Davis, M.I., Hunt, J.P., Herrgard, S., Ciceri, P., Wodicka, L.M., Pallares, G., Hocker, M., Treiber, D.K., Zarrinkar, P.P., 2011. Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* 29, 1046–1051.

DiMasi, J.A., Grabowski, H.G., Hansen, R.W., 2016. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. Health Econ.* 47, 20–33.

Elnaggar, A., Heinzinger, M., Dallago, C., Rhalmi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angermüller, C., Batzner, S., Kelly, R.V., Dunner, C., Chen, M., Mackowiak, L., Rost, B., 2021. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 7112–7127.

Gao, Z., Tan, C., Gao, L., Wu, Y., Lin, J., Zheng, X., Li, S.Z., 2024. A survey on deep learning for drug discovery. *IEEE Trans. Knowl. Data Eng.* 36, 666–688.

Halgren, T.A., 1996. Merck molecular force field. i. basis, form, scope, parameterization, and performance of MMFF94. *Journal of computational chemistry* 17, 490–519.

Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A., 1982. Evaluating the yield of medical tests. *JAMA* 247, 2543–2546.

Hendrycks, D., Gimpel, K., 2016. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*.

Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., Coley, C.W., Xiao, C., Sun, J., Zitnik, M., 2021a. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development, in: NeurIPS Datasets and Benchmarks Track.

Huang, K., Fu, T., Glass, L.M., Zitnik, M., Xiao, C., Sun, J., 2020. DeepPurpose: a deep learning library for drug-target interaction prediction. *Bioinformatics* 36, 5545–5547.

Huang, K., Xiao, C., Glass, L.M., Sun, J., 2021b. MolTrans: molecular interaction transformer for drug-target interaction prediction. *Bioinformatics* 37, 830–836.

582 Ji, Y., Zhang, L., Wu, J., Wu, B., Huang, L.K., Xu, T., Rong, Y., Li, L., Ren, J., Xue, D., Lai, H., Xu, S., Feng, J., Liu,
583 W., Luo, P., Zhou, S., Huang, J., Zhao, P., Bian, Y., 2023. DrugOOD: Out-of-distribution (OOD) dataset curator
584 and benchmark for AI-aided drug discovery, in: Proceedings of the AAAI Conference on Artificial Intelligence,
585 pp. 8023–8031.

586 Jiang, D., Hsieh, C.Y., Wu, J., Kang, Y., Wang, J., Wang, E., Liao, B., Shen, C., Xu, L., Wu, J., Cao, D., Hou, T.,
587 2020. InteractionGraph: protein–ligand interaction prediction using graph neural networks. *J. Med. Chem.* 63,
588 14451–14463.

589 Jing, B., Eismann, S., Suriana, P., Townshend, R.J., Dror, R., 2021. Learning from protein structure with geometric
590 vector perceptrons, in: International Conference on Learning Representations (ICLR).

591 Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek,
592 A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov,
593 S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M.,
594 Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D.,
595 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589.

596 Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

597 Koh, H.Y., Nguyen, A.T.N., Pan, S., May, L.T., Webb, G.I., 2024. Physicochemical graph neural network for learning
598 protein–ligand interaction fingerprints from sequence data. *Nature Machine Intelligence* 6, 673–687. doi:10.1038/
599 s42256-024-00847-1.

600 Krivák, R., Hoksza, D., 2018. P2Rank: knowledge-based ligand binding site prediction using aggregated local fea-
601 tures. *J. Cheminf.* 10, 1–24.

602 Landrum, G., Tosco, P., Kelley, B., Vianello, R., 2006. RDKit: Open-source cheminformatics. [Http://www.rdkit.org](http://www.rdkit.org).

603 Li, S., Wan, F., Shu, H., Jiang, T., Zhao, D., Zeng, J., 2022. Structure-aware interactive graph neural networks for the
604 prediction of protein–ligand binding affinity, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge
605 Discovery and Data Mining, pp. 975–984.

606 Li, Y., Rezaei, M.A., Li, C., Li, X., Qi, D., 2019. DeepAtom: a framework for protein–ligand binding affinity
607 prediction, in: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE. pp. 303–
608 310.

609 Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection, in: Proceedings of
610 the IEEE International Conference on Computer Vision, pp. 2980–2988.

Lin, X., Li, X., Lin, X., 2020. Deep learning for drug-target interaction prediction: current status and future challenges. *Curr. Opin. Struct. Biol.* 61, 152–158.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., Rives, A., 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379, 1123–1130.

Liu, S., Wang, H., Liu, W., Lasenby, J., Guo, H., Tang, J., 2022. Pre-training molecular graph representation with 3D geometry, in: *International Conference on Learning Representations (ICLR)*.

Liu, T., Lin, Y., Wen, X., Jorissen, R.N., Gilson, M.K., 2007. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res.* 35, D198–D201.

Liu, Z., Li, Y., Han, L., Li, J., Liu, J., Zhao, Z., Nie, W., Liu, Y., Wang, R., 2015. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* 31, 405–412.

Lu, W., Wu, Q., Zhang, J., Rao, J., Li, C., Zheng, S., 2022. TANKBind: Trigonometry-aware neural networks for drug-protein binding structure prediction, in: *Advances in Neural Information Processing Systems*, pp. 7236–7249.

Nguyen, T., Le, H., Quinn, T.P., Nguyen, T., Le, T.D., Venkatesh, S., 2021. GraphDTA: predicting drug–target binding affinity with graph neural networks. *Bioinformatics* 37, 1140–1147.

Öztürk, H., Özgür, A., Ozkirimli, E., 2018. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics* 34, i821–i829.

Riniker, S., Landrum, G.A., 2015. Better informed distance geometry: using what we know to generate 3D conformers and their ensembles. *J. Chem. Inf. Model.* 55, 2562–2574.

Schrödinger, LLC, 2015. The PyMOL Molecular Graphics System, Version 1.8.

Schütt, K., Unke, O., Gastegger, M., 2021. Equivariant message passing for the prediction of tensorial properties and molecular spectra, in: *International Conference on Machine Learning (ICML)*, PMLR. pp. 9377–9388.

Tsubaki, M., Tomii, K., Sese, J., 2019. Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics* 35, 309–318.

Wallach, I., Dzamba, M., Heifets, A., 2015. AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv preprint arXiv:1510.02855*.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., Barzilay, R., 2019. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* 59, 3370–3388.

- 640 Zhao, Q., Yang, M., Zhao, Y., Wang, W., Li, J., Kurgan, L., 2022. HyperAttentionDTI: improving drug–protein
641 interaction prediction by sequence-based deep learning with attention mechanism. *Bioinformatics* 38, 655–662.
- 642 Zhao, Y., Xing, Y., Zhang, Y., Wang, Y., Wu, C., 2025. Evidential deep learning-based drug-target interaction predic-
643 tion. *Nat. Commun.* 16, 6915.
- 644 Zhu, Z., Yao, Z., Qi, G., Mazur, N., Yang, P., Cong, B., 2023. Associative learning mechanism for drug-target
645 interaction prediction. *CAAI Transactions on Intelligence Technology* 8, 1558–1577. doi:10.1049/cit2.12194.

ORCID Information

Manuscript Title: GADTI: A Geometry-Aware Dual-Tower Framework for Drug-Target Interaction Prediction

Corresponding Authors:

Junlin Xu: <https://orcid.org/0000-0003-1057-1504>

Taoyang Wu: <https://orcid.org/0000-0002-2663-2001>