molecules

MDPI

Article

HierL2GDTA: Hierarchical Interaction Modeling with Local and Global Features for Drug–Target Affinity Prediction

Kaili Wang [1,*], Yaoxiang Yin [1], Yanglan Gan [1], Wentao Dai [2] and Zhijun Fang [3,*]

[1] School of Information and Intelligent Science, Donghua University, Shanghai 201620, China
[2] NHC Key Laboratory of Reproduction Regulation, Shanghai Institute for Biomedical and Pharmaceutical Technologies (SIBPT), Fudan University, Shanghai 200237, China
[3] Institute of Trustworthy Embodied AI, Fudan University, Shanghai 200433, China
* Correspondence: kailiwang@dhu.edu.cn (K.W.); zjfang@fudan.edu.cn (Z.F.)

Abstract

Accurate prediction of drug–target binding affinity is essential for drug discovery. However, existing methods face limitations in capturing local and global molecular features and in effectively integrating protein and ligand representations for interaction modeling. Here, we present HierL2GDTA, a sequence-based hierarchical framework for predicting binding affinity. Pretrained protein and molecular representations are processed by multi-scale convolutions to extract local features, followed by a bidirectional gated recurrent unit and dilated convolutions to capture broader protein and ligand context, respectively. A modified Transformer encoder integrates global protein context and ligand features through sequential cross-attention to model intramolecular and intermolecular dependencies. HierL2GDTA outperformed the evaluated baseline methods on both BindingDB and Davis. Further evaluations demonstrated the effectiveness of the model components and favorable predictive performance across selected kinase families. Attention visualization highlighted molecular substructures associated with protein binding, while retrospective screening ranked known inhibitors of disease-associated targets among the leading candidates. These results indicate that HierL2GDTA is an effective deep learning method for binding affinity prediction, with potential application for target-based drug screening.

Keywords: drug–target affinity prediction; drug discovery; sequence-based hierarchical framework; modified transformer encoder; intramolecular and intermolecular dependencies; deep learning; target-based drug screening

Check for updates

Academic Editor: Lantian Yao

Received: 26 September 2026
Revised: 23 October 2026
Accepted: 23 October 2026
Published: 26 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The interaction between proteins and small molecules underlies a wide range of biological processes and constitutes the molecular basis of many therapeutic interventions. Quantifying the strength of such interactions, generally referred to as drug–target binding affinity, is therefore a fundamental task in drug discovery [1]. Accurate binding affinity prediction can facilitate the identification and prioritization of promising compounds during virtual screening, guide lead optimization, and reduce the experimental burden associated with large-scale biochemical assays [2,3]. Experimental measurements, such as the dissociation constant K_d, inhibition constant K_i, and half-maximal inhibitory concentration *IC50*, provide reliable estimates of binding strength but are costly and time-consuming when applied to large chemical libraries. Molecular docking [4] and physics-based free-energy calculations provide computational alternatives, but their practical applications are

often constrained by the trade-off between computational efficiency and predictive accuracy [5,6]. Consequently, developing accurate and efficient computational approaches for drug–target binding affinity prediction remains an important challenge in computer-aided drug discovery.

With the rapid development of artificial intelligence, deep learning (DL) has been extensively applied to drug–target binding affinity prediction. According to the types of input representations, current DL methods for drug–target binding affinity prediction [7,8] can be primarily categorized into sequence-based and structure-based approaches.

Sequence-based approaches typically predict binding affinity from protein sequences and ligand SMILES strings without relying on drug–target complex structures [9,10]. For instance, DeepDTA [11] is an early sequence-based affinity model that employs two separate one-dimensional convolutional neural networks (CNNs) to learn representations from protein sequences and ligand SMILES strings. DeepAffinity [12] combined recurrent and convolutional neural networks within a unified framework to capture sequential patterns from protein and compound representations. AttentionDTA [13] used separate 1D CNNs to encode ligand and protein sequences, together with a two-sided multi-head attention mechanism for ligand–protein interaction modeling. More recently, pretrained protein and molecular language models have been increasingly employed in affinity prediction [14]. MutualDTA [15] employs pretrained molecular and protein models to derive context-aware representations, followed by downstream feature extraction and mutual attention-based interaction modeling. Similarly, GeneralizedDTA [16] incorporates pretraining tasks for both molecules and proteins and further fine-tunes the pretrained representations for affinity prediction.

Structure-based approaches explicitly incorporate molecular topology and three-dimensional structural information as input representations [17,18]. Initially, structure-based approaches commonly represented protein–ligand complexes as three-dimensional grids and applied 3D CNNs to learn spatial interaction patterns. Pafnucy [19], for example, encodes protein–ligand complexes into voxelized representations and extracts interaction features through three-dimensional convolutions. Subsequently, graph neural networks (GNNs) have become increasingly powerful and widely used in drug discovery and drug design. GraphDTA [20] represents ligands as molecular graphs and employs GNNs to learn their topological features, while CNNs are used to extract hidden features from protein sequences. Recently, complex-based methods directly construct graphs from 3D protein–ligand complexes. Specifically, PLANET [21] represents the binding pocket as a 3D graph to capture its spatial and structural information. Meanwhile, multimodal frameworks that integrate molecular information across multiple levels have also been increasingly applied to drug–target affinity prediction. In particular, GIF-PLA [22] combines meta-path-enhanced representations of heterogeneous protein–ligand graphs with sequence information to capture higher-order structural and semantic relationships.

Although existing models have achieved promising results in affinity prediction, they still face certain limitations. Structure-based approaches provide direct access to geometrical and intermolecular interaction information, but their performance depends on the availability and quality of the structures. The computational cost also increases with the complexity and level of detail of structural representations. Recent studies have shown that even geometrically plausible protein–ligand poses may fail to recover key intermolecular interactions [23]. Intermolecular Interaction Network Fidelity (I-INF) has been introduced to quantify the agreement between predicted and reference interchain interactions in macromolecular complexes [24], providing a useful basis for evaluating modeled interfaces and informing the development of structure-based interaction prediction models. Sequence-based approaches are more broadly applicable, as they do not explicitly depend on complex

structures. However, existing methods remain limited in modeling multi-scale local features and long-range dependencies among different local regions. In addition, the capacity to effectively integrate protein–ligand interaction information is still insufficient.

To address these challenges, we developed HierL2GDTA, a framework for drug–target binding affinity prediction that combines local and global feature extraction with hierarchical modeling of intramolecular and intermolecular interactions. Specifically, pre-trained protein and molecular language models are first employed to provide context-aware representations. Multi-scale convolutional modules with different receptive fields are subsequently introduced for both proteins and ligands to capture complementary local patterns at multiple contextual scales. To capture dependencies beyond local patterns, distinct encoding strategies are employed for proteins and ligands. A bidirectional gated recurrent unit (BiGRU) is employed to capture long-range sequential dependencies within proteins, whereas dilated convolutions are introduced for ligands to progressively enlarge the receptive field and characterize broader molecular context without excessively increasing network depth. Finally, a variant of the Transformer encoder is constructed to further refine and integrate the learned protein and ligand representations, allowing the model to capture high-level dependencies associated with intermolecular interactions. The overall architecture of the proposed model is illustrated in Figure 1.

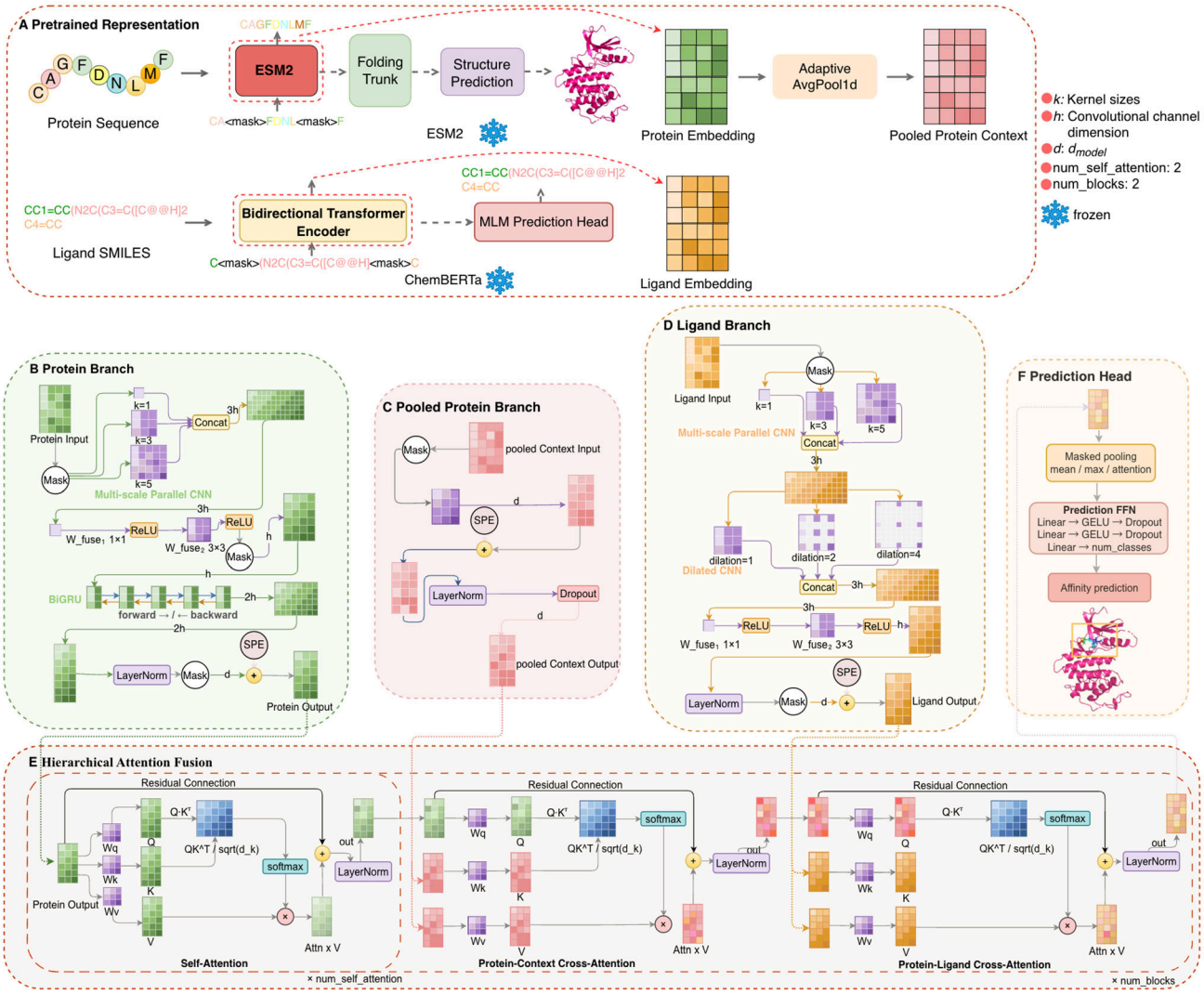


Figure 1. The framework of HierL2GDTA. (**A**) ESM-2 and ChemBERTa encode protein sequences and ligand SMILES, respectively; adaptive average pooling generates a fixed-length protein context.

(**B**) Multi-scale convolutions and a BiGRU extract local and long-range protein features, respectively. (**C**) The pooled protein context undergoes projection and positional encoding. (**D**) Multi-scale and dilated convolutions capture local and broader ligand context, respectively. (**E**) Self-attention models intramolecular dependencies, followed by protein-to-context and protein-to-ligand cross-attention for hierarchical fusion. (**F**) Masked attention pooling and an FFN produce the binding affinity prediction.

2. Results

2.1. Parameter Setting

All experiments were implemented in PyTorch 2.13.0 and conducted on a single NVIDIA RTX PRO 6000 GPU with 96 GB of memory. ESM-2 650M and ChemBERTa-77M-MTR were used as the pretrained protein and ligand encoders, with embedding dimensions of 1280 and 384, respectively. Each branch of the multi-scale convolutional modules used 640 output channels. The common representation dimension d_{model} was set to 384, and each multi-head attention layer used four heads. The dropout rate was set to 0.15.

The model was optimized using AdamW with a batch size of 150, a learning rate of 1×10^{-4}, and a weight decay of 1×10^{-4}. Mean squared error (MSE) was used as the loss function. Gradients were clipped to a maximum norm of 5.0. Training was run for up to 70 epochs on Davis and 100 epochs on BindingDB, with early-stopping patience values of 15 and 20 epochs, respectively. The model checkpoint with the lowest validation MSE was selected for final evaluation. The detailed parameter settings are summarized in Table S1.

2.2. Comparison with Competing Methods

HierL2GDTA was evaluated on the training, validation, and test sets of BindingDB and Davis (Table 1), showing comparable performance between the validation and test sets. We further compared HierL2GDTA with six representative drug–target binding affinity prediction methods, including DeepDTA, DeepAffinity, GraphDTA, MRBDTA, AttentionDTA, and MutualDTA. For fair comparisons, all baseline methods were trained, validated, and evaluated using the same data splits as HierL2GDTA, with identical test samples, affinity labels, and evaluation metrics. Table 2 presents the performance comparison on the BindingDB and Davis datasets in terms of MSE, mean absolute error (MAE), root mean squared error (RMSE), coefficient of determination (R^2), Pearson correlation coefficient (Pearson), and Spearman rank correlation coefficient (ρ). Lower MSE, MAE, and RMSE values indicate smaller prediction errors, whereas higher R^2, Pearson, and ρ values indicate greater explanatory power, stronger linear correlation, and stronger monotonic correlation, respectively. On the BindingDB dataset, HierL2GDTA achieved the best performance across all six metrics, with MSE, MAE, RMSE, R^2, Pearson, and ρ values of 0.568, 0.483, 0.754, 0.830, 0.917, and 0.890, respectively. Compared with the best competing result for each metric, HierL2GDTA reduced MSE, MAE, and RMSE by approximately 21.33%, 14.36%, and 11.29%, respectively, while increasing R^2, Pearson, and ρ by 0.046, 0.024, and 0.013, respectively. HierL2GDTA also achieved the best performance across all metrics on the Davis dataset, with corresponding values of 0.251, 0.278, 0.501, 0.688, 0.844, and 0.677. Compared with the strongest competing method, HierL2GDTA reduced MSE, MAE, and RMSE by approximately 11.62%, 4.47%, and 6.00%, respectively, while increasing R^2, Pearson, and ρ by 0.041, 0.024, and 0.014, respectively. Although the relative reductions in prediction errors were smaller on Davis, improvements were observed across all metrics on both benchmarks (Figure 2).

Table 1. Performance of HierL2GDTA on the training, validation, and test sets of BindingDB and Davis.

Dataset	Data Split	MSE ↓	MAE ↓	RMSE ↓	R^2 ↑	Pearson ↑	ρ ↑
BindingDB	Training	0.308	0.351	0.555	0.909	0.956	0.933
	Validation	0.621	0.504	0.788	0.817	0.909	0.882
	Test	0.568	0.483	0.754	0.830	0.917	0.890
Davis	Training	0.146	0.210	0.383	0.817	0.912	0.728
	Validation	0.260	0.276	0.510	0.679	0.842	0.680
	Test	0.251	0.278	0.501	0.688	0.844	0.677

↑ indicates that higher values are better, while ↓ indicates that lower values are better.

Table 2. Performance comparison of HierL2GDTA with baseline methods on the BindingDB and Davis test sets.

Dataset	Methods	MSE ↓	MAE ↓	RMSE ↓	R^2 ↑	Pearson ↑	ρ ↑
BindingDB	AttentionDTA	0.771	0.564	0.878	0.769	0.886	0.863
	DeepAffinity	0.796	0.615	0.892	0.761	0.875	0.855
	DeepDTA	0.826	0.636	0.909	0.752	0.877	0.850
	GraphDTA	0.722	0.565	0.850	0.784	0.893	0.877
	MRBDTA	0.765	0.601	0.875	0.771	0.885	0.864
	MutualDTA	0.759	0.571	0.871	0.772	0.880	0.866
	HierL2GDTA	**0.568**	**0.483**	**0.754**	**0.830**	**0.917**	**0.890**
Davis	AttentionDTA	0.284	0.291	0.533	0.647	0.820	0.663
	DeepAffinity	0.333	0.458	0.577	0.586	0.778	0.639
	DeepDTA	0.332	0.363	0.576	0.588	0.784	0.642
	GraphDTA	0.337	0.316	0.581	0.581	0.797	0.630
	MRBDTA	0.315	0.325	0.561	0.609	0.815	0.658
	MutualDTA	0.290	0.309	0.539	0.640	0.819	0.660
	HierL2GDTA	**0.251**	**0.278**	**0.501**	**0.688**	**0.844**	**0.677**

Within each metric column, the best value is shown in bold. ↑ indicates that higher values are better, while ↓ indicates that lower values are better.

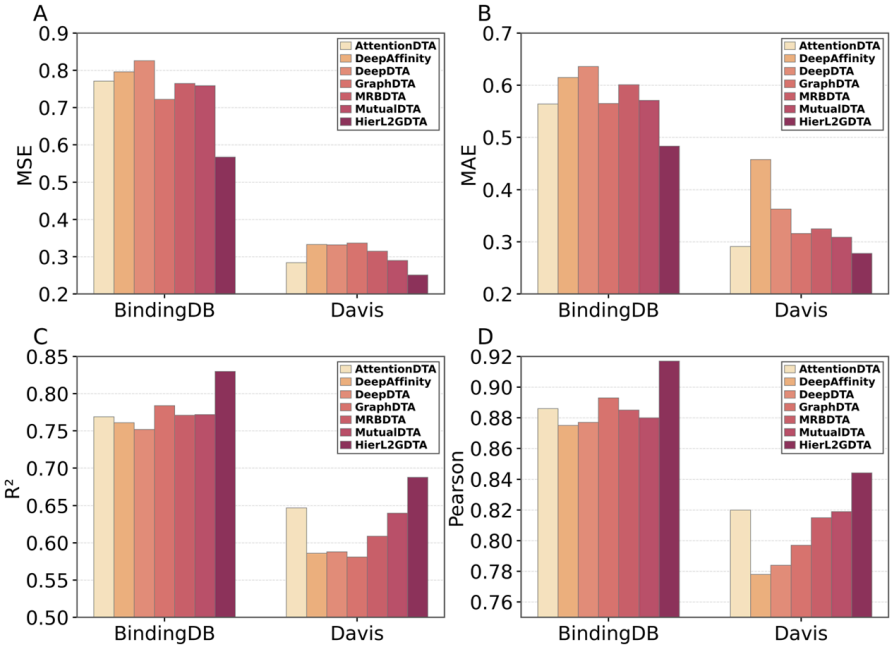


Figure 2. Performance of HierL2GDTA and the competitive methods on BindingDB and Davis test sets for four metrics: (**A**) MSE, (**B**) MAE, (**C**) R^2, and (**D**) Pearson.

To further assess the robustness and generalization of HierL2GDTA, we conducted additional experiments using repeated random splits and a protein-based cold split. Specifically, three independent experiments were conducted using different random splits, and the mean and standard deviation of each performance metric across the three runs are reported in Table 3 to assess model stability across different data partitions. The corresponding training and validation loss curves are presented in Figure S1. In addition, a protein-based cold split was adopted to assess the model's predictive performance on unseen protein targets, as presented in Table 4. For this evaluation, the core framework was retained, while two hyperparameters were adjusted: the maximum protein input length was set to 1024, and masked mean pooling was employed. The random seed was set to 2029, and all other experimental settings remained unchanged. Overall, these experiments further demonstrate the robustness of HierL2GDTA across different random splits and its predictive capability on unseen protein targets.

Table 3. Performance comparison of HierL2GDTA with baseline methods across three independent random splits on BindingDB and Davis.

Dataset	Methods	MSE ↓	MAE ↓	RMSE ↓	R^2 ↑	Pearson ↑	ρ ↑
BindingDB	AttentionDTA	0.661 (0.096)	0.509 (0.047)	0.812 (0.058)	0.803 (0.029)	0.900 (0.012)	0.880 (0.015)
	DeepAffinity	0.769 (0.029)	0.609 (0.023)	0.877 (0.016)	0.771 (0.010)	0.879 (0.004)	0.858 (0.008)
	DeepDTA	0.786 (0.042)	0.625 (0.019)	0.886 (0.024)	0.766 (0.014)	0.879 (0.004)	0.855 (0.006)
	GraphDTA	0.753 (0.032)	0.610 (0.040)	0.868 (0.018)	0.775 (0.008)	0.885 (0.008)	0.865 (0.010)
	MRBDTA	0.755 (0.039)	0.619 (0.042)	0.869 (0.022)	0.775 (0.011)	0.883 (0.006)	0.854 (0.015)
	MutualDTA	0.749 (0.021)	0.603 (0.034)	0.865 (0.012)	0.777 (0.006)	0.883 (0.006)	0.863 (0.008)
	HierL2GDTA	**0.590 (0.020)**	**0.497 (0.013)**	**0.768 (0.013)**	**0.824 (0.005)**	**0.912 (0.004)**	**0.888 (0.002)**
Davis	AttentionDTA	0.270 (0.019)	0.261 (0.027)	0.520 (0.019)	0.661 (0.022)	0.840 (0.018)	0.668 (0.004)
	DeepAffinity	0.290 (0.039)	0.361 (0.085)	0.537 (0.035)	0.636 (0.045)	0.803 (0.023)	0.652 (0.012)
	DeepDTA	0.302 (0.026)	0.338 (0.022)	0.549 (0.023)	0.621 (0.029)	0.799 (0.013)	0.650 (0.013)
	GraphDTA	0.300 (0.035)	0.325 (0.010)	0.547 (0.032)	0.624 (0.040)	0.804 (0.009)	0.640 (0.011)
	MRBDTA	0.297 (0.032)	0.318 (0.006)	0.544 (0.029)	0.627 (0.038)	0.805 (0.022)	0.652 (0.010)
	MutualDTA	0.286 (0.003)	0.300 (0.008)	0.535 (0.003)	0.640 (0.001)	0.813 (0.006)	0.656 (0.007)
	HierL2GDTA	**0.257 (0.015)**	0.279 (0.006)	**0.507 (0.014)**	**0.677 (0.019)**	0.834 (0.009)	**0.673 (0.004)**

Bold and underlined values indicate the best and second-best results, respectively. The standard deviations are given in parentheses. ↑ indicates that higher values are better, while ↓ indicates that lower values are better.

Table 4. Performance comparison of HierL2GDTA with baseline methods under a protein-based cold split on BindingDB and Davis.

Dataset	Methods	MSE ↓	MAE ↓	RMSE ↓	R^2 ↑	Pearson ↑	ρ ↑
BindingDB	AttentionDTA	1.856	0.967	1.362	0.312	0.577	0.597
	DeepAffinity	2.162	1.101	1.470	0.199	0.482	0.505
	DeepDTA	2.005	1.007	1.416	0.257	0.550	0.557
	GraphDTA	2.302	1.124	1.517	0.147	0.440	0.465
	MRBDTA	2.019	1.049	1.421	0.252	0.531	0.544
	MutualDTA	2.382	1.215	1.543	0.117	0.388	0.397
	HierL2GDTA	**1.850**	**0.947**	**1.360**	**0.315**	**0.608**	**0.600**
Davis	AttentionDTA	0.359	0.329	0.600	0.421	0.652	0.587
	DeepAffinity	0.464	0.447	0.681	0.253	0.503	0.435
	DeepDTA	0.444	0.424	0.666	0.284	0.534	0.484
	GraphDTA	0.570	0.458	0.755	0.081	0.523	0.455
	MRBDTA	0.484	0.436	0.695	0.220	0.514	0.463
	MutualDTA	1.147	0.991	1.071	−0.849	0.411	0.356
	HierL2GDTA	**0.348**	0.332	**0.590**	**0.440**	**0.683**	**0.592**

Bold and underlined values indicate the best and second-best results, respectively. ↑ indicates that higher values are better, while ↓ indicates that lower values are better.

2.3. Ablation Study

Eleven ablation experiments were conducted to assess the contributions of key components in HierL2GDTA (Figure 3). Three experiments evaluated the cross-attention modules: w/o cross_attn removed both modules, whereas w/o ligand_attn and w/o context_attn removed protein-to-ligand and protein-to-context cross-attention, respectively. To assess the contributions of pretrained representations, w/o pretraining replaced both protein and ligand representations with one-hot encodings, while w/o pretrain_protein and w/o pretrain_ligand applied the same replacement to protein and ligand representations, respectively. The remaining five experiments separately removed the protein BiGRU (w/o GRU), ligand dilated convolutions (w/o dilated_conv), initial protein self-attention (w/o self_attn), protein multi-scale convolutions (w/o protein_conv), and ligand multi-scale convolutions (w/o ligand_conv). All other experimental settings remained unchanged.

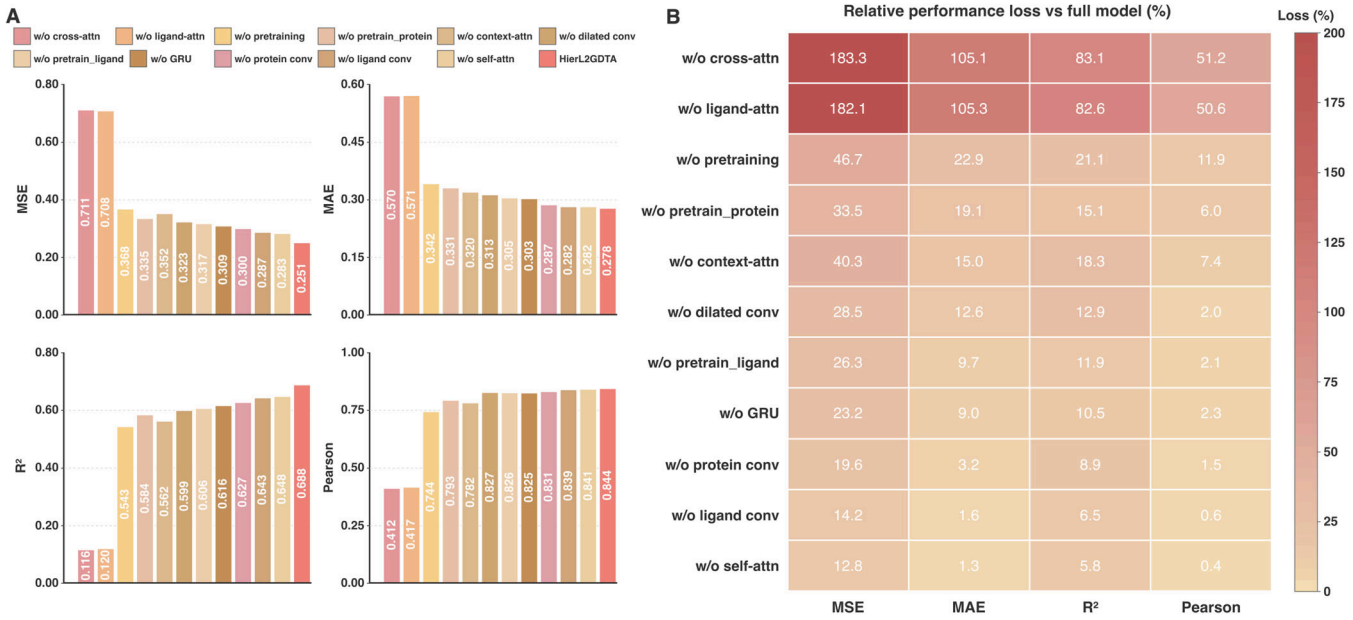


Figure 3. Ablation study of HierL2GDTA on the Davis dataset. (**A**) Performance comparison between the full model and eleven ablated variants. (**B**) Relative performance loss of each ablated variant compared with the full model. Darker shading indicates greater performance degradation.

The full model achieved the lowest MSE and MAE and the highest R^2 and Pearson correlation. Removing both cross-attention modules increased MSE from 0.251 to 0.711 and decreased Pearson from 0.844 to 0.412. Removing protein-to-ligand cross-attention alone resulted in a comparable decline, with an MSE of 0.708 and a Pearson of 0.417. Removing protein-to-context cross-attention also resulted in a notable performance decline, yielding an MSE of 0.352 and a Pearson of 0.782. Replacing both pretrained representations with one-hot encodings increased MSE to 0.368 and reduced Pearson to 0.744. Individually replacing protein or ligand pretrained representations increased MSE to 0.335 and 0.317, respectively, highlighting the importance of both representations. Removing the remaining components caused smaller performance decreases, with MSE values ranging from 0.283 to 0.323. These results indicate that all evaluated components contribute to predictive performance, particularly the two cross-attention modules and pretrained representations.

2.4. Parameter Analysis

We conducted a parameter sensitivity analysis on the Davis dataset to evaluate the effects of fusion depth, cross-attention layer allocation, and representation dimension d_{model} (Table 5). Two fusion blocks achieved the best overall performance. Fewer blocks reduced

predictive performance, whereas additional blocks yielded no consistent improvement. Increasing the number of protein-to-context or protein-to-ligand cross-attention layers individually provided no further benefit.

Table 5. Results of the parameter analysis.

Parameters	MSE ↓	MAE ↓	R^2 ↑	Pearson ↑
One fusion blocks	0.271	0.289	0.664	0.839
Three fusion blocks	0.298	0.294	0.629	0.826
Four fusion blocks	0.253	0.269	0.686	0.838
Fusion block (1, 1, 2)	0.272	0.274	0.662	0.824
Fusion block (1, 2, 1)	0.269	0.267	0.665	0.831
d_{model} = 192	0.276	0.287	0.656	0.843
d_{model} = 288	0.257	0.266	0.680	0.838
d_{model} = 480	0.253	**0.264**	0.686	0.841
HierL2GDTA	**0.251**	0.278	**0.688**	**0.844**

The default HierL2GDTA configuration consists of two fusion blocks, a (1, 1, 1) fusion configuration, and d_{model} = 384. The best value in each metric column is shown in bold. ↑ indicates that higher values are better, while ↓ indicates that lower values are better.

Performance remained relatively stable across moderate representation dimensions but declined at smaller dimensions. A dimension of 384 provided the best overall balance across metrics, with no consistent improvement at larger dimensions. Accordingly, the final model used two fusion blocks, each containing one protein-to-context and one protein-to-ligand cross-attention layer, with d_{model} = 384.

2.5. Substructure-Aware Drug Feature Extraction

The interpretability of the model was further assessed through atom-level attention visualization (Figure 4). Four protein–ligand pairs were selected from the Davis and BindingDB test sets. For each pair, the 15 highest-ranked ligand atoms were examined in relation to the corresponding binding-site interactions (Table S2). Ligand atom importance scores were derived from the attention weights of the protein-to-ligand cross-attention module. The importance maps highlighted several functional groups and aromatic regions involved in protein binding, providing a structural perspective on the model predictions.

In the carbonic anhydrase II–FSB complex (Figure 4A), high-scoring atoms were located in the sulfonamide group (atoms 0–3) and the attached benzene ring (atoms 4–7, 19, and 20). The sulfonamide group participates in two hydrogen bonds with Thr199, whereas the benzene ring forms a π–π contact with His94. The correspondence therefore extended to both polar and aromatic interactions. A similar pattern was observed for ABL2–imatinib (Figure 4B), where several highly ranked atoms belonged to the substituted 2-aminopyrimidine moiety (atoms 24–29 and 36) and the amide linker (atoms 5–7). These regions participate in hydrogen bonding with Thr361, Glu332, and Asp427, connecting the highlighted substructures to specific contacts within the binding pocket.

Aromatic substructures were also emphasized in the ABL1–PD173955 complex (Figure 4C). All six carbon atoms of the benzene ring (atoms 12–17) were highly ranked, consistent with the cation–π contact identified between this ring and Lys271. Several high-scoring atoms were also located within the 3-(methylthio)anilino group (atoms 20–28), while two hydrogen bonds with Met318 involved the anilino–pyrimidine junction. In the KIT–sunitinib complex (Figure 4D), the highlighted 5-fluorooxindole moiety (atoms 17–27) was fully represented among the top 15 atoms. Its carbonyl oxygen O27 and lactam NH group at N25 participate in hydrogen bonds with Cys673 and Glu671, respectively, linking the importance profile to the hinge-binding region of the ligand.

Although not every atom with a high importance score could be directly associated with an annotated protein contact, as illustrated by C28 in PD173955 and sunitinib, several highlighted substructures were involved in protein–ligand interactions. The identified functional groups and aromatic moieties participate in hydrogen bonding and aromatic contacts, emphasizing the relevance of these substructures to molecular recognition. These observations support the model's ability to identify binding-relevant substructures, providing a chemically meaningful basis for interpreting its affinity predictions.

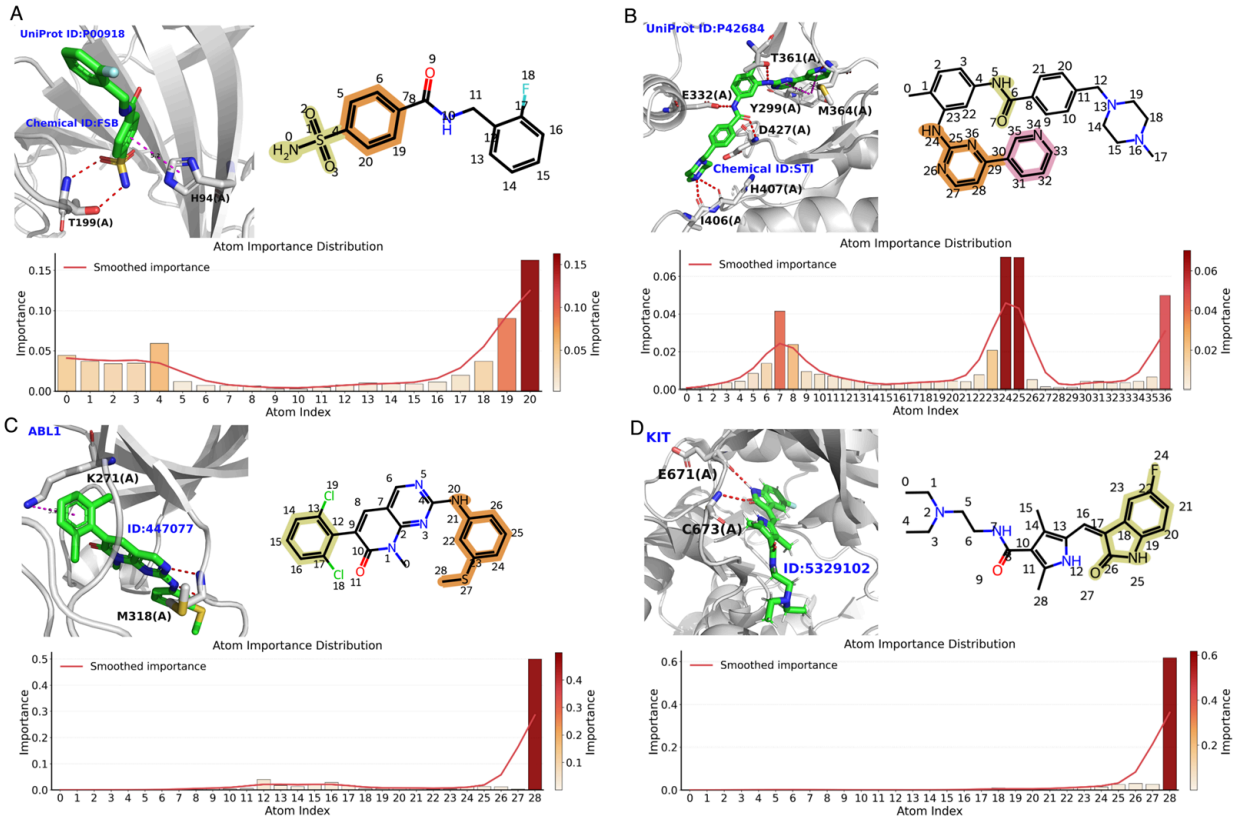


Figure 4. Atom-level attention visualization. Four representative protein–ligand pairs were selected from the BindingDB (**A**,**B**) and Davis (**C**,**D**) test sets: (**A**) P00918–FSB; (**B**) P42684–imatinib (STI); (**C**) ABL1–PD173955 (CID 447077); and (**D**) KIT–sunitinib (CID 5329102). Each panel shows the binding site, indexed ligand structure with highlighted substructures, and atom importance distribution. Proteins are shown in gray and ligands as green sticks; dashed lines indicate selected interactions with labeled residues. Bar heights and colors indicate attention-derived atom importance, with darker colors representing higher scores. Red curves show smoothed importance profiles.

2.6. Prediction Performance Across Protein Families

Binding affinity prediction was evaluated across kinase families in the Davis test set. Twelve families with at least 20 protein–ligand pairs were randomly selected for analysis (Figure 5). The selected families included EGFR, PDGFR, FGFR, JAK, class I PI3K, RET, CAMKL, CDK, STE20, Abl, RAF, and Src. Strong correlations were observed between predicted and experimental binding affinities, with Pearson correlation exceeding 0.85 for 11 of the 12 families (Table S3). Pearson correlations above 0.90 were obtained for class I PI3K, RET, FGFR, JAK, EGFR, PDGFR, and CAMKL. Among these families, EGFR, PDGFR, and CAMKL comprised 114, 131, and 116 test pairs, respectively, spanning 15, 18, and 19 kinase entries. Correlations of 0.885, 0.865, and 0.857 were obtained for CDK, Abl, and STE20, respectively. STE20 had the largest test subset among the selected families, comprising 188 drug–target pairs across 27 kinase entries, with an MSE of 0.202. For RAF, a high Pearson correlation of 0.904 was accompanied by a comparatively high

MSE of 0.469. The affinities of two high-affinity pairs were underestimated, and these predictions accounted for 75.7% of the total squared error, indicating that the elevated MSE was largely attributable to these individual predictions. Src exhibited the lowest Pearson correlation (0.784) among the selected families, with experimental pK_d values spanning a relatively narrow range of 5.000–7.824. These observations suggest that differences in experimental affinity distributions may affect the observed correlations across protein families, highlighting the need to interpret linear correlation coefficients in the context of the underlying data variance. Overall, the model showed favorable predictive performance across the selected kinase families, suggesting that it is useful for estimating drug–target binding affinities within these families.

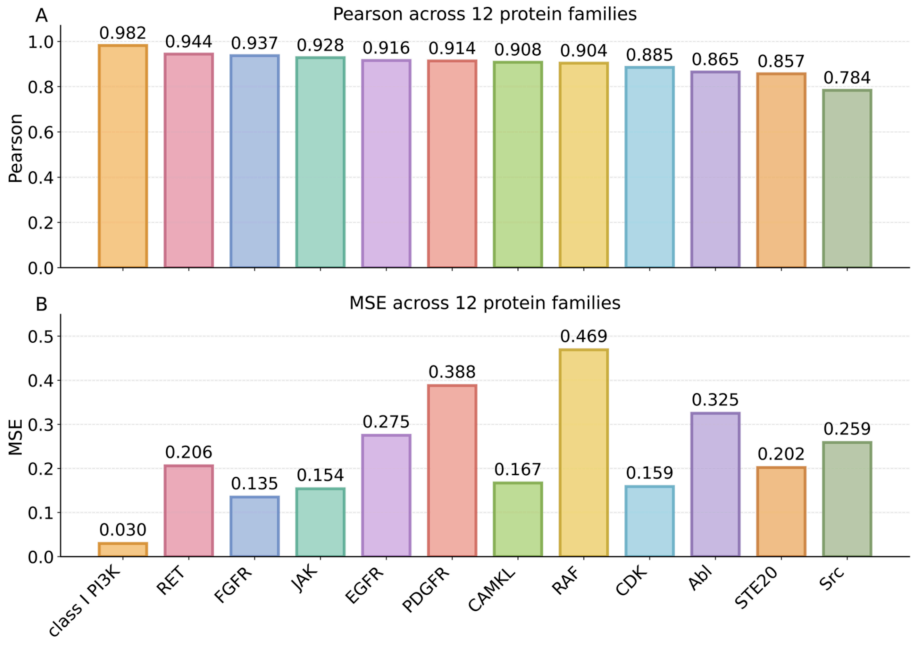


Figure 5. Binding affinity prediction performance of HierL2GDTA across 12 protein families: (**A**) Pearson's correlation and (**B**) MSE.

2.7. Compound Prioritization for Disease-Associated Targets

The utility of HierL2GDTA for disease-oriented compound prioritization was assessed in a screening analysis (Figure 6). All protein–ligand pairs included in this analysis were obtained from the Davis test set. Candidate compounds were ranked according to their predicted binding affinities within the corresponding target subsets (Table S4). EGFR and its activating variants are therapeutic targets in non-small-cell lung cancer (NSCLC) [25]. Among seven pairs involving EGFR(E746A750del), gefitinib was ranked first, with a predicted pK_d of 8.65 compared with an experimental value of 9.27 (Figure 6A). Similarly, erlotinib was ranked first among 22 pairs involving EGFR proteins harboring the L858R mutation (Figure 6B). Its predicted and experimental pK_d values were 8.45 and 9.01, respectively. Canertinib, an investigational pan-ErbB inhibitor previously evaluated in NSCLC clinical trials [26], ranked first among ten pairs involving EGFR(S752I759del). Its predicted pK_d of 9.69 closely matched the experimental value of 9.72 (Figure 6C). Across all 96 pairs involving EGFR and its variants, canertinib, gefitinib, and erlotinib were the three highest-ranked distinct compounds.

The same analysis was applied to ABL1, the kinase component of the BCR–ABL1 fusion protein that drives chronic myeloid leukemia (CML) [27]. Among seven ABL1–ligand pairs, imatinib was ranked second, with predicted and experimental pK_d values of 7.47 and 8.96, respectively (Figure 6D). The highest-ranked compound was PD173955 (CID

447077), an experimentally characterized Abl inhibitor [28], with corresponding values of 8.91 and 8.82. These compounds had the two highest experimental affinities in the subset, although their relative order was reversed in the predictions. The recovery of known inhibitors among the highest-ranked candidates supports the potential use of HierL2GDTA in prioritizing compounds for experimental screening against disease-associated targets.

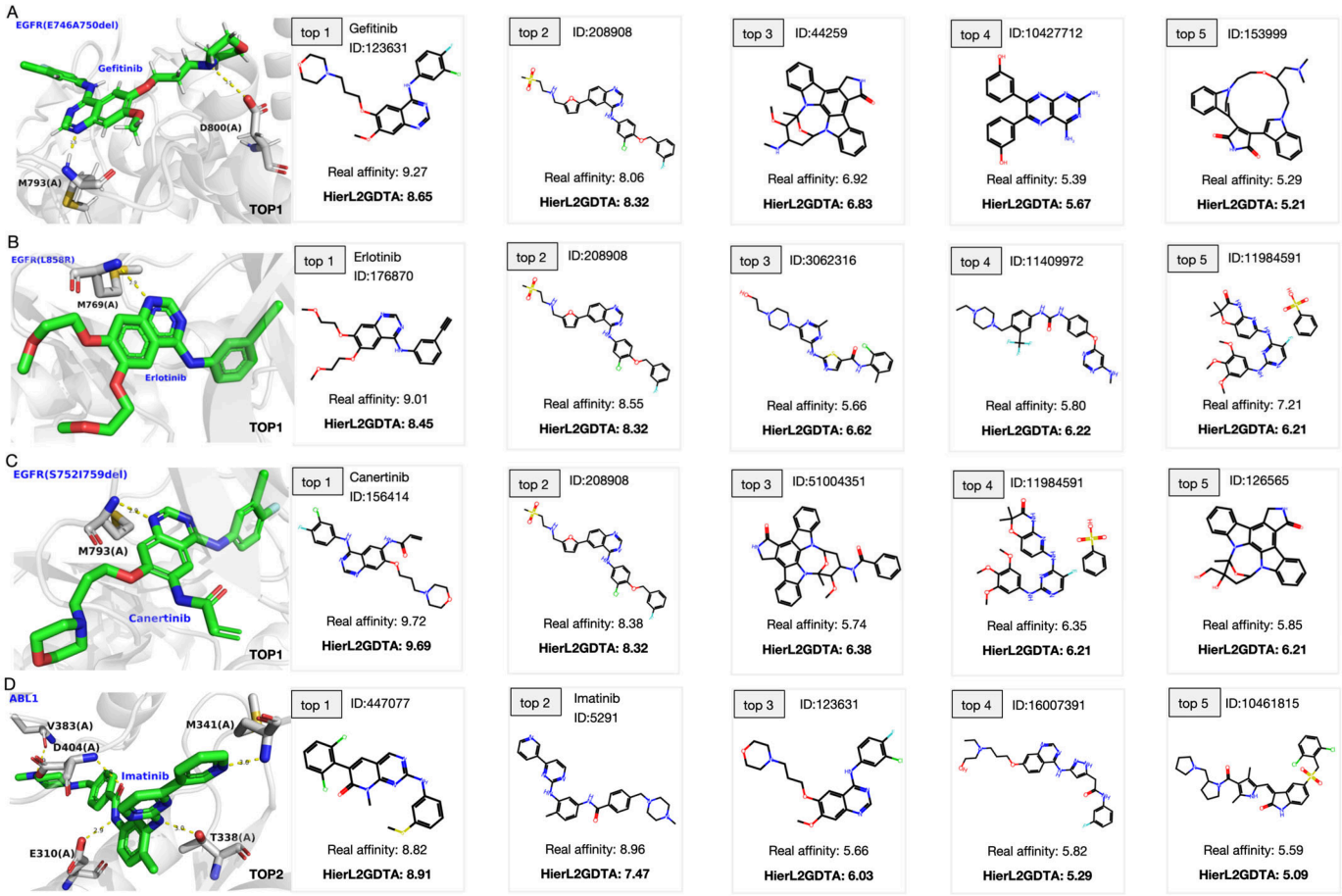


Figure 6. The five highest-ranked compounds predicted by HierL2GDTA within the Davis test involving (**A**) EGFR(E746A750del), (**B**) EGFR proteins harboring L858R, (**C**) EGFR(S752I759del), and (**D**) ABL1. Molecular structures, identifiers, and experimental and predicted (pK_d) values are shown. Binding-site views depict the first-ranked compounds gefitinib (**A**), erlotinib (**B**), and canertinib (**C**), and the second-ranked compound imatinib (**D**). Proteins are shown in gray and ligands as green sticks; dashed lines indicate selected interactions with labeled residues.

3. Discussion

In this study, HierL2GDTA was developed to integrate local and global sequence information for drug–target binding affinity prediction. The framework first generates protein and molecular representations using pretrained language models. These representations are then processed through multi-scale local feature extraction, followed by distinct encoding mechanisms to capture broader contextual dependencies. Hierarchical attention-based fusion subsequently integrates global protein information and ligand features into the protein representation, connecting intramolecular dependency modeling with intermolecular interaction learning.

Benchmark evaluations on BindingDB and Davis showed lower prediction errors and stronger correlations with experimental affinities than the evaluated baseline methods. Ablation analyses supported the contributions of the feature extraction and fusion components, with protein-to-ligand cross-attention making a substantial contribution to

predictive performance. Replacing both pretrained representations with one-hot encodings also reduced performance, supporting their joint use within the framework. Beyond benchmark evaluation, atom-level attention maps highlighted functional groups and aromatic substructures associated with protein binding, offering qualitative interpretations of the predictions. Retrospective screening further recovered known EGFR and ABL1 inhibitors among the highest-ranked candidates. Although the pretrained protein and molecular embeddings provide informative contextual representations, some physicochemical and binding-related information is only implicitly captured in the current framework. Incorporating complementary structural and interaction features in future work may enable a more comprehensive characterization of protein–ligand recognition. Interaction recovery metrics and adapted interface agreement measures will be investigated to assess the fidelity of modeled binding interfaces and explore their potential as prior-informed scoring signals to guide model training and optimization [23,24]. These extensions may facilitate the explicit modeling of key intermolecular interactions and further improve predictive performance. Together, these findings support hierarchical local-to-global representation learning for affinity prediction and indicate that HierL2GDTA can help prioritize compounds for screening against disease-associated targets.

4. Materials and Methods

4.1. Datasets

Two public drug–target affinity datasets, BindingDB [29] and Davis [30], were used to train and evaluate HierL2GDTA. Following the preprocessing protocol described previously [31], the original BindingDB dataset containing 3,048,285 records was processed as follows. First, records involving multiple protein chains were excluded, leaving 2,871,422 entries. Records with missing SMILES or protein sequences were then removed, resulting in 2,871,348 records. Next, K_d values were restricted to the range of 0.01–10^7 nM, with values reported using "<" or ">" approximated by dividing or multiplying the reported thresholds by 3.16, respectively, yielding 112,647 records. Filtering based on target organisms and assay temperatures retained 108,226 records. Specifically, records from mammalian and avian targets were retained, while viral targets were considered only if the assay temperature was between 33 and 40 °C. Furthermore, Protein sequences outside the range of 50–2560 residues were excluded, leaving 108,004 records. SMILES strings were subsequently canonicalized using RDKit (version 2026.09.1), and invalid SMILES or those exceeding 300 characters were removed, resulting in 106,738 records. Finally, duplicate protein–ligand records with identical SMILES and protein sequences were consolidated by taking the 20th percentile of their K_d measurements. The resulting BindingDB dataset contained 69,856 unique protein–ligand pairs. The Davis dataset was obtained directly from a publicly available preprocessed benchmark, comprising 30,056 kinase–inhibitor pairs involving 442 kinase targets and 68 inhibitors. For both datasets, the binding affinity labels were derived from K_d values. Specifically, K_d values expressed in nM were converted to pK_d using the following equation:

$$pK_d = -log_{10}\left(\frac{K_d}{10^9}\right) \quad (1)$$

The length distributions of the protein sequence and ligand SMILES strings were analyzed before model training. Protein sequences were limited to 1400 residues, covering approximately 95% of the proteins. Sequences exceeding this length were truncated. Given the relatively small variation in SMILES length compared with protein sequences, ligand SMILES strings were encoded without truncation. Table 6 summarizes the detailed dataset statistics. Data density was defined as the proportion of protein–ligand pairs with available

affinity labels among all possible combinations. BindingDB exhibited a highly sparse affinity matrix with a density of 0.16%, whereas Davis had complete affinity label coverage (100%). Given these differences in data density, the model was trained separately on each dataset to evaluate its performance across different data distributions. Following TEFDTA [32], drug–target pairs were randomly divided into training, validation, and test sets at a ratio of 7:2:1 using a fixed random seed.

Table 6. Summary of the datasets.

Datasets	Number of Proteins	Number of Ligands	Number of Interactions	Max Length for Proteins (Coverage)	Max Length for Ligands (Coverage)
BindingDB	1557	27,480	69,856	1400 (95%)	244 (100%)
Davis	442	68	30,056	1400 (95%)	81 (100%)

4.2. Model Construction

Given a protein–ligand pair, the protein sequence and ligand SMILES string are first encoded by the pretrained ESM-2 [33] and ChemBERTa [34] models, respectively. The resulting residue and token embeddings provide contextualized representations for subsequent feature extraction. Multi-scale local features are then extracted separately from the protein and ligand representations using three parallel 1D convolutional layers with kernel sizes of 1, 3, and 5. The outputs of the three convolutions are fused to integrate local patterns at different receptive fields. To further capture long-range dependencies, different feature extraction modules are designed for proteins and ligands. For ligand representation, three parallel dilated convolutions [35] with a kernel size of 3 and dilation rates of 1, 2, and 4 are applied. The different dilation rates enlarge the receptive fields and capture dependencies between distant SMILES tokens at multiple scales. For protein representation, a BiGRU [36] is employed to capture bidirectional dependencies among amino acid residues. Finally, a modified Transformer [37] encoder is designed for hierarchical intramolecular and intermolecular interaction learning. Specifically, sinusoidal positional encoding (SPE) is first added to the protein, pooled protein context, and ligand representations to preserve sequential information. The protein representation is then processed by multi-head self-attention to capture intramolecular dependencies among protein residues. The resulting features pass through two fusion blocks, each applying protein self-attention, protein-to-context cross-attention, and protein-to-ligand cross-attention in sequence. These operations progressively integrate residue-level features, pooled protein context, and ligand information. The final representation is aggregated by masked attention pooling and passed to a three-layer feed-forward neural network (FFN) for binding affinity prediction.

4.3. Pretrained Sequence Representation

Protein sequences and ligand SMILES strings are first encoded using the frozen pretrained ESM-2 and ChemBERTa models, respectively. The ESM-2 encoder contains 33 Transformer layers and 650 million parameters, producing 1280-dimensional contextualized residue embeddings. Given a protein sequence $P = \left(p_1, p_2, \ldots, p_{L_p}\right)$, its representation is expressed as:

$$E_P = ESM2(P) = \left[e_1^p, e_2^p, \ldots e_{L_p}^p\right] \quad (2)$$

where p_i denotes the amino acid residue at position i, and L_p is the protein sequence length. $e_i^p \in \mathbb{R}^{1280}$ represents the contextualized embedding of the i-th residue.

ChemBERTa is a Transformer-based molecular language model pretrained on large-scale SMILES data. Given a tokenized SMILES sequence $D = (d_1, d_2, \ldots, d_{L_d})$, ChemBERTa generates the token-level representation:

$$E_D = \text{ChemBERTa}(D) = \left[e_1^d, e_2^d, \ldots e_{L_d}^d\right] \tag{3}$$

where L_d denotes the number of SMILES tokens and $e_i^d \in \mathbb{R}^{384}$ represents the contextualized embedding of the i-th token.

A compact protein context is constructed by applying adaptive average pooling to the pretrained residue embeddings along the sequence dimension. This operation produces L_c = 16 context vectors, denoted by $C \in \mathbb{R}^{L_c \times 1280}$. After projection into the common feature space, these vectors serve as the keys and values in protein-to-context cross-attention.

4.4. Multi-Scale Local Feature Extraction

Separate convolutional modules extract local features from the protein and ligand embeddings. Each module consists of three parallel one-dimensional convolutions with kernel sizes of 1, 3, and 5. For an input representation $X \in \{E_P, E_D\}$, the feature map obtained using a convolutional kernel of size k is expressed as:

$$H_k = \sigma(Conv1D_k(X)), k \in \{1, 3, 5\} \tag{4}$$

where k denotes the convolutional kernel size, σ denotes the Gaussian Error Linear Units (GELU) activation function. Each convolution produces 640 feature channels. The resulting three feature maps, H_1,H_3,H_5, are subsequently concatenated along the channel dimension:

$$H_{multi} = Concat(H_1, H_3, H_5) \tag{5}$$

Accordingly, the three parallel convolutions capture complementary local patterns with different receptive fields. The kernel of size 1 transforms features independently at each position, whereas kernels of sizes 3 and 5 aggregate information from neighboring positions. Concatenating the branch outputs combines local sequence patterns captured at different receptive-field sizes. The concatenated representation therefore contains 1920 feature channels and is further processed for subsequent feature extraction.

4.5. Long-Range Dependency Modeling

To capture broader contextual information, protein and ligand features are further processed using a BiGRU and parallel dilated convolutions, respectively. The ligand features H_{multi}^D obtained from the multi-scale convolutional module are processed by three parallel one-dimensional dilated convolutions, with a kernel size K of 3 and dilation rates d of 1, 2, and 4.

$$H_{multi}^D = \left[h_1^D, h_2^D, \ldots, h_{L_d}^D\right] \in \mathbb{R}^{L_d \times C_{in}} \tag{6}$$

$$h_i^{(d)} = \sigma\left(\sum_{m=0}^{K-1} W_m^{(d)} h_{i+d(m-\lfloor K/2 \rfloor)}^D + b^{(d)}\right) \tag{7}$$

where $h_i^{(d)}$ denotes the output feature vector at position i for dilated rate d, C_{in} is the input feature dimension. $W_m^{(d)}$ is the weight matrix, $b^{(d)}$ is the bias term, and σ denotes the activation function. Appropriate padding is applied at sequence boundaries to preserve the sequence length. The receptive field of dilated convolution is calculated as:

$$R = 1 + (K - 1)d \tag{8}$$

For a kernel size of 3, dilation rates of 1, 2, and 4 correspond to receptive-field spans of 3, 5, and 9 positions in the input feature sequence, respectively. The parallel branches therefore aggregate information over different sequence ranges without increasing the kernel size. Each of the three parallel dilated convolutions produces features with 640 channels. The outputs are concatenated along the channel dimension as follows:

$$H^{(d)} = DilatedConv_{K=3,d}\left(H^{D}_{multi}\right), d \in \{1,2,4\} \tag{9}$$

$$H_{dilat} = Contat(H^{(1)},H^{(2)},H^{(4)}) \tag{10}$$

where $H^{(d)} = \left[h_1^{(d)}, h_2^{(d)}, \ldots, h_{L_d}^{(d)}\right]$. The concatenated features H_{dilat}, with a feature dimension of 1920, are subsequently projected to a common representation dimension of 384 for downstream interaction modeling.

For protein encoding, the BiGRU processes the residue features in both sequence directions to model long-range dependencies. The hidden states are updated through reset and update gates, as defined in Equations (11)–(14):

$$z_i = f(W_z x_i + U_z h_{i-1} + b_z) \tag{11}$$

$$r_i = f(W_r x_i + U_r h_{i-1} + b_r) \tag{12}$$

$$\widetilde{h}_i = tanh\left[W_h x_i + U_h\left(r_i \bigodot h_{i-1}\right) + b_h\right] \tag{13}$$

$$h_i = (1 - z_i) \bigodot h_{i-1} + z_i \bigodot \widetilde{h}_i \tag{14}$$

here, x_i denotes the input feature of the residue at position i, h_{i-1} is the hidden state from the previous position, and h_i is the updated hidden state. z_i and r_i represent the update and reset gates, respectively, while $\widetilde{h}_i$ denotes the candidate hidden state. W_z, W_r, W_h, U_z, U_r, and U_h are weight matrices. b_z, b_r, and b_h denote the corresponding bias vectors. f and $tanh$ denote the sigmoid and hyperbolic tangent activation functions, respectively, and $\odot$ represents element-wise multiplication. In the bidirectional architecture, the protein sequence is processed independently in the forward and backward directions:

$$\overrightarrow{h}_i = GRU_f\left(x_i, \overrightarrow{h}_{i-1}\right) \tag{15}$$

$$\overleftarrow{h}_i = GRU_b\left(x_i, \overleftarrow{h}_{i+1}\right) \tag{16}$$

where $\overrightarrow{h}_i$ and $\overleftarrow{h}_i$ denote the 640-dimensional forward and backward hidden states, respectively. Finally, the two hidden states are concatenated to obtain a bidirectional contextual representation of dimension 1280 at position i.

4.6. Hierarchical Intramolecular and Intermolecular Interaction Learning

The protein, ligand, and pooled protein-context features are independently projected into a common feature space of dimension d = 384, followed by layer normalization and dropout. The resulting representations are denoted as $P \in \mathbb{R}^{L_p \times d}$, $D \in \mathbb{R}^{L_d \times d}$, and $C \in \mathbb{R}^{L_c \times d}$, respectively. SPE is then added to each representation:

$$\widetilde{P} = P + SPE_p, \ \widetilde{C} = C + SPE_C, \ \widetilde{D} = D + SPE_d \tag{17}$$

The protein representation is first processed by two stacked self-attention layers to model dependencies among residues. Queries, keys, and values are obtained through

separate learnable projections, and multi-head scaled dot-product attention is applied with a protein padding mask.

$$Q_p = \widetilde{P}W_Q^p, \;\; K_p = \widetilde{P}W_K^p, \;\; V_p = \widetilde{P}W_V^p \tag{18}$$

$$Attn(Q_p, K_p, V_p) = softmax\left(\frac{Q_p K_p^T}{\sqrt{d_k}} + M_p\right)V_p \tag{19}$$

$$MHA(P) = Concat(head_1, \ldots, head_h)W_O \tag{20}$$

where W_Q^p, W_K^p, W_V^p, W_O are the projection matrices. d_k denotes the dimension of each attention head. M_p is the protein padding mask. h denotes the number of attention heads.

The resulting protein features are processed by two stacked hierarchical fusion blocks, each comprising three attention layers applied sequentially: protein self-attention, protein-to-context cross-attention, and protein-to-ligand cross-attention. Following self-attention, the updated protein representation P_s serves as the query, while the pooled protein context $\widetilde{C}$ provides the keys and values:

$$Q_c = P_s W_Q^c, \;\; K_c = \widetilde{C}W_K^c, \;\; V_c = \widetilde{C}W_V^c \tag{21}$$

$$Attn_c = softmax\left(\frac{Q_c K_c^T}{\sqrt{d_k}} + M_c\right)V_c \tag{22}$$

where W_Q^c, W_K^c, W_V^c are the projection matrices. M_c denotes the padding mask for the pooled protein context. This asymmetric attention mechanism enables each residue-level protein feature to selectively retrieve complementary information from the compressed protein context.

The context-enhanced protein features P_c subsequently serve as queries for protein-to-ligand cross-attention, with ligand features providing the keys and values. Each attention sublayer uses a residual connection followed by layer normalization [38,39]. This operation allows each protein position to selectively aggregate ligand information, thereby modeling intermolecular dependencies. The final hierarchical feature representation is aggregated using masked attention pooling [40], which computes a weighted sum over valid residue positions:

$$e_i = W_b^T tanh\left(W_a h_i^H + b_a\right) \tag{23}$$

$$\alpha_i = \frac{exp(e_i)}{\sum_{j:m_j=1} exp(e_j)}, \quad m_i = 1 \tag{24}$$

$$h_{init}^H = \sum_{i:m_i=1} \alpha_i h_i^H \tag{25}$$

where h_i^H denotes the i-th feature vector in the intermolecular hierarchical representation, W_a, W_b, and b_a are learnable parameters, m_i and m_j are binary indicators of valid residues at positions i and j, respectively. All self-attention and cross-attention layers in the hierarchical interaction learning module have an output feature dimension of 384.

Finally, the intermolecular feature representation with a dimension of 384 is fed into a three-layer FFN to predict drug–target binding affinity.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/molecules31213856/s1, Figure S1: Training and validation loss curves of HierL2GDTA across three independent random splits on BindingDB and Davis; Table S1: Model parameters; Table S2: Ligand atom importance rankings and binding interactions for four representative protein–ligand pairs; Table S3: Binding affinity prediction performance of HierL2GDTA

across 12 protein families in the Davis test set; Table S4: Predicted and experimental binding affinities of compounds evaluated against selected EGFR variants and ABL1 in the Davis test set.

Author Contributions: Conceptualization, K.W.; methodology, K.W.; software, Y.Y.; validation, K.W. and Y.Y.; formal analysis, K.W. and Y.Y.; investigation, Y.Y.; resources, Y.G.; data curation, Y.Y.; writing—original draft preparation, K.W.; writing—review and editing, K.W. and Z.F.; visualization, K.W. and Y.Y.; supervision, K.W.; project administration, K.W., W.D. and Z.F.; funding acquisition, K.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (No. 62402099, No. 62572114) and the Shanghai Sailing Program (No. 24YF2701500).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All datasets used in this study were obtained from publicly available repositories. The BindingDB dataset (release 202604) was obtained from https://www.bindingdb.org/ (accessed on 20 October 2026), and the Davis dataset was obtained from the PaddleHelix distribution at https://baidu-nlp.bj.bcebos.com/PaddleHelix/datasets/dti_datasets/davis_v1.tgz (accessed on 20 October 2026). The resource codes are available at https://github.com/WangKLLab/HierL2GDTA (accessed on 20 October 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Gilson, M.K.; Zhou, H.-X. Calculation of Protein-Ligand Binding Affinities. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42. [CrossRef] [PubMed]
2. Kairys, V.; Baranauskiene, L.; Kazlauskiene, M.; Matulis, D.; Kazlauskas, E. Binding affinity in drug design: Experimental and computational techniques. *Expert Opin. Drug Discov.* **2019**, *14*, 755–768. [CrossRef] [PubMed]
3. Wang, D.D.; Zhu, M.; Yan, H. Computationally predicting binding affinity in protein–ligand complexes: Free energy-based simulations and machine learning-based scoring functions. *Brief. Bioinform.* **2021**, *22*, bbaa107. [CrossRef] [PubMed]
4. Kitchen, D.B.; Decornez, H.; Furr, J.R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat. Rev. Drug Discov.* **2004**, *3*, 935–949. [CrossRef] [PubMed]
5. Mobley, D.L.; Gilson, M.K. Predicting Binding Free Energies: Frontiers and Benchmarks. *Annu. Rev. Biophys.* **2017**, *46*, 531–558. [CrossRef] [PubMed]
6. Cournia, Z.; Allen, B.; Sherman, W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *J. Chem. Inf. Model.* **2017**, *57*, 2911–2937. [CrossRef] [PubMed]
7. Wang, K.; Zhou, R.; Li, Y.; Li, M. DeepDTAF: A deep learning method to predict protein–ligand binding affinity. *Brief. Bioinform.* **2021**, *22*, bbab072. [CrossRef] [PubMed]
8. Lodh, E.; Majumder, S.; Chowdhury, T.; De, M. RLBindDeep: A ResNet-LSTM based novel framework for protein–ligand binding affinity prediction. *J. Mol. Graph. Model.* **2026**, *144*, 109282. [CrossRef] [PubMed]
9. Zhang, L.; Wang, C.-C.; Chen, X. Predicting drug–target binding affinity through molecule representation block based on multi-head attention and skip connection. *Brief. Bioinform.* **2022**, *23*, bbac468. [CrossRef] [PubMed]
10. Zhu, Z.; Yao, Z.; Qi, G.; Mazur, N.; Yang, P.; Cong, B. Associative learning mechanism for drug-target interaction prediction. *CAAI Trans. Intell. Technol.* **2023**, *8*, 1558–1577. [CrossRef]
11. Öztürk, H.; Özgür, A.; Ozkirimli, E. DeepDTA: Deep drug–target binding affinity prediction. *Bioinformatics* **2018**, *34*, i821–i829. [CrossRef] [PubMed]
12. Karimi, M.; Wu, D.; Wang, Z.; Shen, Y. DeepAffinity: Interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics* **2019**, *35*, 3329–3338. [CrossRef] [PubMed]
13. Zhao, Q.; Duan, G.; Yang, M.; Cheng, Z.; Li, Y.; Wang, J. AttentionDTA: Drug–target binding affinity prediction by sequence-based deep learning with attention mechanism. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2022**, *20*, 852–863. [CrossRef] [PubMed]
14. Bidgoli, A.H.; Mahdavi, M.; Malek, H. Structure-free drug–target affinity prediction using protein and molecule language models. *J. Cheminformatics* **2026**, *18*, 21. [CrossRef] [PubMed]
15. Yuan, Y.; Chen, S.; Hu, R.; Wang, X. MutualDTA: An Interpretable Drug–Target Affinity Prediction Model Leveraging Pretrained Models and Mutual Attention. *J. Chem. Inf. Model.* **2025**, *65*, 1211–1227. [CrossRef] [PubMed]

16. Lin, S.; Shi, C.; Chen, J. GeneralizedDTA: Combining pre-training and multi-task learning to predict drug-target binding affinity for unknown drug discovery. *BMC Bioinform.* **2022**, *23*, 367. [CrossRef] [PubMed]
17. Yang, Z.; Zhong, W.; Lv, Q.; Dong, T.; Yu-Chian Chen, C. Geometric Interaction Graph Neural Network for Predicting Protein–Ligand Binding Affinities from 3D Structures (GIGN). *J. Phys. Chem. Lett.* **2023**, *14*, 2020–2033. [CrossRef] [PubMed]
18. Kyro, G.W.; Smaldone, A.M.; Shee, Y.; Xu, C.; Batista, V.S. T-ALPHA: A Hierarchical Transformer-Based Deep Neural Network for Protein–Ligand Binding Affinity Prediction with Uncertainty-Aware Self-Learning for Protein-Specific Alignment. *J. Chem. Inf. Model.* **2025**, *65*, 2395–2415. [CrossRef] [PubMed]
19. Stepniewska-Dziubinska, M.M.; Zielenkiewicz, P.; Siedlecki, P. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics* **2018**, *34*, 3666–3674. [CrossRef] [PubMed]
20. Nguyen, T.; Le, H.; Quinn, T.P.; Nguyen, T.; Le, T.D.; Venkatesh, S. GraphDTA: Predicting drug–target binding affinity with graph neural networks. *Bioinformatics* **2021**, *37*, 1140–1147. [CrossRef] [PubMed]
21. Zhang, X.; Gao, H.; Wang, H.; Chen, Z.; Zhang, Z.; Chen, X.; Li, Y.; Qi, Y.; Wang, R. Planet: A multi-objective graph neural network model for protein–ligand binding affinity prediction. *J. Chem. Inf. Model.* **2024**, *64*, 2205–2220. [CrossRef] [PubMed]
22. Zhu, Y.; Wang, C.; Wang, J.; Zhao, L.; Zou, Q. Structure-Aware Heterogeneous Information Fusion Framework for Protein–Ligand Binding Affinity Prediction. *J. Chem. Inf. Model.* **2026**, *66*, 503–517. [CrossRef] [PubMed]
23. Errington, D.; Schneider, C.; Bouysset, C.; Dreyer, F.A. Assessing interaction recovery of predicted protein-ligand poses. *J. Cheminformatics* **2025**, *17*, 76. [CrossRef] [PubMed]
24. Ludwiczak, O.; Antczak, M.; Szachniuk, M. Assessing interface accuracy in macromolecular complexes. *PLoS ONE* **2025**, *20*, e0319917. [CrossRef] [PubMed]
25. Lynch, T.J.; Bell, D.W.; Sordella, R.; Gurubhagavatula, S.; Okimoto, R.A.; Brannigan, B.W.; Harris, P.L.; Haserlat, S.M.; Supko, J.G.; Haluska, F.G.; et al. Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non–Small-Cell Lung Cancer to Gefitinib. *N. Engl. J. Med.* **2004**, *350*, 2129–2139. [CrossRef] [PubMed]
26. Jänne, P.A.; von Pawel, J.; Cohen, R.B.; Crino, L.; Butts, C.A.; Olson, S.S.; Eiseman, I.A.; Chiappori, A.A.; Yeap, B.Y.; Lenehan, P.F.; et al. Multicenter, Randomized, Phase II Trial of CI-1033, an Irreversible Pan-ERBB Inhibitor, for Previously Treated Advanced Non–Small-Cell Lung Cancer. *J. Clin. Oncol.* **2007**, *25*, 3936–3944. [CrossRef] [PubMed]
27. Druker, B.J.; Talpaz, M.; Resta, D.J.; Peng, B.; Buchdunger, E.; Ford, J.M.; Lydon, N.B.; Kantarjian, H.; Capdeville, R.; Ohno-Jones, S.; et al. Efficacy and Safety of a Specific Inhibitor of the BCR-ABL Tyrosine Kinase in Chronic Myeloid Leukemia. *N. Engl. J. Med.* **2001**, *344*, 1031–1037. [CrossRef] [PubMed]
28. Nagar, B.; Bornmann, W.G.; Pellicena, P.; Schindler, T.; Veach, D.R.; Miller, W.T.; Clarkson, B.; Kuriyan, J. Crystal Structures of the Kinase Domain of c-Abl in Complex with the Small Molecule Inhibitors PD173955 and Imatinib (STI-571). *Cancer Res.* **2002**, *62*, 4236–4243. [CrossRef]
29. Liu, T.; Lin, Y.; Wen, X.; Jorissen, R.N.; Gilson, M.K. BindingDB: A web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198–D201. [CrossRef] [PubMed]
30. Davis, M.I.; Hunt, J.P.; Herrgard, S.; Ciceri, P.; Wodicka, L.M.; Pallares, G.; Hocker, M.; Treiber, D.K.; Zarrinkar, P.P. Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **2011**, *29*, 1046–1051. [CrossRef] [PubMed]
31. Khokhlov, I.; Tashchilova, A.; Bugaev-Makarovskiy, N.; Glushkova, O.; Yudin, V.; Keskinov, A.; Yudin, S.; Svetlichnyy, D.; Skvortsova, V. DrugForm-DTA: Towards real-world drug-target binding affinity model. *Comput. Struct. Biotechnol. J.* **2025**, *27*, 4106–4120. [CrossRef] [PubMed]
32. Li, Z.; Ren, P.; Yang, H.; Zheng, J.; Bai, F. TEFDTA: A transformer encoder and fingerprint representation combined prediction method for bonded and non-bonded drug–target affinities. *Bioinformatics* **2024**, *40*, btad778. [CrossRef] [PubMed]
33. Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science* **2023**, *379*, 1123–1130. [CrossRef] [PubMed]
34. Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *arXiv* **2020**. [CrossRef]
35. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv* **2015**. [CrossRef]
36. Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Moschitti, A., Pang, B., Daelemans, W., Eds.; Association for Computational Linguistics: Doha, Qatar, 2014; pp. 1724–1734. [CrossRef]
37. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30. Available online: https://proceedings.neurips.cc/paper_files/paper/7181-attention-is (accessed on 28 August 2026).

38. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2016; pp. 770–778. Available online: https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html (accessed on 20 October 2026).
39. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer Normalization. *arXiv* **2016**. [CrossRef]
40. Lin, Z.; Feng, M.; Santos, C.N.D.; Yu, M.; Xiang, B.; Zhou, B.; Bengio, Y. A Structured Self-attentive Sentence Embedding. *arXiv* **2017**. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.