Graph-Based Cross-Modal Learning for Drug-Target Affinity Prediction

Aakriti Attri [1] **Jashn Khemani** [1] **Arush Gumber** [1] **Anubha Gupta** [*1]

Abstract

Drug-target binding affinity prediction is a fundamental task in structure-based drug discovery. But, in general, sequence and structure-based deep learning methods lack explicit modeling of the physicochemical interactions between drug atoms and protein residues at the binding site. We present HeteroBindNet, a heterogeneous graph neural network that combines 1) GINE-based convolutions over RDKit-derived atomic graphs, 2) GCN-based protein contact graph learning over ESM-2 contact maps, 3) auxiliary global features enriched by Morgan fingerprints, 4) multi-scale 1-D CNN sequence encoders within a shared embedding space, and 5) a novel cross-modal HeteroBindNet which learns an atom-residue interaction matrix through scaled, temperature-gated attention and bidirectional gated message passing, capturing the local structural complementarity that governs binding. Even without any binding-site supervision, this interaction matrix is able to identify a subset of ligand atoms and protein residues (top 5%) that are enriched near known binding regions, consistent with established pharmacophoric contacts. Across multiple evaluation splits, HeteroBindNet achieves a CI of 0.791 and MSE of 0.347 on the KIBA benchmark cold-drug split, outperforming the baseline model, DMFF-DTA, by 14.95% on MSE and improving CI by 5.05%, with consistent performance on the Davis benchmark and strong generalization in cold split scenarios, demonstrating its utility for real-world drug discovery applications.

1. Introduction

Drug discovery is one of the most resource-intensive scientific endeavors, which requires the identification of a suitable drug that binds to a biological target, a protein that plays a key role in the disease process (Mohs & Greig, 2017). On average, it takes around 10-15 years and more than USD 2.8 billion before the use of a novel drug is clinically translated. (Wouters et al., 2020). One of the key areas of research is determining which candidate compounds show high affinity to a disease-relevant protein and to what extent. This is more commonly referred to as drug-target affinity (DTA) prediction.

Traditionally, affinity was measured through the screening of large chemical libraries to identify and optimize lead candidates. However, these experimental approaches are inherently time-consuming, costly, and burdened by high failure rates, creating a significant bottleneck in the early drug discovery pipeline. Classical machine learning methods such as KronRLS (Pahikkala et al., 2015) and SimBoost (He et al., 2017) encode drugs and proteins through pre-computed similarity matrices and apply kernel regression or gradient boosting. The deep learning era brought a paradigm shift. DeepDTA (Öztürk et al., 2018) demonstrated that 1-D convolutional networks applied directly to raw SMILES strings and amino acid sequences could surpass similarity-based methods on the Davis (Davis et al., 2011) and KIBA (Tang et al., 2014) benchmarks. Graph neural network (GNN) based approaches-including GraphDTA (Nguyen et al., 2021), represent drug molecules as molecular graphs, enabling the capture of topological features from chemical structures that sequence encodings are unable to capture. 3DProtDTA (Voitsitskyi et al., 2023) utilizes residue-level protein graphs from AlphaFold (Jumper et al., 2021) predicted structures, encoding inter-residue contacts as typed edges. More recently, advanced graph-based models have been proposed, with a few prominent ones being Drug-BAN (Bai et al., 2023), which introduces bilinear attention for cross-modal drug-target interaction modeling; PocketDTA (Zhao et al., 2024) integrates evolutionary scale modeling (ESM-2) (Lin et al., 2023) protein representations with P2Rank (Krivák & Hoksza, 2018)-predicted binding pocket geometry through geometric vector perceptron-graph neural network (GVP-GNN) (Jing et al., 2020) layers; DMFF-DTA (He et al., 2025) fuses sequence and graph modalities around binding-site-focused graph construction; multi-feature fusion methods, including SMFF-DTA (Wang et al., 2025), GS-DTA (Luo et al., 2025), and MGF-DTA (Ni

* Corresponding Author [1]SBILab, Indraprastha Institute of Information Technology Delhi (IIIT-Delhi), India. Correspondence to: Anubha Gupta <anubha@iiitd.ac.in>.

Accepted at the AI for Science Workshop at the 43rd International Conference on Machine Learning (ICML), 2026., Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

et al., 2026), further advance representational integration through hierarchical attention and transformer-graph fusion strategies. Despite these architectural and domain advances, all of the above models share a fundamental limitation: drug and protein representations are computed independently and merged only at the final concatenation-and-regression stage. We present HeteroBindNet, a multimodal drug-protein binding affinity prediction framework that jointly encodes structural and sequential representations across both modalities. On the protein side, ESM-2 (Lin et al., 2023) contextual embeddings are fed into a three-layer GCN over a residue contact graph, yielding local and global protein embeddings; a parallel multi-scale CNN over the raw sequence with kernels $k \in \{7, 15, 31\}$ captures motif-level features at varying resolutions.

On the drug side, RDKit (Landrum et al., 2026)-derived atom and bond features are encoded via a three-layer graph isomorphism network with edge features (GINE) over the molecular graph, producing local and global drug embeddings; Morgan fingerprints (ECFP4) provide an independent substructure-based representation projected to 128 dimensions. These unimodal representations are coupled through a hybrid cross-modal graph module- the key innovation of HeteroBindNet- which projects atom and residue embeddings into a shared 64-dimensional alignment space, computes a temperature-annealed soft interaction matrix over all atom-residue pairs, and performs bidirectional relational message passing to produce cross-modal embeddings. Finally, all six embeddings are concatenated into a unified 768-dimensional vector and passed through an MLP prediction head to regress the binding affinity score.

2. Background and Related Work

A series of methodological advancements for the task of DTA prediction have been proposed over the years and are broadly classified as machine learning-based approaches, sequence-based deep learning approaches, graph-based structure-aware approaches, knowledge-augmented methods, and more recently, contrastive learning and large language model (LLM)-driven approaches.

2.1. Machine Learning-Based Approaches

KronRLS (Pahikkala et al., 2015)used the Kronecker regularized least squares method for achieving competitive results on early benchmarks. SimBoost (He et al., 2017) introduced a gradient boosting framework based on the pairwise similarity signals derived from the structure of drugs and amino acid sequences of the proteins. While both the approaches use non-neural representations for predicting a continuous affinity value, their performance is limited by their reliance on hand-engineered similarity kernels.

2.2. Sequence-Based Deep Learning

DeepDTA uses parallel convolutional neural networks for the protein and drug sequences and marked one of the earliest successful applications of deep learning to the task of DTA. WideDTA(Öztürk et al., 2019) proposed biophysical feature integration such as protein domain and motif annotation alongside the SMILES sequences of drugs to improve, predictive performance, particularly on the KIBA (He et al., 2017) dataset. AttentionDTA(Zhao et al., 2022) enhanced the convolutional framework using attention mechanisms, by enabling the model to focus on the region of interaction between the drug and target sequences. The biochemical structures of the protein and drug molecules play a major role in molecular binding, and almost all the sequence-based approaches rely on one-dimensional textual representations, which limits their representational capacity.

2.3. Graph Neural Network-Based Approaches

GraphDTA (Nguyen et al., 2021) introduced the representation of drug molecules as graphs to take into account the structural topology and outperformed sequence-only baselines across Davis and KIBA benchmarks. DGraphDTA (Jiang et al., 2020) further extended this approach to include a graph representation of contact maps derived from protein sequences, along with molecular drug graphs. GS-DTA(Luo et al., 2025) addressed the challenge of capturing long-range dependencies between sequentially distant amino acid fragments of target sequences by using a multi-scale sequence model and a graph attention network (GAT)v2-GCN for drug molecular graphs. GEFormerDTA (Liu et al., 2024) proposed a transformer-based graph representation of drugs along with protein secondary structure for an early fusion strategy to highlight the importance of embedding biophysical context in the learning framework. Despite the enhanced representational expressiveness offered by GNNs, these models share some common limitations: lack of access to the broader biochemical context of binding affinity, including pathway memberships, gene ontology features, and homogeneous interaction networks of proteins.

2.4. Structure-Aware and 3D-Informed Approaches

3DProtDTA (Voitsitskyi et al., 2023) was among the first models to use the predicted 3D structure of proteins derived from AlphaFold (Jumper et al., 2021) for DTA prediction to improve affinity estimates even in the absence of experimentally resolved protein structures. PocketDTA (Zhao et al., 2024) advanced this direction by focusing on binding pocket geometry, with the help of (GVP-GNN) (Jing et al., 2020) layers and a graph multi-view pre-training (GraphMVP) (Liu et al., 2021) decoder. DMFF-DTA (He et al., 2025) proposed a dual-modality framework by inte-

grating sequence and structure information from both drugs and proteins, along with a binding site-focused graph construction approach to model the drug-target complex as a unified entity. While these approaches represent meaningful progress, they remain dependent on the availability of accurate structural data and cannot fully incorporate the pharmacological, pathway-level, or adverse-event context as encoded in biomedical knowledge graphs.

2.5. Attention, Fusion, and Multi-Modal Architectures

SMFF-DTA (Wang et al., 2025) proposed a multi-feature fusion method that combines structural information and physicochemical properties of drugs and targets through multiple attention blocks. DrugBAN (Bai et al., 2023) proposed a bilinear attention network that explicitly models cross-modal drug-target interactions, enabling the model to learn interaction-aware unified representations. It was among the first methods to systematically evaluate under cold-drug and cold-target splits, establishing a rigorous generalization benchmark. While these multimodal attention-based methods have substantially improved representational quality, they struggle to reason over the relational biochemical context behind the mechanism of drug-target interaction.

3. Methodology

3.1. Datasets

We evaluate our model on two widely used benchmark datasets for drug-target affinity prediction: Davis (Davis et al., 2011) and KIBA (Tang et al., 2014) (Table 1). Both datasets provide experimentally derived continuous affinity measurements for drug-protein pairs and have been extensively used to compare sequence-based, graph-based, and multimodal DTA models.

Table 1. Drug-Target Binding Affinity Datasets

DATASET	DRUGS	PROTEINS	INTERACTIONS
DAVIS	68	442	30,056
KIBA	2,111	229	118,254
METZ	1423	170	35259
BINDINGDB	9886	1087	42236

Davis: The Davis (Davis et al., 2011) dataset contains kinase inhibitor binding affinities measured as dissociation constants (K_d). Following common practice in DTA prediction, the raw affinity values are transformed into the logarithmic pK_d scale:

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

where K_d is given in nanomolar units. Lower raw K_d values correspond to stronger binding, while higher pK_d values indicate stronger drug-target affinity. Davis is a relatively small and sparse benchmark, containing 68 drugs, 442 proteins, and 30,056 interactions. Its limited number of unique drugs makes cold-drug evaluation particularly challenging, since the model must generalize to chemically unseen compounds from a small drug pool.

KIBA: The KIBA (Tang et al., 2014) dataset integrates heterogeneous bioactivity measurements, including K_i, K_d, and IC_{50}, into a unified KIBA score. Compared with Davis, KIBA contains a substantially larger drug set, with 2,111 drugs, 229 proteins, and 118,254 drug-target interactions. This makes KIBA a larger and denser benchmark for evaluating affinity prediction models. Because the KIBA dataset features a significantly higher number of distinct compounds compared to Davis, it serves as a more robust benchmark for evaluating cold-drug generalization. Simultaneously, it maintains the complexity needed to test a model's capacity to learn protein-specific binding interactions.

3.2. Overview

Most of the models in the drug-target affinity prediction domain treat sequence and structure modalities independently, concatenate them, and make predictions. We instead frame it as a structured communication problem rather than a representational alignment problem. Our model reasons about which atoms bind which residues through a learned interaction field rather than just predicting whether or not a pair interacts.

We propose HeteroBindNet (Figure 1), a multi-modal, dual-resolution framework that takes into account both local atomic geometry and global contextual representations for drug and protein entities. A novel cross-modal gating mechanism operating at the node level fuses these representations before pooling and models this interaction at atomic resolution, ensuring that the final affinity prediction is based on atomic contacts rather than just driven by data modality. There are two core limitations that our model aims to address: (1) the inability of the local-context models to account for domain-level functional context, and (2) the failure of the global-context models to identify and utilize fine-grained binding-site geometry.

The interaction matrix $\mathbf{C}$ produced by our cross-modal module identifies the top 5% of entries that correspond to known binding-site residues and pharmacophoric atoms, without any explicit post-hoc attribution method, thereby enhancing the approach's interpretability.

3.3. Drug Graph Backbone

A drug molecule is represented as a graph $\mathcal{G}_d = (\mathcal{V}_d, \mathcal{E}_d)$ with node features $\mathbf{x}_v \in \mathbb{R}^{41}$ and edge features $\mathbf{e}_{ij} \in \mathbb{R}^4$. These features have been sourced from the RdKit (Landrum et al., 2026) library, with detailed descriptions provided in Appendix Table 3. The node and edge features are first

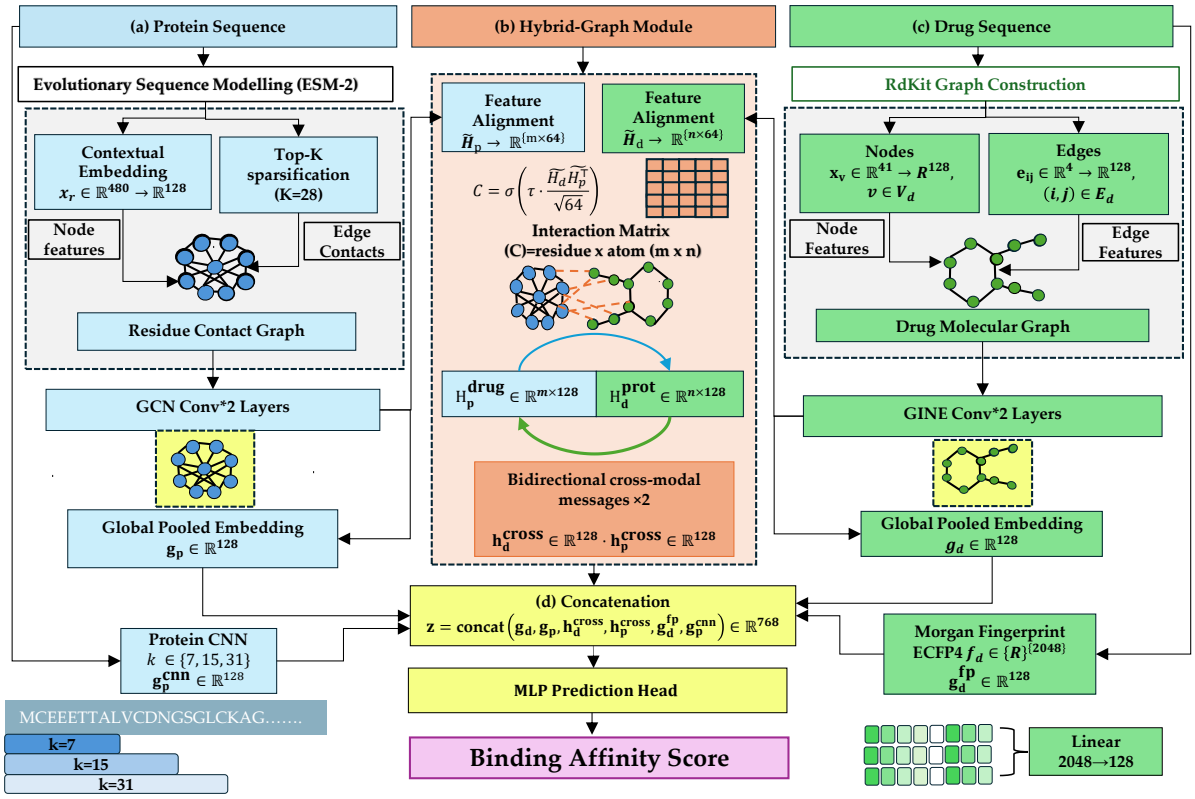

Figure 1. **Architecture of the proposed hybrid drug-protein binding affinity prediction model:** The framework encodes protein sequences via ESM-2 and GCN over a residue contact graph (a), computes atom-residue cross-modal interactions through a scaled dot-product interaction matrix with bidirectional message passing (b), and encodes drug molecules via GINE convolution over RDKit molecular graphs augmented with ECFP4 fingerprints (c). All resulting embeddings are concatenated into a 768-dimensional vector and decoded by an MLP to predict binding affinity (d).

projected to a shared d=128 dimensional space via learned linear maps. We then apply a three-layer Graph Isomorphism Network with Edge features (GINE) (Yue et al., 2024) backbone, where each propagation step computes:

$$\mathbf{h}_v^{(k)} = \mathrm{MLP}^{(k)}\left((1+\epsilon^{(k)})\mathbf{h}_v^{(k-1)} + \sum_{u\in\mathcal{N}(v)} \mathbf{h}_u^{(k-1)} + \mathbf{e}_{uv}\right) \tag{1}$$

where $\epsilon^{(k)}$ is a learnable scalar per layer, followed by an internal MLP, $\mathrm{Linear}(128, 256) \rightarrow \mathrm{ReLU} \rightarrow \mathrm{Linear}(256, 128), \rightarrow \mathrm{BatchNorm1d}(128)$ and $\mathrm{Dropout}(0.1)$. GINE's theoretical equivalence to the 1-WL graph isomorphism test (Chen et al., 2019) allows the model to distinguish non-isomorphic substructures, such as ortho versus para substitution patterns, that GCN and GAT are unable to identify and collapse to identical embeddings. For pharmacophore identification, where substituent position helps to determine binding activity, this is not a theoretical advantage but a practical requirement. This backbone generates local node embeddings, given by $\mathbf{H}_d \in \mathbb{R}^{|\mathcal{V}_d|\times 128}$ a global graph embedding, represented as $\mathbf{g}_d = \mathrm{MeanPool}(\mathbf{H}_d) \in \mathbb{R}^{128}$.

3.4. Protein Graph Backbone

Protein structure is encoded as a contact-map graph, given by $\mathcal{G}_p = (\mathcal{V}_p, \mathcal{E}_p)$, constructed using ESM-2 (Lin et al., 2023) predicted contact probabilities with a hard threshold of 28 neighbors per each node (Lin et al., 2022). This makes the pipeline fully sequence-driven with no dependence on the availability of experimental crystal structures of the proteins. Residue nodes are initialized with ESM-2 (Lin et al., 2023) contextual embeddings $\mathbf{x}_r \in \mathbb{R}^{480}$- providing an evolution-guided feature space from a model pre-trained on 250 million protein sequences. These embeddings are kept frozen during training to preserve the evolutionary signal. A linear projection maps these to $\mathbb{R}^{128}$, after which three layers of GCNConv propagation are applied:

$$\mathbf{H}_p^{(k)} = \sigma\left(\hat{D}^{-1/2}\hat{A}\hat{D}^{-1/2}\mathbf{H}_p^{(k-1)}\mathbf{W}^{(k)}\right) \tag{2}$$

followed by $\mathrm{BatchNorm1d}(128)$ and $\mathrm{Dropout}(0.1)$ after each layer. The contact map topology helps to preserve 3D spatial proximity, making the model robust to structural noise. By initializing from ESM-2 rather than one-hot encodings, the protein GCN produces residue embeddings that are simultaneously structurally aware and evolutionarily in-

formed, a hybrid representation not captured independently by any sequence-only or structure-only baseline. The enoder learns both, fine-grained local residue embeddings $\mathbf{H}_p \in \mathbb{R}^{|\mathcal{V}_p| \times 128}$ along with a global pooled embedding $\mathbf{g}_p \in \mathbb{R}^{128}$ for the protein sequences.

3.5. Cross-Modal Interaction (HeteroBindNet Module)

The central architectural contribution is the HeteroBindNet module, which performs bidirectional atom-residue message passing without explicitly constructing a joint heterogeneous graph. Given drug atom embeddings $\mathbf{H}_d \in \mathbb{R}^{n \times 128}$ and protein residue embeddings $\mathbf{H}_p \in \mathbb{R}^{m \times 128}$, both modalities are first projected into a shared 64-dimensional alignment space:

$$\tilde{\mathbf{H}}_d = \mathbf{H}_d \mathbf{W}_d^{\text{align}} \in \mathbb{R}^{n \times 64}, \tag{3}$$

$$\tilde{\mathbf{H}}_p = \mathbf{H}_p \mathbf{W}_p^{\text{align}} \in \mathbb{R}^{m \times 64}. \tag{4}$$

A scaled dot-product interaction matrix is then computed over all atom–residue pairs:

$$\mathbf{C} = \sigma\left(\frac{\tilde{\mathbf{H}}_d \tilde{\mathbf{H}}_p^\top}{\sqrt{64}\,\tau}\right) \in \mathbb{R}^{n \times m}, \tag{5}$$

where τ is a temperature parameter and $\sigma(\cdot)$ denotes the sigmoid activation. This matrix assigns a soft interaction score to every atom-residue pair. To encourage sparse local communication, we retain only the top-K entries of $\mathbf{C}$, where $K = n \cdot K_{\text{atom}}$ and $K_{\text{atom}} = 1$ in our implementation. Thus, approximately one cross-modal interaction is selected per drug atom, although the selection is performed globally over the flattened atom-residue matrix rather than independently for each atom.

The resulting sparse interaction matrix is used for bidirectional message passing. Protein-to-drug and drug-to-protein messages are computed as:

$$\mathbf{M}_{d \leftarrow p} = \mathbf{C}\mathbf{H}_p \mathbf{W}_{p \to d}^m \in \mathbb{R}^{n \times 128}, \tag{6}$$

$$\mathbf{M}_{p \leftarrow d} = \mathbf{C}^\top \mathbf{H}_d \mathbf{W}_{d \to p}^m \in \mathbb{R}^{m \times 128}. \tag{7}$$

These cross-modal messages are then gated and added back to the original node embeddings through residual updates:

$$\mathbf{H}_d' = \text{LayerNorm}\left(\mathbf{H}_d + \sigma(\mathbf{M}_{d \leftarrow p}\mathbf{W}_d^g) \odot \frac{\mathbf{M}_{d \leftarrow p}}{\mathbf{N}_d}\right), \tag{8}$$

$$\mathbf{H}_p' = \text{LayerNorm}\left(\mathbf{H}_p + \sigma\big(\mathbf{M}_{p \leftarrow d}\mathbf{W}_p^g\big) \odot \frac{\mathbf{M}_{p \leftarrow d}}{\mathbf{N}_p}\right), \tag{9}$$

where $\mathbf{N}_d$ and $\mathbf{N}_p$ denote the number of selected cross-modal neighbors for each atom and residue, respectively. The sigmoid gates suppress irrelevant cross-modal signals adaptively, allowing the model to preserve useful unimodal node information while injecting interaction-specific context.

Finally, the interaction matrix is recomputed over the updated node states and used to obtain interaction-weighted local embeddings:

$$\mathbf{h}_d^{\text{cross}} = \frac{\sum_{i=1}^n \bar{C}_i^{\text{atom}} \mathbf{H}_d'^{(i)}}{\sum_{i=1}^n \bar{C}_i^{\text{atom}}} \in \mathbb{R}^{128}, \tag{10}$$

$$\mathbf{h}_p^{\text{cross}} = \frac{\sum_{j=1}^m \bar{C}_j^{\text{res}} \mathbf{H}_p'^{(j)}}{\sum_{j=1}^m \bar{C}_j^{\text{res}}} \in \mathbb{R}^{128}, \tag{11}$$

where

$$\bar{C}_i^{\text{atom}} = \sum_{j=1}^m C_{ij}, \qquad \bar{C}_j^{\text{res}} = \sum_{i=1}^n C_{ij}.$$

Here, $\bar{C}_i^{\text{atom}}$ measures the cumulative interaction strength of atom i with all selected residues, while $\bar{C}_j^{\text{res}}$ measures the cumulative interaction strength of residue j with all selected atoms. Therefore, the final cross-modal embeddings $\mathbf{h}_d^{\text{cross}}$ and $\mathbf{h}_p^{\text{cross}}$ are not raw passed messages; they are interaction-weighted pooled representations of the cross-updated atom and residue embeddings. This makes the pooling mechanism directly interpretable, since highly weighted atoms and residues correspond to nodes that participate most strongly in the learned atom-residue interaction field.

3.6. Auxiliary Global Backbones

In order to complement the local features extracted by graph message passing, two global encoders are employed:

Morgan Fingerprint MLP: A 2048-bit ECFP4 fingerprint is projected into a 128-dimensional representation $\mathbf{g}_d^{\text{fp}} \in \mathbb{R}^{128}$ using a two-layer MLP comprising $\text{Linear}(2048, 512)$, BatchNorm, ReLU, $\text{Dropout}(0.1)$, and $\text{Linear}\ (512, 128)$ with ReLU activation. Morgan fingerprints encode global topological structural patterns via circular substructure enumeration. These descriptors capture ring-system membership and long-range substituent relationships that may fall beyond the receptive field of a 3-layer GINE encoder.

Protein CNN: Raw amino acid sequences are embedded via $\text{Embedding}(21, 128)$ and processed through three 1D convolutional blocks Table 2 to explicitly encode local sequential motifs complementary to the evolutionary information encoded by the ESM-2 embeddings. Each block

Table 2. Architecture details for the Protein 1D-CNN

BLOCK	IN CH.	OUT CH.	KERNEL	STRIDE
1	128	256	7	1
2	256	256	15	2
3	256	256	31	4

applies BatchNorm, GELU, and $\text{Dropout}(0.1)$. Progressively wider kernels capture motifs at multiple scales simultaneously to encode secondary structure patterns that

contact-map GCN may overlook due to sparse long-range edges. $\mathrm{AdaptiveMaxPool1d}$ extracts the dominant activation, projected to $\mathbf{g}_p^{\text{cnn}} \in \mathbb{R}^{128}$ via $\mathrm{Linear}(256, 128) \rightarrow \mathrm{LayerNorm}(128) \rightarrow \mathrm{GELU}$. The ESM-2 and CNN protein streams are thus complementary: the former captures evolutionary covariation, and the latter captures local sequential characteristics.

3.7. Prediction Head and Full Model

All six 128-dimensional streams are concatenated into a unified representation:

$$\mathbf{z} = \left[\mathbf{g}_d \parallel \mathbf{g}_p \parallel \mathbf{h}_d^{\text{cross}} \parallel \mathbf{h}_p^{\text{cross}} \parallel \mathbf{g}_d^{\text{fp}} \parallel \mathbf{g}_p^{\text{cnn}}\right] \in \mathbb{R}^{768} \tag{12}$$

where each stream is normalized (via $\mathrm{BatchNorm}$ or $\mathrm{LayerNorm}$) prior to concatenation, ensuring all six components contribute at comparable scale. The MLP head then maps $\mathbf{z}$ to a scalar affinity prediction. The fused representation is transformed as $\mathbf{h} = \phi(\mathbf{W}_1\mathbf{z} + \mathbf{b}_1)$, followed by $\hat{y} = \mathbf{W}_2\mathbf{h} + b_2$, where $\phi(\cdot)$ denotes ReLU followed by dropout.

No single stream can be dropped without measurable performance loss, as confirmed by our ablations: the MLP head learns to implicitly weight contributions from all four information scales: atomic topology ($\mathbf{g}_d$, $\mathbf{h}_d^{\text{cross}}$), residue contact structure ($\mathbf{g}_p$, $\mathbf{h}_p^{\text{cross}}$), global fingerprint topology ($\mathbf{g}_d^{\text{fp}}$), and sequential motifs ($\mathbf{g}_p^{\text{cnn}}$). The total model comprises approximately **5 M** trainable parameters, with ESM-2 weights frozen throughout. Appendix Table 1 summarizes all embedding dimensions.

4. Experiments

4.1. Evaluation Pipeline

We evaluate the model under warm split and three cold split settings: cold-drug, cold-target, and cold-both. For each setting, we use 10-fold cross-validation with an 80/10/10 train/validation/test split. The validation set is used for model selection and early stopping, while the test set is held out for final evaluation.

In the cold-drug setting, unique drugs are partitioned into 10 disjoint folds. For fold k, the drugs in fold k are used for testing, the drugs in fold $(k+1) \bmod 10$ are used for validation, and the remaining eight folds are used for training. Thus, no drug appearing in the test or validation set appears in the training set, and the train, validation, and test drug sets are mutually disjoint.

In the cold-target setting, the same protocol is applied over unique targets instead of drugs. Therefore, no target appearing in the test or validation set appears in the training set, and the train, validation, and test target sets are mutually disjoint.

In the cold-both setting, both unique drugs and unique targets are partitioned into disjoint folds. For each fold, the test set contains only drug-target pairs whose drugs and targets are assigned to the test folds, while the validation set contains only pairs whose drugs and targets are assigned to the validation folds. The training set contains only pairs formed from the remaining drug and target folds. Consequently, neither drugs nor targets are shared across the train, validation, and test partitions.

For all settings, we report the mean performance across the 10 folds. Model performance is evaluated using mean squared error (MSE) and concordance index (CI), measuring point-wise regression accuracy and ranking consistency, respectively. Details of all the splits are provided in Appendix Table 2.

4.2. Results

We evaluate HeteroBindNet under two complementary settings. First, a standard warm interaction split to establish baseline competitiveness against prior methods on Davis and KIBA. Second, three cold splits: cold-drug, cold-target, and cold-both designed to assess generalisation to unseen chemical and biological spaces. Table 3 compare HeteroBindNet against representative baselines spanning sequence-based, graph-based, and structure-aware methods on warm splits. Tables 4 and 5 present cold split results, where we additionally report cold-both performance.

Table 3. DTA regression performance under warm split on Davis and KIBA datasets

MODEL	DAVIS		KIBA	
	CI↑	MSE↓	CI↑	MSE↓
KRONRLS (2015)	0.871	0.379	0.782	0.411
SIMBOOST (2017)	0.872	0.282	0.836	0.222
DEEPDTA (2018)	0.878	0.261	0.863	0.194
GRAPHDTA (2021)	0.893	0.229	0.891	0.139
MGRAPHDTA (2022)	0.900	0.207	0.902	0.128
FUSIONDTA (2022)	0.913	0.208	0.906	0.130
COLDDTA (2023)	0.900	0.216	0.892	0.138
3DPROTDTA (2023)	0.908	0.195	0.904	0.122
POCKETDTA (2024)	0.903	0.177	0.892	0.140
HCAF-DTA (2025)	0.908	0.198	0.907	0.122
GS-DTA (2025)	0.903	0.213	0.905	0.124
CDTA (2026)	0.917	0.199	0.882	0.132
(OURS)	**0.900**	**0.176**	**0.906**	**0.133**

4.3. Ablation Study

To verify the contribution of each component in the model to the accurate drug-target affinity prediction ability, we decomposed different components of the model and conducted ablation experiments. Table 6 examines the effect of cross-modal bonds under the Davis cold-drug setting across five folds. Removing cross-modal bonds consistently degrades CI across all folds (mean 0.771 vs. 0.755), confirming

Table 4. Performance comparison on the Davis dataset under cold split settings.

Model	Cold-Drug		Cold-Target		Cold-Both	
	CI↑	MSE↓	CI↑	MSE↓	CI↑	MSE↓
FusionDTA (2022)	0.747	0.681	0.826	0.331	–	–
ColdDTA (2023)	0.768	0.549	0.819	0.393	–	–
LKE-DTA (2025)	0.733	0.546	0.853	0.324	–	–
MixingDTA (2025)	0.754	0.538	0.874	0.231	–	–
DMFF-DTA (2025)	0.742	0.548	0.840	0.330	0.655	0.759
CDTA (2026)	0.767	0.780	0.819	0.463	–	–
Ours	**0.749**	**0.592**	**0.850**	**0.356**	**0.687**	**0.683**

Table 5. Performance comparison on the KIBA dataset under cold split settings.

Model	Cold-Drug		Cold-Target		Cold-Both	
	CI ↑	MSE↓	CI↑	MSE↓	CI↑	MSE↓
FusionDTA (2022)	0.748	0.429	0.685	0.439	0.641	0.587
ColdDTA (2023)	0.788	0.380	0.739	0.372	–	–
LKE-DTA (2025)	–	–	–	–	–	–
MixingDTA (2025)	–	–	–	–	–	–
DMFF-DTA (2025)	0.753	0.408	0.748	0.410	0.667	0.567
CDTA (2026)	0.746	0.410	0.729	0.389	–	–
Ours	**0.791**	**0.340**	**0.751**	**0.377**	**0.669**	**0.547**

Table 6. Effect of cross-modal bonds on Davis cold-drug setting (**CB**:Cross Bonds)

Fold	**CI**		**MSE**	
	With CB	**Without CB**	**With CB**	**Without CB**
1	0.693	0.682	1.120	1.050
2	0.780	0.753	0.678	0.605
3	0.785	0.773	0.353	0.298
4	0.794	0.777	0.433	0.375
5	0.805	0.792	0.623	0.633
Mean	**0.771**	**0.755**	**0.641**	**0.592**

that the interaction matrix captures ranking-relevant binding signals that transfer to unseen drugs. The marginal MSE improvement without cross-modal bonds (0.592 vs. 0.641) reflects a known CI-MSE trade-off in affinity prediction; richer local atom-residue signals improve ranking consistency at the cost of point-wise regression precision. Since candidate ranking governs hit selection in virtual screening, we consider CI the more clinically relevant metric here. Table 7 confirms the contribution of auxilary global features in the the model on the KIBA dataset after removing Morgan fingerprints and 1-D CNN features in the cold-drug setting.

Table 7. Contribution of global features on the KIBA cold-drug setting

Model Variant	**CI ↑**	**MSE ↓**
Full HeteroBindNet	**0.791**	**0.340**
w/o Cross-Modal Bonds	0.779	0.338
w/o Global Auxiliary Features (Morgan Fingerprints + CNN Encoder)	0.743	0.410

Removing Morgan fingerprints and the CNN encoder produces the largest performance drop, with CI decreasing by 6.1% (0.791 → 0.743) and MSE increasing by 20.6% (0.340 → 0.410). This impact is substantially larger than removing cross-modal bonds alone (CI: -1.5%, MSE: -0.6%). Together, the two ablation studies establish a complementary division of labor within HeteroBindNet: cross-modal bonds drive ranking consistency while global auxiliary encoders provide the distribution-aware representation necessary for accurate point-wise affinity estimation under cold split conditions.

5. Discussion and Conclusion

On standard random-split benchmarks, HeteroBindNet achieves competitive performance across both datasets. On Davis, our model attains the lowest MSE of 0.176 across all compared methods, including structure-aware models such as PocketDTA (0.177) and 3DProtDTA (0.195), while achieving a CI of 0.900, consistent with leading graph-based methods. On KIBA, HeteroBindNet achieves a CI of 0.906, competitive with FusionDTA (0.906) and HCAF-DTA (0.907), with an MSE of 0.133 that remains within the performance envelope of current state-of-the-art methods. These results are achieved without binding-site supervision, suggesting that structural signal from 2D graphs and sequence-derived protein representations is sufficient to match or exceed specialized baselines on point-wise affinity estimation.

HeteroBindNet demonstrates strong and consistent generalization under cold split evaluation conditions. On the cold-drug split of the KIBA benchmark, our model achieves the highest CI (0.791 ± 0.047) and lowest MSE (0.340 ± 0.014) across all compared methods, outperforming DMFF-DTA by 14.95% on MSE and 5.05% on CI. On the cold-target split, HeteroBindNet remains competitive, achieving a CI of 0.751 and MSE of 0.377, comparable to DMFF-DTA and CDTA. In the cold-both setting, where neither drugs nor targets are shared across partitions, HeteroBindNet consistently outperforms DMFF-DTA on both benchmarks, reducing MSE from 0.759 to 0.683 on Davis and from 0.567 to 0.547 on KIBA.

The Davis cold-drug setting remains a limitation, where HeteroBindNet obtains an MSE of 0.592, higher than MixingDTA (0.538) and ColdDTA (0.549). We attribute this in part to the smaller size and lower affinity variance of Davis relative to KIBA, which amplifies MSE sensitivity to outlier predictions under strict cold-drug partitioning. However, Davis cold-target performance remains strong (CI: 0.850, MSE: 0.356), and the cold-both results indicate that the learned representations retain meaningful signal even under simultaneous drug and target distribution shifts.

The generalization advantage of HeteroBindNet is mechanistically grounded in its architecture. Unlike DMFF-DTA,

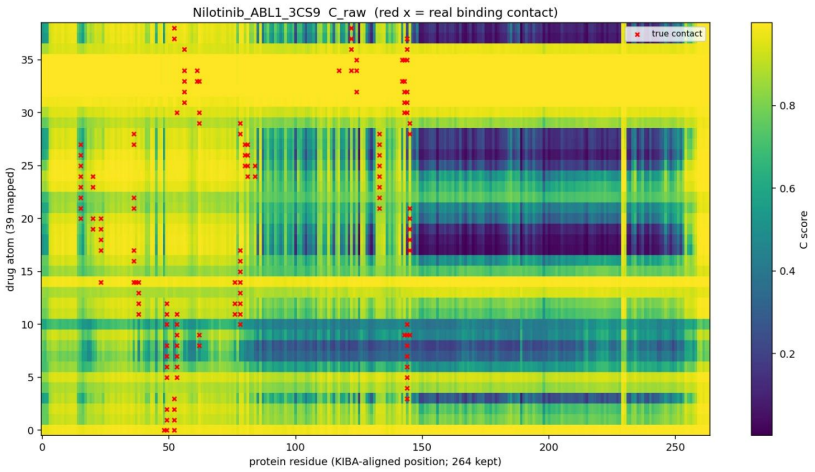

Figure 2. Visualisation of the learned interaction matrix $\mathbf{C}$ for Nilotinib with ABL1 (PDB: 3CS9). The heatmap shows cross-modal interaction scores across drug atom and protein residue pairs, with protein residues mapped to KIBA-aligned positions. Red markers ($\times$) denote true binding contacts derived from the co-crystal structure.

which constructs binding site-focused protein graphs using AlphaFold2 as structural supervision, HeteroBindNet recovers binding-relevant atom-residue contacts from affinity labels through the learned interaction matrix. This indicates that the Hybrid Graph Module operates over feature space rather than memorized drug or target identities. The ablation studies support this interpretation: cross-modal bonds improve ranking consistency by capturing local structural complementarity at the binding interface, while Morgan fingerprints and the CNN encoder provide the distribution-aware representation needed for accurate point-wise affinity estimation under cold split conditions.

Figure 2 visualizes the learned interaction matrix $\mathbf{C}$ for the Nilotinib with ABL1 (PDB: 3CS9) (Cowan-Jacob et al., 2008) complex, with ground-truth binding contacts from the co-crystal structure overlaid as red markers. The true contacts do not distribute uniformly, but localize within structured regions of the latent interaction matrix. This supports the claim that the cross-modal architecture can recover biologically meaningful atom-residue interaction patterns without explicit 3D spatial coordinates or binding-site labels during training. A more comprehensive structural validation across five protein-ligand complexes, including quantitative contact-recall analysis, is provided in Appendix Tables 4 and 5 and Appendix Figures 1 and 2.

Overall, HeteroBindNet shows that explicit cross-modal interaction modeling between atom and residue spaces improves drug-target affinity prediction under realistic generalization conditions. The strong cold-split performance across Davis and KIBA suggests that the learned interaction matrix captures binding-relevant substructures rather than dataset-specific correlations. Furthermore, the identification of pharmacophoric atoms and residues without site-level labels points toward a broader principle: affinity optimization can induce sufficiently expressive cross-modal graph models to recover biologically meaningful interaction patterns. Future work could extend the Hybrid Graph Module to include 3D structural graphs, explicit binding pocket geometries, larger protein language models, and conformational flexibility. Beyond DTA prediction, the approach may extend to protein-protein interaction, RNA-ligand binding, and other heterogeneous biological graph settings where inter-type connectivity is latent and must be inferred from data.

Impact Statement

This work advances computational methods for predicting drug-target binding affinity, with direct implications for accelerating early-stage drug discovery by enabling more efficient virtual screening of candidate compounds against protein targets. The emergent binding site interpretability of our model further offers a mechanism for hypothesis generation that could guide medicinal chemists toward more targeted lead optimization strategies. We do not anticipate direct negative societal consequences of this work. However, as with all computational drug discovery tools, predictions should be treated as hypothesis-generating rather than conclusive and should be validated experimentally before informing clinical or therapeutic decisions.

Acknowledgement

This research was supported by project grant no. IMP-013, awarded by the Center of Excellence in Healthcare (CoEHe), Indraprastha Institute of Information Technology Delhi (IIIT-Delhi), India. The authors are affiliated with SBILab, IIIT-Delhi (https://sbilab.iiitd.edu.in/).

References

Bai, P., Miljković, F., John, B., and Lu, H. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. *Nature Machine Intelligence*, 5 (2):126–136, 2023.

Chen, Z., Villar, S., Chen, L., and Bruna, J. On the equivalence between graph isomorphism testing and function approximation with gnns. *Advances in neural information processing systems*, 32, 2019.

Cowan-Jacob, S. W., Fendrich, G., Manley, P., Liebetanz, J., and Fabbro, D. Human ABL kinase in complex with nilotinib, April 2008.

Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., Hocker, M., Treiber, D. K., and Zarrinkar, P. P. Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology*, 29(11):1046–1051, 2011.

He, H., Chen, G., Tang, Z., and Chen, C. Y.-C. Dual modality feature fused neural network integrating binding site information for drug target affinity prediction. *NPJ Digital Medicine*, 8(1):67, 2025.

He, T., Heidemeyer, M., Ban, F., Cherkasov, A., and Ester, M. Simboost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. *Journal of cheminformatics*, 9(1):24, 2017.

Jiang, M., Li, Z., Zhang, S., Wang, S., Wang, X., Yuan, Q., and Wei, Z. Drug–target affinity prediction using graph neural network and contact maps. *RSC advances*, 10(35): 20701–20712, 2020.

Jing, B., Eismann, S., Suriana, P., Townshend, R. J., and Dror, R. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*, 2020.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Krivák, R. and Hoksza, D. P2rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of cheminformatics*, 10(1):39, 2018.

Landrum, G., Tosco, P., Kelley, B., Rodriguez, R., Cosgrove, D., Vianello, R., sriniker, Gedeck, P., Jones, G., Kawashima, E., NadineSchneider, Nealschneider, D., tadhurst-cdd, Dalke, A., Swain, M., Cole, B., Turk, S., Savelev, A., Maeder, N., Pechersky, Y., Vaucher, A., Wójcikowski, M., Walker, R., Faara, H., Take, I., Scalfani, V. F., Probst, D., Ujihara, K., Monat, J., and Lehtivarjo, J. rdkit/rdkit: 2026_03_2 (q1 2026) release, 2026.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022:500902, 2022.

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*, 379(6637): 1123–1130, 2023.

Liu, S., Wang, H., Liu, W., Lasenby, J., Guo, H., and Tang, J. Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728*, 2021.

Liu, Y., Xing, L., Zhang, L., Cai, H., and Guo, M. Geformerdta: drug target affinity prediction based on transformer graph for early fusion. *Scientific Reports*, 14 (1):7416, 2024.

Luo, J., Zhu, Z., Xu, Z., Xiao, C., Wei, J., and Shen, J. Gsdta: integrating graph and sequence models for predicting drug-target binding affinity. *BMC genomics*, 26(1):105, 2025.

Mohs, R. C. and Greig, N. H. Drug discovery and development: Role of basic biological research. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 3(4):651–657, 2017.

Nguyen, T., Le, H., Quinn, T. P., Nguyen, T., Le, T. D., and Venkatesh, S. Graphdta: predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37 (8):1140–1147, 2021.

Ni, Z., Wei, B., and Zeng, Y. Mgf-dta: A multi-granularity fusion model for drug–target binding affinity prediction. *International Journal of Molecular Sciences*, 27(2):947, 2026.

Öztürk, H., Özgür, A., and Ozkirimli, E. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

Öztürk, H., Ozkirimli, E., and Özgür, A. Widedta: prediction of drug-target binding affinity. *arXiv preprint arXiv:1902.04166*, 2019.

Pahikkala, T., Airola, A., Pietilä, S., Shakyawar, S., Szwajda, A., Tang, J., and Aittokallio, T. Toward more realistic drug–target interaction predictions. *Briefings in bioinformatics*, 16(2):325–337, 2015.

Tang, J., Szwajda, A., Shakyawar, S., Xu, T., Hintsanen, P., Wennerberg, K., and Aittokallio, T. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *Journal of chemical information and modeling*, 54(3):735–743, 2014.

Voitsitskyi, T., Stratiichuk, R., Koleiev, I., Popryho, L., Ostrovsky, Z., Henitsoi, P., Khropachov, I., Vozniak, V., Zhytar, R., Nechepurenko, D., et al. 3dprotdta: a deep learning model for drug-target affinity prediction based on residue-level protein graphs. *RSC advances*, 13(15): 10261–10272, 2023.

Wang, X., Xia, Z., Feng, R., Han, T., Wang, H., Yu, W., and Wang, X. Smff-dta: using a sequential multi-feature fusion method with multiple attention mechanisms to predict drug-target binding affinity. *BMC biology*, 23(1): 120, 2025.

Wouters, O. J., McKee, M., and Luyten, J. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *Jama*, 323(9):844–853, 2020.

Yue, X., Liu, B., Zhang, F., and Qu, G. Edged weisfeiler-lehman algorithm. In *Artificial Neural Networks and Machine Learning – ICANN 2024: 33rd International Conference on Artificial Neural Networks, Lugano, Switzerland, September 17–20, 2024, Proceedings, Part V*, pp. 93–109, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72343-8. doi: 10.1007/978-3-031-72344-5_7. URL https://doi.org/10.1007/978-3-031-72344-5_7.

Zhao, L., Wang, H., and Shi, S. Pocketdta: an advanced multimodal architecture for enhanced prediction of drug-target affinity from 3d structural data of target binding pockets. *Bioinformatics*, 40(10):btae594, 2024.

Zhao, Q., Duan, G., Yang, M., Cheng, Z., Li, Y., and Wang, J. Attentiondta: Drug–target binding affinity prediction by sequence-based deep learning with attention mechanism. *IEEE/ACM transactions on computational biology and bioinformatics*, 20(2):852–863, 2022.

A. Appendix.

Table 1. Embedding dimensions across all model components.

Component	Output	Dim
Drug node / edge projection	$\mathbf{x}_v^{(0)}$, $\mathbf{e}_{ij}$	128
Drug GINE (×2)	$\mathbf{H}_d$	$n \times 128$
Drug global pool	$\mathbf{g}_d$	128
Protein ESM-2 embedding	$\mathbf{x}_r$	480
Protein node projection	$\mathbf{h}_r^{(0)}$	128
Protein GCN (×2)	$\mathbf{H}_p$	$m \times 128$
Protein global pool	$\mathbf{g}_p$	128
Drug / protein align projection	$\tilde{\mathbf{H}}_d$, $\tilde{\mathbf{H}}_p$	128×64
Interaction matrix	$\mathbf{C}$	$n \times m$
Cross-modal drug message	$\mathbf{h}_d^{\text{cross}}$	128
Cross-modal protein message	$\mathbf{h}_p^{\text{cross}}$	128
Morgan fingerprint input	$\mathbf{f}_d$	2048
Morgan MLP output	$\mathbf{g}_d^{\text{fp}}$	128
CNN embedding	token $\rightarrow$ channel	$128 \rightarrow 256$
CNN output	$\mathbf{g}_p^{\text{cnn}}$	128
Final concatenation	$\mathbf{z}$	768

Table 2. Cold split setting: distribution of drug-target pairs in training and testing sets for the Davis and KIBA datasets.

Davis dataset				**KIBA dataset**			
Drug cold split		**Target cold split**		**Drug cold split**		**Target cold split**	
Total Drug	68	Total Target	442	Total Drug	2111	Total Target	229
Training Drug	55	Training Target	354	Training Drug	1669	Training Target	184
Testing Drug	13	Testing Target	88	Testing Drug	442	Testing Target	45
Training Set	5746	Training Set	24072	Training Set	94599	Training Set	94528
Testing Set	24310	Testing Set	5984	Testing Set	23655	Testing Set	23726
Total Set	30056	Total Set	30056	Total Set	118254	Total Set	118254

Table 3. RDKit-derived node and edge features used for drug graph construction.

Feature type	Encoded values	Dimension
Atom / node features		
Atom type	C, N, O, S, F, Si, P, Cl, Br, I, other	11
Atom degree	0, 1, 2, 3, 4, 5, 6, other	8
Formal charge	-3, -2, -1, 0, 1, 2, 3, other	8
Total number of hydrogens	0, 1, 2, 3, 4, other	6
Hybridization	SP, SP2, SP3, SP3D, SP3D2, other	6
Aromaticity	`IsAromatic` flag	1
Ring membership	`IsInRing` flag	1
Total atom feature dimension	–	41
Bond / edge features		
Bond type	SINGLE, DOUBLE, TRIPLE, AROMATIC	4
Total edge feature dimension	–	4
Total RDKit-derived feature dimensions	41 node + 4 edge	45

Structural Validation of Cross-Modal Interactions. To evaluate whether the learned cross-modal interaction matrix captures physically meaningful binding relationships, we compare the predicted interaction scores against experimentally

Table 4. The five KIBA pairs used for structural validation. "Atoms" is the number of heavy atoms in the drug after RDKit-to-PDB matching; "Residues kept" is the number of protein residues that appear in both the PDB structure and the model's 1022-residue input window; "Contacts" is the number of atom-residue pairs at ≤ 4.5 Å heavy-atom distance after alignment. The model's predicted KIBA score is in the high-affinity regime for all five inputs, confirming that the model is confident on these inputs.

Drug	Target	PDB	Atoms	Residues kept	Contacts (4.5 Å)	Pred. KIBA
Nilotinib	ABL1	3CS9	39	264	134	12.27
Ponatinib	ABL1	3OXZ	39	268	146	12.31
Sunitinib	VEGFR2	4AGD	29	157	58	11.39
Bosutinib	ABL1	3UE4	36	270	92	11.78
Afatinib	EGFR	4G5J	34	307	94	13.44

Table 5. Per-pair recall of true contacts at four cross-bond score thresholds, plus the median C score over true contact positions only. Higher is better; the bottom row is the mean across the five pairs.

Pair	# Contacts	median C	R@0.5	R@0.7	R@0.8	R@0.9
Nilotinib–ABL1 (3CS9)	134	0.93	0.94	0.86	0.82	0.66
Ponatinib–ABL1 (3OXZ)	146	0.90	0.91	0.74	0.65	0.49
Sunitinib–VEGFR2 (4AGD)	58	1.00	1.00	1.00	1.00	0.93
Bosutinib–ABL1 (3UE4)	92	0.91	0.91	0.72	0.66	0.51
Afatinib–EGFR (4G5J)	94	0.93	0.83	0.67	0.60	0.53
mean $\pm$ std	—	0.93	$\mathbf{0.92 \pm 0.06}$	0.80 ± 0.13	0.75 ± 0.17	0.62 ± 0.18

observed atom-residue contacts derived from crystal structures. The cross-modal module produces a dense interaction matrix $\mathbf{C} \in [0,1]^{N_{\text{atom}} \times N_{\text{res}}}$ that assigns a soft interaction score to every drug atom-protein residue pair. Rather than treating this as a binary contact predictor, we interpret $\mathbf{C}$ as a learned interaction field and ask whether experimentally verified binding contacts receive consistently high interaction scores.

We evaluate five drug-target-PDB tuples from the KIBA test set (Table 4). For each tuple, we retrieve the corresponding PDB structure, align the crystallographic protein chain to the KIBA protein sequence using Needleman-Wunsch alignment, and map ligand atoms to the SMILES-derived molecular graph using RDKit substructure matching. Ground-truth contact maps $\mathbf{T} \in \{0,1\}^{N_{\text{atom}} \times N_{\text{res}}}$ are constructed using a heavy-atom distance threshold of 4.5 Å. The trained model is then executed in inference mode to extract the post-sigmoid interaction matrix $\mathbf{C}$. Both $\mathbf{C}$ and $\mathbf{T}$ are restricted to residues present in the resolved PDB structure and within the model's 1022-residue inference window.

Drug graphs are constructed using heavy atoms only, consistent with standard cheminformatics pipelines employing implicit hydrogens. Consequently, the validated ligands contain between 29 and 39 graph nodes depending on molecular size. On the protein side, the validated residue counts correspond only to residues successfully aligned between the crystallographic structure and the sequence window processed by the model. Since kinase crystal structures typically resolve only the catalytic kinase domain rather than the full-length protein, validation is restricted to the structurally resolved region for which experimental ground truth is available. The model nevertheless produces interaction scores over the full protein sequence, including residues outside the crystallographically observed domain.

Metrics. For each pair we report two interpretability-focused quantities:

- **Recall at threshold** τ: of the true contacts in the crystal structure, what fraction have $C \geq \tau$? We sweep $\tau \in \{0.5, 0.7, 0.8, 0.9\}$.
- **Distribution of** C **at true contacts**: the median, IQR, and full spread of cross-bond scores assigned to actually-bonded atom-residue pairs. The right benchmark is $C = 0.5$, which is what an untrained sigmoid output would average to.

Findings. Across the five pairs, the model recovers $\mathbf{91.9\% \pm 6.1\%}$ of real 4.5 Å atom–residue contacts at $C \geq 0.5$ (Figure 1, Table 5). The median cross-bond score *at* true contact positions is 0.93 (Figure 2); for the Sunitinib-VEGFR2 pair it is ≥ 0.99. Even at the strict threshold $C \geq 0.9$ the model still recovers 62.3% of real contacts on average, and at $C \geq 0.8$ the recall is 74.6%. The lowest single-pair recall at $\tau = 0.5$ is Afatinib-EGFR at 83%.

Interpretation. The cross-bond matrix concentrates its mass on the residues that actually participate in binding. This is a non-trivial result: the model is trained *only* on the scalar pK_i-equivalent KIBA score and never sees a 3D structure or a

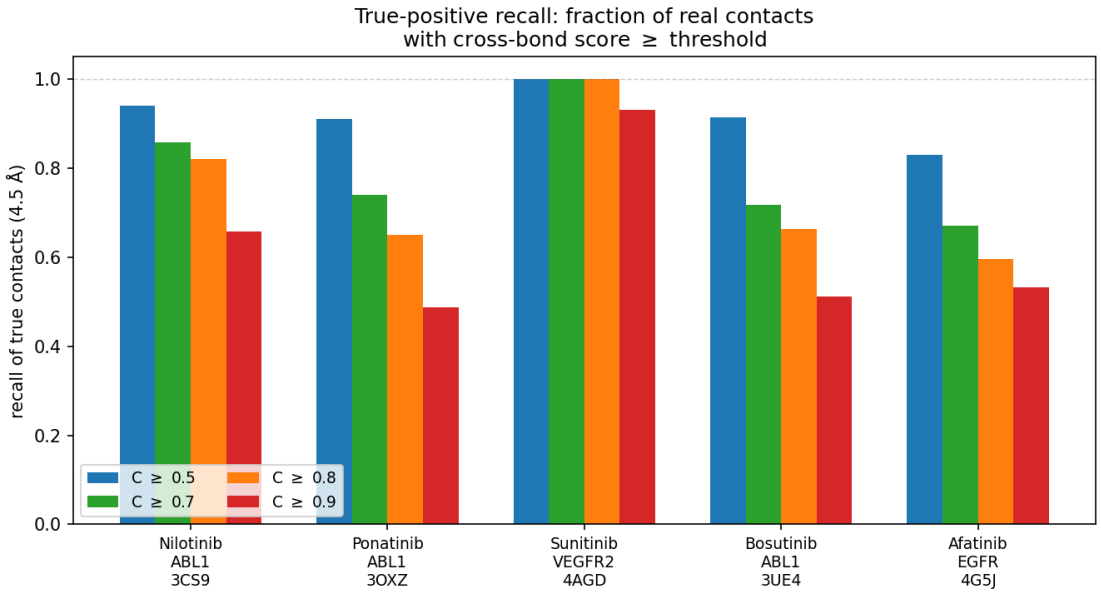

Figure 1. **True-positive recall of real binding contacts.** For each of the five validation pairs, the bars show the fraction of real 4.5 Å atom-residue contacts whose cross-bond score C exceeds a given threshold (blue 0.5, green 0.7, orange 0.8, red 0.9). At a permissive threshold of $C \geq 0.5$ the model recovers $\mathbf{91.9\%} \pm \mathbf{6.1\%}$ of real contacts on average across the five pairs; at the strict threshold $C \geq 0.9$ it still recovers 62.3%. The Sunitinib-VEGFR2 pair is at 100% recall for thresholds up to 0.8. This is our central evidence that the cross-bond mechanism is not random with respect to true binding geometry.

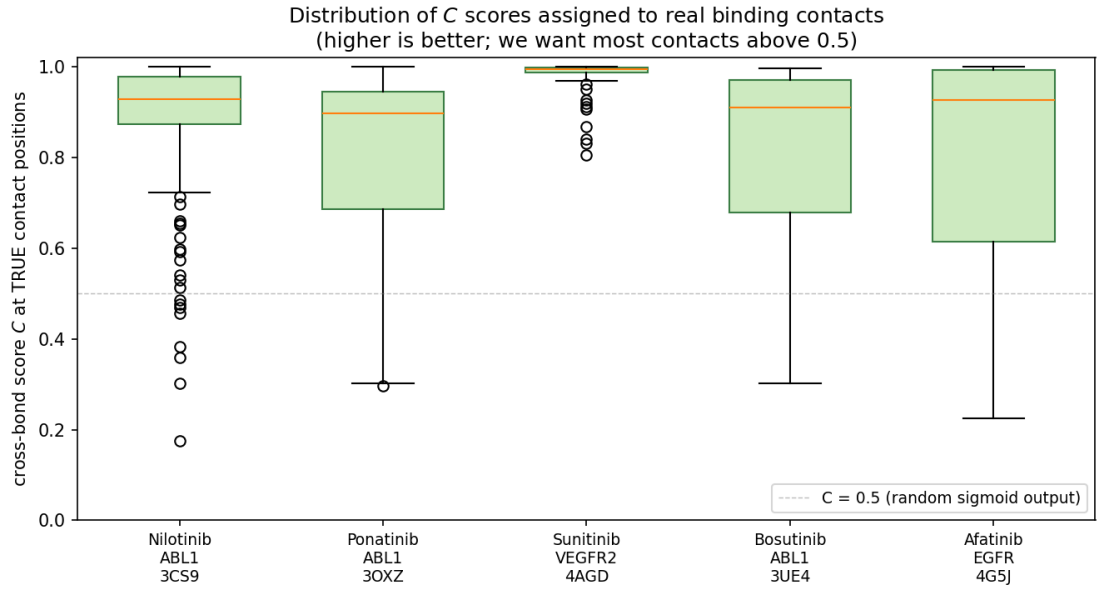

Figure 2. **Distribution of C scores at true contact positions only.** Box plots show the spread of cross-bond scores assigned by the trained model to actually-bonded atom-residue pairs. The median C score at true contacts is 0.93 across all pairs, well above the $C = 0.5$ neutral line; for Sunitinib-VEGFR2 the median is ≥ 0.99. The presence of low-C outliers in the Nilotinib and Afatinib panels marks the $\sim$ 5-15% of true contacts that the model misses, typically peripheral van der Waals contacts on flexible side chains, where the model's contact estimate from the ESM-derived graph is most uncertain.

contact label, yet at inference time the high-C entries align with the crystallographic atom-residue contacts in $\sim 92\%$ of cases. We read the cross-bond mechanism as a *learned, soft binding-pocket attention* that is high-recall on real interactions:

a usable substrate for downstream tasks like binding-site annotation, mutation-effect attribution, and ligand-redirection studies, without any structural supervision during training. The remaining gap between $\sim 92\%$ recall and a hypothetical 100% is concentrated on peripheral van der Waals contacts on flexible side chains, where even crystallographic ground truth is itself somewhat noisy; tightening this gap is a clean direction for follow-up work using a small contact-aware auxiliary loss.