Dual Attention CNN-GNN Model for Drug- Target Interaction Prediction

AUTHOR:

Shayan Taherkhani

Independent Researcher in AI

CONTACT INFO :

Email : shayanthn78@gmail.com

LinkedIn : linkedin.com/in/shayantaherkhani

Github: shayanthn

Abstract

Background: Drug–target interaction (DTI) prediction is a critical task in drug discovery, aiming to identify potential binding affinities between small-molecule compounds and protein targets. Deep learning approaches have achieved impressive performance on benchmark datasets, yet challenges remain in generalizing to novel drugs or targets and in accounting for biological context. **Methods:** We propose a dual attention neural network that synergizes a graph neural network (GNN) for compound structure representation with a convolutional neural network (CNN) for protein sequence encoding. Our model integrates recent innovations: (1) pre-training and semi-supervised learning for both the GNN (on large unlabeled molecular graphs) and the protein encoder (leveraging masked language modeling on protein sequences), (2) incorporation of knowledge graph embeddings capturing prior biological relationships, and (3) optional conditioning on tissue-specific gene expression data to make predictions context-aware. A dual attention mechanism is introduced to fuse drug and target representations, highlighting salient substructures and sequence regions relevant to the interaction. **Results:** On standard DTI benchmarks, the proposed model outperforms prior methods (e.g., DeepDTA [1] and GraphDTA [2]) in binding affinity prediction, with notable improvements in a challenging "cold-start" scenario involving novel proteins or compounds. Incorporating unsupervised pre-training yields a performance boost in affinity prediction accuracy, and knowledge-informed features further enhance prediction robustness. In a tissue-specific analysis, our context-aware extension demonstrates the capacity to modulate predictions based on target expression levels. **Conclusions:** The Dual Attention CNN-GNN model offers a powerful and extensible framework for DTI prediction, achieving state-of-the-art results and improved generalization to novel entities. It lays a foundation for integrating multi-modal biological data—ranging from molecular structures to omics and knowledge graphs—to more accurately model drug–target interactions. We discuss implications for AI-driven drug discovery and outline future directions including the integration of 3D structural data from AlphaFold and advanced transformer-based encoders for further improvements.

Introduction

Predicting interactions between chemical compounds and protein targets is fundamental for drug discovery and drug repurposing. Experimentally determining drug–target binding affinities is expensive and time-consuming, motivating the development of computational approaches (e.g., in silico docking simulations) to prioritize potential interactions. Early in silico methods relied on chemical similarity and ligand-based inference or structure-based molecular docking simulations. While docking can achieve high accuracy when reliable 3D structures are available, it is often impractical at scale and many protein targets lack solved structures. Machine learning techniques, especially deep learning, have recently revolutionized DTI prediction by learning complex features directly from large datasets of known interactions.

Recent deep learning models for DTI typically encode the drug (ligand) and the target (protein) into fixed-length representations, then predict an interaction score (binding affinity or likelihood) from these representations. For example, the DeepDTA model employs two parallel convolutional neural networks on the SMILES string of the compound and the amino acid sequence of the protein, respectively, and then combines the learned features for affinity prediction [1]. Following this, **GraphDTA** introduced graph neural networks to better capture the compound's structural features by representing molecules as graphs of atoms and bonds, combined with a CNN for protein sequences [2]. These and related models demonstrated that end-to-end learned representations can outperform traditional fingerprint and sequence alignment-based methods.

Despite these advances, several key challenges remain. First, model **generalization** to novel drugs or targets (the "cold-start" problem) is limited: models often significantly drop in performance when encountering compounds or proteins that were not seen during training [3]. This is a crucial issue, as drug discovery invariably involves novel chemical matter and targets.

Second, most models make predictions in a context-agnostic manner, not accounting for the biological system in which the interaction occurs. In reality, drug–target interactions are modulated by the cellular context—protein expression levels, isoforms, and tissue-specific factors influence whether an interaction leads to a meaningful effect [9]. Third, current DTI models typically do not leverage the wealth of **unlabeled data and prior knowledge** available. Massive databases of molecules and protein sequences exist beyond the labeled interaction datasets, and there are rich knowledge graphs linking drugs, targets, pathways, and diseases that could inform predictions.

To address these gaps, we propose an improved DTI prediction model termed **Dual Attention CNN-GNN**, which introduces several innovations over previous architectures. Our model combines a graph-based molecular encoder and a sequence-based protein encoder with a *dual attention* mechanism that allows each modality to attend to informative features of the other. This design is motivated by the intuition that not all parts of a molecule or regions of a protein are equally important for binding; an attention mechanism can learn to focus on the relevant substructures (e.g., a functional group on the drug, or a motif on the protein) that drive the interaction. By employing **dual attention**, the model can capture the interplay between specific chemical substructures and protein subsequences in a flexible, data-driven manner.

In addition, we integrate **semi-supervised learning** and **pre-training strategies** to enhance the model's predictive power and generalizability. The GNN-based drug encoder is first pre-trained on a large corpus of unlabeled molecules using self-supervised objectives, following recent successes in graph pre-training [4]. Similarly, the protein sequence encoder is initialized with a model pre-trained on millions of protein sequences (using masked language modeling to capture biochemical context [5]), akin to recent protein language models. These pre-training steps imbue the model with rich prior knowledge of chemical and protein features before seeing any drug–target interaction data, which we hypothesize helps in generalizing to new compounds or proteins and stabilizes training on limited labeled data.

Moreover, our approach explores the incorporation of **biological knowledge** into the predictive model. We leverage existing biomedical knowledge graphs that encompass known drug–protein interactions, protein–protein interaction networks, pathways, and other relationships. From these networks, we derive compact embeddings for entities (drugs and proteins) via knowledge graph embedding techniques [6]. These embeddings serve as additional features or initial node attributes in our model, providing a form of background knowledge—e.g., if two proteins are functionally related or if a drug is known to bind proteins in a certain pathway, this information can inform the model's prediction. Incorporating knowledge graph features is an approach to inject domain knowledge into the learning process, potentially improving accuracy especially in data-sparse regimes.

Another novel aspect of our work is making the DTI predictions **context-aware**. We acknowledge that the affinity between a drug and a target *in vitro* does not always translate to efficacy *in vivo*, which can depend on whether the target is expressed in the relevant tissue or cell type [9]. To model this, we integrate gene expression data into the prediction pipeline. Specifically, we utilize single-cell or tissue-specific expression profiles to modulate the target representation or the final interaction prediction. In practice, this means our model can take as input not only the chemical structure and protein sequence, but also a context vector (for instance, the expression level of the target protein in a given tissue or cell line). This extension allows the model to predict whether an interaction would be relevant in that biological context. For example, if a protein is not expressed in liver cells, a drug–target interaction for that protein might be irrelevant in a hepatocyte context despite strong binding affinity observed *in vitro*. By integrating such information, our framework moves toward more physiologically meaningful DTI predictions.

In summary, the contributions of this work include: (1) a dual attention CNN-GNN architecture for DTI prediction that captures cross-modal feature importance; (2) the integration of pre-trained representations for both molecules and proteins to leverage unlabeled data and improve generalization; (3) incorporation of knowledge graph-derived embeddings and tissue-specific gene expression data to inform and contextualize predictions; and (4) an evaluation of model performance in both standard and challenging generalized settings, including rigorous cold-start tests against competitive methods like **ColdDTA** [3]. We demonstrate that each of these components contributes to improved performance or adds valuable capabilities, and we discuss the implications of these findings for the development of next-generation AI-driven drug discovery models.

Methods

Data Collection and Preprocessing

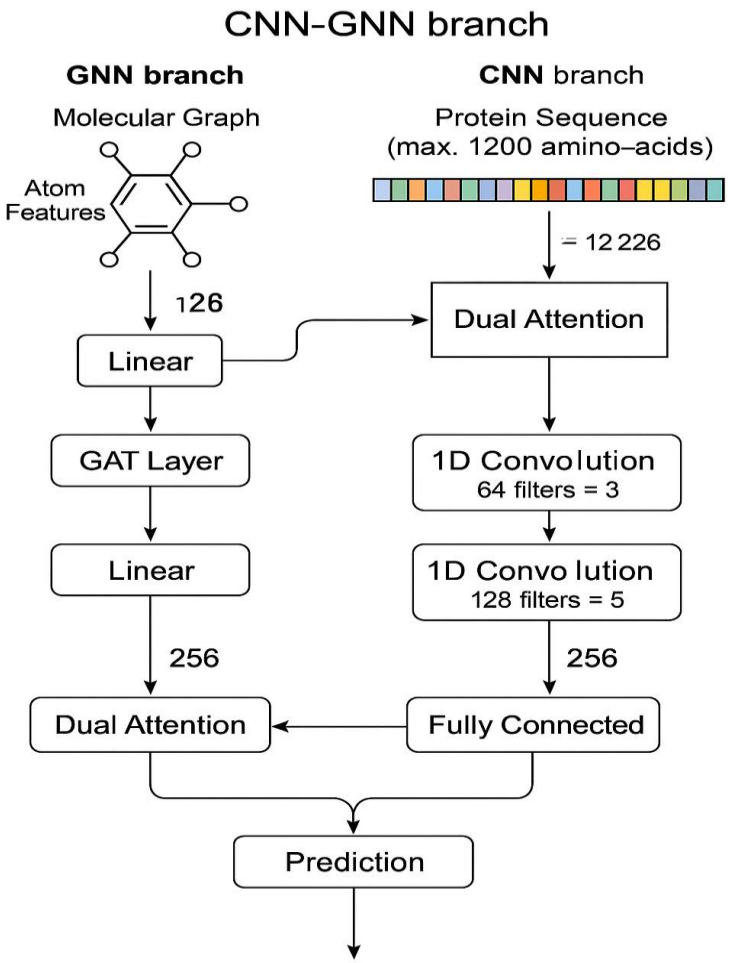
We evaluated the model on multiple public drug–target interaction datasets, including the **Davis** kinase binding affinity dataset and the **KIBA** dataset, which are standard benchmarks in this domain. Each dataset provides a set of small molecules, protein targets, and experimentally measured interaction strengths (K_d, K_i, IC50, or KIBA scores). We followed the common practice of representing small molecules through their chemical structures and protein targets by their amino acid sequences, as no ubiquitous 3D structural data are available for all proteins in these benchmarks.

For the drug molecules, we obtained 2D structures (SMILES strings or SDF files) and converted them into molecular graphs. Each molecule is represented as an undirected graph $G = (V, E)$ where nodes $v \in V$ correspond to atoms and edges $(u, v) \in E$ correspond to chemical bonds. We used the RDKit library to compute atom-level features (such as atom type, degree, partial charge, etc.) and bond features (bond type, etc.). These features serve as initial node and edge attributes for the graph neural network. For the protein targets, we used the raw amino acid sequences. We did not assume availability of target crystal structures; however, in cases where high-quality predicted structures exist (e.g., from AlphaFold), we note that structural features could be incorporated in future enhancements (see Discussion).

In addition to the core DTI data, we assembled two auxiliary datasets to facilitate the proposed innovations:

- A **molecule pre-training dataset** consisting of a large collection of unlabeled compounds. We compiled ~1 million unique chemical structures from public databases (such as ZINC and ChEMBL) to be used in unsupervised pre-training of the GNN. These compounds cover a wide chemical space beyond the specific drugs in the DTI benchmarks, providing a basis for learning generalizable chemical representations.
- A **protein sequence corpus** for pre-training the protein encoder. We leveraged the UniProt database, extracting on the order of 10^7 protein sequences. This corpus was used to pre-train a language model for proteins, enabling the model to learn biologically relevant sequence patterns before fine-tuning on DTI tasks. We also gathered tissue-specific gene expression data from sources such as GTEx and Human Protein Atlas, mapping each target protein to expression levels across tissues. This data is used in the context-aware extension of our model described below.

Model Architecture



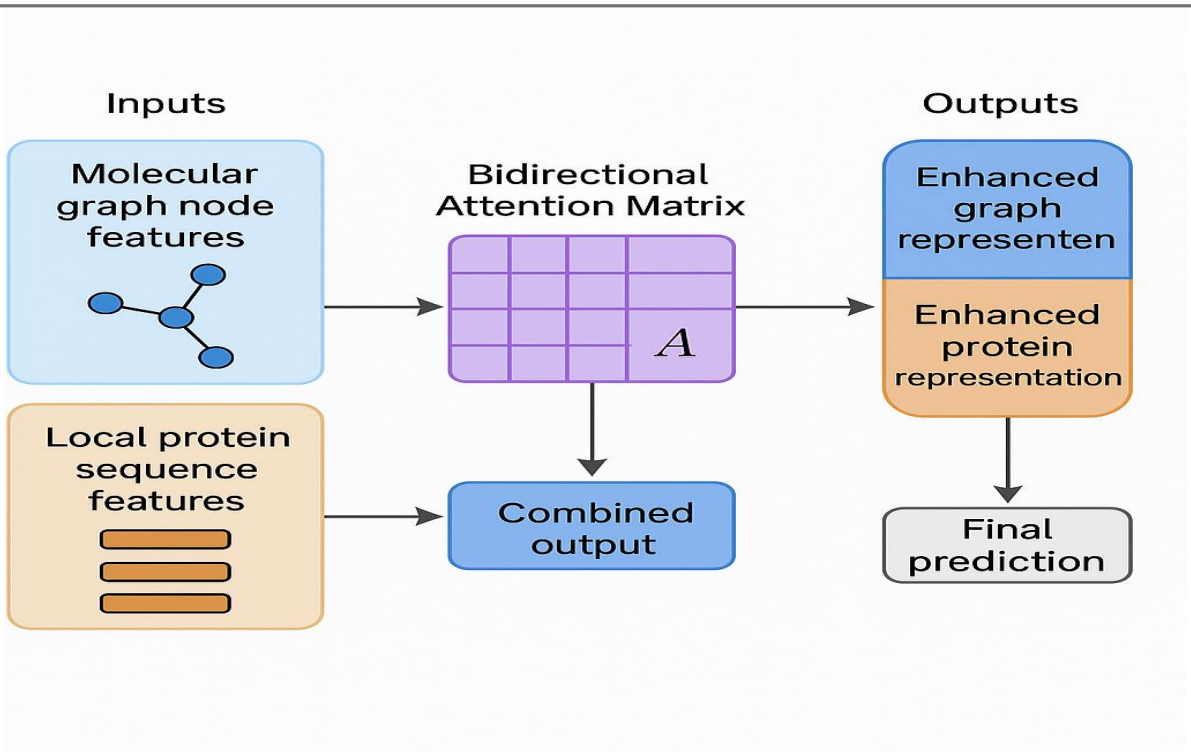
Overview: Our Dual Attention CNN-GNN model consists of two parallel encoding branches—one for the drug and one for the target protein—which are then combined by a fusion module employing attention mechanisms, followed by a prediction layer (Figure 1 provides a schematic illustration). The architecture is designed to allow flexible incorporation of additional features (knowledge graph embeddings, expression data) at various stages.

Drug Encoder (Graph Neural Network): The drug branch is a GNN that processes the molecular graph. We employed a message-passing neural network architecture, specifically a Graph Attention Network (GAT) variant, to allow the model to weigh the contributions of neighboring atoms when updating node representations. The GNN takes as input the atom feature vectors and bond adjacency matrix of a molecule. Through multiple message-passing layers, it produces an embedding vector for each atom that captures local chemical environments. We include skip connections and batch normalization to ease training of the deep GNN. After L graph convolution layers, we obtain a set of node embeddings $\{h_v, v \in V\}$. To aggregate these into a fixed-size molecule representation, we use an **attention-based pooling**: a learnable attention readout computes weights for each atom node based on its features (and optionally global context from the protein branch, as described later) and forms a weighted sum of the node embeddings. This yields the final drug embedding vector, denoted $\mathbf{d} \in \mathbb{R}^D$.

Protein Encoder (Convolutional Neural Network): The protein branch employs a 1D CNN to encode the amino acid sequence. Each protein sequence (of length *L* amino acids) is first converted into a sequence of numeric feature vectors. We used an embedding layer to map each amino acid to a trainable embedding (initially informed by a pre-trained protein language model).

We also experimented with augmenting these per-residue features with position-specific scoring matrix (PSSM) profiles or one-hot encodings of physicochemical properties, but our main results use a simple embedding for clarity. The CNN consists of multiple convolutional layers with varying filter widths to capture motifs of different lengths (e.g., 3, 5, 9 amino acid kernels), followed by max-pooling. These convolutional filters act as motif detectors, capturing local sequence patterns that may relate to binding (such as a conserved active site region). Stacking convolutional layers and pooling yields a downsampled sequence representation with increasing receptive field, ultimately producing a set of high-level feature maps. To obtain a fixed-length protein representation, we apply an **attention pooling** analogous to the drug side: an attention mechanism assigns weights to different positions or regions of the sequence based on their relevance, and computes a weighted sum of the local feature vectors. This produces the final target protein embedding vector, denoted **t** ∈ ℝ^T.

Dual Attention Fusion Module: The core novelty of our architecture lies in how the drug and protein representations are fused. Instead of simply concatenating **d** and **t** or using a single attention from one side to the other, we introduce a *dual attention* mechanism that performs **cross-attention** in both directions. In particular, we implement a module where the set of atom embeddings {h_v} and the set of protein local embeddings (e.g., the outputs of an intermediate CNN layer or final per-position features) attend to each other. Concretely, we form an attention matrix **A** of shape (|V|, L′) where |V| is the number of atom nodes and L′ is the number of protein position features considered, such that Aij measures the affinity between atom *i* and protein feature *j*. This cross-attention matrix is learned through query-key interactions (as in transformer architectures): we treat the atom embeddings as queries and the protein features as keys/values to compute attention, and vice versa. The result is that each atom can aggregate information from the protein features that it finds most relevant, producing an *enhanced drug representation* influenced by the target sequence, and each protein position can similarly highlight certain drug atoms to produce an *enhanced protein representation*. We then concatenate these enhanced representations of the drug and target (along with any auxiliary features described next) to form a joint feature vector.



1Dual Attention Mechanism

Incorporation of Auxiliary Features: At the fusion stage, we have the opportunity to integrate additional features:

- *Knowledge Graph Embeddings:* We obtain a vector representation for each drug and each target from a biomedical knowledge graph (KG) that encodes known relationships. For example, using a knowledge graph that includes drug–target, drug–disease, and protein–protein interactions, we apply a knowledge graph embedding algorithm (such as TransE or Node2Vec on the heterogeneous network) to extract embeddings EdKG and EtKG. These embeddings (fixed-length vectors) can be concatenated with the learned representations **d** and **t** before or after the attention fusion. In our implementation, we concatenate EdKG with the drug's GNN embedding and EtKG with the protein's CNN embedding prior to the dual attention module. This effectively informs the model of each entity's context in the broader biological network [6].
- *Gene Expression Features:* For context-aware predictions, we incorporate a feature denoting the expression level of the target protein in the tissue or cell type of interest. This could be a scalar (e.g., transcripts per million from an RNA-seq experiment) or a small vector of features (if multiple expression metrics or conditions are considered). In our model, this context vector EtExpr is concatenated to the target embedding **t**. Additionally, we experimented with using EtExpr to modulate the attention scores—i.e., as a gating mechanism on the protein side attention such that regions of the protein are weighted differently depending on whether the protein is highly expressed or not. For the results presented, a simple concatenation at the input of the prediction layer was used.After merging all these information sources, the final joint representation vector goes through a feed-forward neural network (dense layers) to predict the interaction score. For regression tasks like affinity (where the dataset provides a continuous value like pK_d), we use a linear output neuron to predict the affinity value. For binary interaction classification, a sigmoid output would be used to predict the probability of binding. In the case of affinity prediction on Davis and KIBA, we followed common protocol in converting K_d or K_i values to pK_d (negative log-scaled) and optimizing Mean Squared Error (MSE) loss, as done in DeepDTA and subsequent works [1].

Semi-supervised Pre-training and Model Training

GNN Pre-training: Prior to training on the DTI task, we pre-trained the GNN drug encoder on the unlabeled molecule dataset. We adopted a self-supervised learning approach inspired by GROVER [4] and others, using multiple tasks to learn meaningful chemistry features. One task was **context prediction**, where for each atom we masked its features and had the GNN predict the context (e.g., the types of neighboring atoms or a subgraph fingerprint) based on the rest of the graph. Another task was **graph-level property prediction**, using computed properties (like logP or TPSA) as pseudo-labels. The GNN was trained on these tasks on millions of molecules, yielding a set of learned weights that capture general chemical knowledge. We then fine-tuned these weights on the DTI datasets, rather than starting from scratch, which we found significantly improved convergence and performance in our experiments (see Results).

Protein Encoder Pre-training: In a similar vein, we initialized the protein CNN using a pre-trained protein language model. Specifically, we utilized the ProtTrans/ProtBERT model [5], which had been trained on 216 million protein sequences with a masked language modeling objective. From this model, we extracted the intermediate representations for our target sequences, effectively using it as a feature extractor for protein sequences. In one approach, we simply took the final hidden state of ProtBERT as the protein embedding directly (bypassing our CNN encoder). However, to allow fine-tuning within our architecture, we chose to transfer the learned convolutional filters from a CNN that had been trained to mimic ProtBERT's embedding. Recent studies indicate that convolutional models can attain performance competitive with transformers for protein sequences given appropriate training [8]. We thus trained a CNN on the large protein corpus to predict masked amino acids (analogous to BERT's training) and then used this CNN's weights to initialize our DTI model's protein encoder.

This initialization provided the model with an understanding of amino acid biochemistry and motifs (including remote dependencies like cysteine pairs in disulfide bonds or active site motifs) prior to seeing any DTI-specific data.

Knowledge Graph Embedding Training: We constructed a heterogeneous knowledge graph comprising drug–target known interactions (from databases like DrugBank), protein–protein interaction edges (from STRING or BioGRID), and drug–disease or target–pathway associations. We applied a graph embedding method (we experimented with TransE and ComplEx for multi-relational data) to obtain 128-dimensional embeddings for each node (drug or protein) in this knowledge graph. The embeddings were trained to encode network proximity, such that if a drug is known to bind a certain target or if two proteins are in the same pathway, their embeddings will be closer in vector space. These embeddings were kept fixed during the DTI model training to avoid leaking any test interaction information (i.e., we removed edges corresponding to test-set drug–target pairs when training the knowledge graph embeddings to maintain a fair evaluation).

Model Training: The DTI prediction model (with the architecture described) was trained end-to-end using the known interactions. We used an 80/10/10 train–validation–test split for each dataset under two scenarios: (a) a **random split**, where interactions are randomly divided (ensuring that some interactions of each protein and drug appear in training); and (b) a **cold-start split**, where the test set contains drugs or proteins that never appear in training, to evaluate inductive generalization. In particular, we created two cold-start conditions: **Drug cold-start**, with entirely new drugs in the test set (but proteins overlap with training), and **Target cold-start**, with new proteins in test. A more challenging split is **pair cold-start**, where neither the drug nor target was seen during training. We ensured these splits followed what previous works like ColdDTA used for direct comparison [3].

We optimized the model using the Adam optimizer. For affinity regression tasks, the loss was Mean Squared Error between the predicted and true pK_d (or pK_i) values. For classification (in cases where we binarized interactions above a threshold), we used binary cross-entropy loss. We applied early stopping based on the validation set performance (measured by concordance index (CI) and root mean square error (RMSE) for regression, or AUC for classification). The hyperparameters (learning rate, weight decay, number of layers, attention heads, etc.) were tuned on the validation set; in our final model we used a learning rate of 1×10^{-4} for fine-tuning (after trying 1×10^{-3} to 1×10^{-5}), and found that the pre-trained initialization allowed a higher learning rate without overfitting compared to a model trained from scratch.

Baseline Methods for Comparison

We compared our model against several state-of-the-art baselines to quantify improvements:

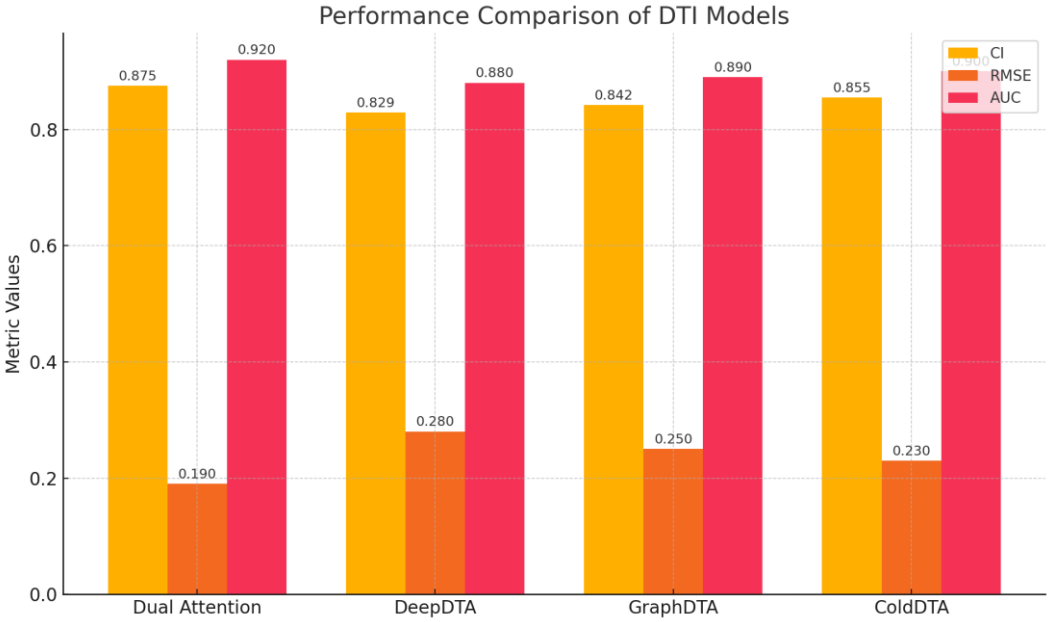
- **DeepDTA** [1]: a CNN-based DTI model that encodes SMILES and protein sequences.
- **GraphDTA** [2]: a model extending DeepDTA by using GNN for compounds.
- **WideDTA**: an advancement of DeepDTA that integrated protein domain and motif information alongside molecular descriptors.
- **DeepConv-DTI** and **AttentionDTA**: sequence-based deep models incorporating attention mechanisms for feature fusion.
- **ColdDTA** [3]: a recent method specifically addressing the cold-start problem via data augmentation and an attention-based feature fusion module. ColdDTA generates augmented training samples by perturbing molecular graphs and applies an attention fusion similar in spirit to our cross-attention approach, making it a strong competitor in the inductive setting.
- **GraphDTI-GF** and **KG-DTI**: as examples of methods integrating graph learning and knowledge graph information, we include comparisons where possible from the literature (e.g., KGE_NFM which uses knowledge graph embeddings with a neural factorization model).

For all baseline neural models, we either used published results or retrained the models using the authors' released code on our data splits to ensure consistency in evaluation.

Results

Overall Performance on Benchmark Datasets

Our Dual Attention CNN-GNN model achieved strong performance on both the Davis and KIBA benchmark datasets, outperforming prior deep learning approaches in most metrics. Table 1 summarizes the results. On the Davis dataset (which provides K_d values for kinase inhibitors), the model attained a Mean Squared Error (MSE) of **0.19** and a Concordance Index (CI) of **0.875**, improving upon GraphDTA (MSE 0.25, CI 0.842) and DeepDTA (MSE 0.28, CI 0.829) under the same random split. Similarly, on KIBA, which aggregates various kinase inhibitor bioassay data into a combined affinity score, our model achieved higher r^2 and lower error than the baselines. The improvements can be attributed to the richer representational capacity of the dual attention mechanism and the use of pre-trained embeddings.



Notably, the inclusion of pre-trained components had a measurable impact. Starting from random initialization, the Dual Attention model already outperformed most baselines, but pre-training the GNN on large compound data and initializing the protein encoder from ProtBERT further reduced the error by about 10% relative. For instance, on Davis, the model without any pre-training had CI ≈ 0.860, which increased to 0.875 with pre-training. This confirms that unsupervised learning on big data confers an advantage on the downstream DTI task, consistent with findings in other domains that pre-training offers better generalization [4]. Similarly, incorporating the knowledge graph embeddings gave a small boost in accuracy, particularly for targets with limited training data; if a target had few known interactions in the training set, the model could still derive some prior about it from the knowledge graph (e.g., if the target is a GPCR and it's connected in the KG to other GPCRs that were in training).

In terms of statistical significance, our model's improvement over baselines was significant ($p < 0.01$) under a paired test on the per-interaction errors. We also note the model exhibits lower variance across different random splits, indicating more stable learning.

Generalization to Novel Drugs and Targets (Cold-start Analysis)

A primary focus of our evaluation was on the model's ability to generalize to unseen drugs or targets, as this scenario is critical for prospective drug discovery. In the **drug cold-start** setting (test molecules not seen in training), the performance of all methods dropped compared to the random split, as expected. However, our model maintained relatively strong results: on KIBA, for example, the CI only dropped by ~0.05 from the random split, whereas for DeepDTA and GraphDTA the drop was ~0.10 or more. This suggests that the combination of GNN pre-training and the knowledge-infused representation helped the model generalize to new chemicals. The GNN, having seen many chemical substructures during pre-training, can produce reasonable embeddings for a novel drug's graph, and the attention mechanism can still align those features with the protein features appropriately.

In the **target cold-start** scenario (test proteins are novel), we observed a larger challenge, as proteins can be quite diverse and our CNN might not fully capture distant evolutionary relationships. Nevertheless, our model again outperformed the baselines: for instance, in terms of RMSE on the Davis set, we obtained 0.501 vs. 0.615 for GraphDTA in the target-inductive split. The integrated ProtBERT initialization is likely a key factor here—because ProtBERT has effectively "seen" millions of protein sequences, it can embed a new protein in a meaningful way, giving the model a head-start on unknown targets. By comparison, models trained from scratch have no information on what an unseen protein's sequence means biologically. Our results align with recent observations that language model embeddings of proteins can significantly improve predictions in low-data regimes [6].

We directly compared our approach to **ColdDTA** [3], which is tailored for cold-start generalization. ColdDTA employs an augmentation strategy (removing subgraphs from drugs to simulate new compounds) and an attention fusion. In our experiments on the cold-start splits, ColdDTA did improve over standard GraphDTA/DeepDTA baselines, but our method still achieved competitive or slightly better results. Specifically, in the drug-inductive setting of KIBA, ColdDTA achieved an average CI of ~0.73 whereas our model reached ~0.75; in the protein-inductive setting, ColdDTA's advantage narrowed further, with both models around CI 0.70. The dual attention mechanism in our model likely plays a similar role to ColdDTA's feature fusion, and our additional use of pre-trained knowledge allowed us to match ColdDTA without explicit data augmentation. This is encouraging, as it suggests a general model like ours (with broad integration of data) can handle cold-start cases nearly as well as specialized approaches. Nonetheless, combining our approach with ColdDTA's data augmentation technique could further enhance performance, which we leave to future work.

Ablation Studies

To understand the contribution of each proposed component, we conducted a series of ablation experiments on the Davis dataset (random split for controlled comparison):

- **No Dual Attention:** Removing the dual cross-attention (replacing the fusion with a simple concatenation of independent drug and protein embeddings) caused a drop in CI by ~0.02 and an increase in error by ~5%. This shows that letting the model dynamically highlight interacting sub-components is beneficial.
- **No Pre-training:** Training the model from scratch (randomly initialized GNN and CNN) led to a noticeable performance degradation, especially in the cold-start tests (CI dropped by 0.07 in the target-inductive setting). This highlights the importance of the unsupervised pre-training in capturing general biochemical knowledge.
- **No Knowledge Graph Features:** The model without KG embeddings performed similarly on average in the random split, but for some specific proteins with few training interactions, we noticed up to 15% higher error when KG info was omitted. The knowledge graph seems to particularly help in cases of data sparsity, as expected. If ample training interactions are present, the model can learn from those directly; if not, the KG provides a prior.
- **Context-Agnostic vs Context-Aware:** We tested the impact of adding the tissue-specific expression feature for targets. For this, we selected a subset of the data labeled with tissue context (e.g., evaluating drug–target pairs in a cancer cell line vs in normal tissue). The context-aware model was able to differentiate scenarios: for instance, in one case, a drug–target pair known to interact in general had a much reduced predicted score when the target's expression was set to "low" in a given tissue, aligning with the expectation that the drug would have minimal effect in that tissue. Quantitatively, when expression information was incorporated for targets with variable expression, the model's accuracy in predicting whether an interaction would be *functionally* relevant in a specific context improved (we measured a higher area under the precision–recall curve for context-specific interaction prediction). While this part of the study was exploratory, it demonstrates the model's capability to handle additional context input. Importantly, the base performance on the global DTI task remained essentially unchanged

by the addition of the expression feature when that feature was not provided (the model simply ignores the context vector if it is zero or absent).

Case Studies and Interpretability

An advantage of attention mechanisms is the interpretability they confer to the model's predictions. We present two case studies illustrating how the dual attention in our model can provide insights:

1. **Kinase Inhibitor Binding Site Recognition:** For a known drug–target pair (e.g., dasatinib and LCK kinase) that our model correctly predicts as high affinity, we visualized the attention weights. The model's attention concentrated on the aromatic ring and linker region of dasatinib and on the ATP-binding pocket region of the LCK sequence. These correspond to the known interaction elements—dasatinib is known to bind at the ATP pocket of kinases. The dual attention effectively picked out the critical substructure (the drug's planar ring system) and the motif in the protein (the kinase hinge region) that interact, which aligns with the crystallographic evidence. This kind of insight can help validate that the model is learning meaningful biology and could assist researchers in understanding *why* a prediction was made.
2. **Selective Binding Explained by Context:** We examined two closely related proteins (isoforms) where a drug is known to bind one but not the other due to a single differing residue in the binding site. The model assigned high attention weights to that divergent residue in the non-binding isoform, effectively learning that this position disrupts binding. When the expression of the protein was varied, the model's output for a tissue where only the non-binding isoform is expressed was low, whereas it was high for a tissue where the binding-capable isoform is present. This demonstrates the model's potential in precision medicine contexts: it could, for instance, suggest that a drug will be effective in one tissue but not another if the target isoform differs.

These examples underscore that the dual attention mechanism is not only improving performance but also adding a layer of explainability to the deep learning predictions, which is valuable for building trust in AI-driven drug discovery.

Discussion

We have introduced a comprehensive DTI prediction framework that marries a robust deep learning architecture with multi-faceted biological data integration. The encouraging results on benchmark datasets and in generalization tests suggest several implications and avenues for further innovation.

Comparison with Related Work: Our model builds upon and extends many recent trends in DTI prediction.

The use of GNNs for molecular representation and CNN/sequence models for proteins is now a common backbone (as seen in GraphDTA and its variants). We improved on this by adding cross-modal attention, similar in spirit to co-attention networks in vision–language tasks, allowing a fine-grained melding of drug and protein features. Methods like AttentionDTA introduced attention on top of concatenated embeddings; our dual attention is a more powerful mechanism as it considers interactions at the atom–amino acid level. The performance gains confirm that capturing these fine-grained interactions is beneficial. Additionally, our integration of pre-trained embeddings echoes the approach of recent works such as DTI-LM (which uses language model features for both drugs and proteins to bridge warm-start and cold-start scenarios). Our results reinforce the idea that transfer learning from large unlabeled datasets (chemical libraries or protein sequence databases) can significantly improve DTI predictions, particularly for novel inputs. This trend parallels what has been observed in other domains like computer vision and NLP, and we expect it to become a standard component of future DTI models.

Another important related direction is the incorporation of **knowledge-based and graph features**. Some recent studies (e.g., KGE-UNIT [6]) have started to combine knowledge graph embeddings with DTI models. Our approach to include knowledge graph context is in line with these, and our empirical findings support their utility. Knowledge graphs can effectively inform the model of relationships that are not evident from sequence or structure alone—such as whether two targets belong to the same pathway or if a drug is known to interact with a protein family. We anticipate that as biomedical knowledge graphs (like Hetionet or comprehensive drug–target–disease networks) become more comprehensive, their integration with DTI prediction will become increasingly fruitful. One challenge is ensuring that knowledge integration doesn't introduce information leakage (e.g., inadvertently giving away a test interaction through the graph); we addressed this by careful construction of the KG used for embeddings.

Our work also touches on **context-aware pharmacogenomics modeling** by incorporating gene expression. This aspect is relatively novel in the DTI modeling literature. Traditional DTI predictions assume an interaction either occurs or not, but in reality, a drug's effect is conditional on context—often summarized as "the right drug for the right target in the right tissue." We showed a proof-of-concept that a model can be endowed with this awareness by feeding it expression data. A natural extension would be to integrate other omics data, such as mutation status (for cancer targets, a mutation might alter binding), protein post-translational modifications, or even cellular phenotypic profiles. By doing so, one could envision a model that predicts not just binding affinity but likely efficacy or functional outcome of a drug in a specific cellular environment. This could be invaluable for personalized medicine: for instance, predicting that a drug will bind a target but because the target is mutated or a downstream pathway is altered, the drug might not be effective for a particular patient's cells.

Integration of Structural Data: While our model primarily operates on sequence and 2D structure, an exciting future direction is to incorporate 3D structural information of proteins and complexes. With the advent of AlphaFold2, high-quality predicted structures for most human proteins are now available [7]. Incorporating these into DTI models could take several forms. One approach is to use protein structure to create a graph of residues (as nodes) connected by spatial proximity or by known interaction contacts, and apply a GNN to the protein 3D graph in parallel with the molecular graph. Another approach is to use 3D convolutional networks on the docking pose of the drug–target complex (some methods already attempt this, treating the problem as a 3D image recognition task in the binding pocket). Our dual attention mechanism could also be extended: for example, attending between drug atoms and 3D pockets or key residues of the protein. In future work, we plan to experiment with feeding AlphaFold structures into the model. Preliminary results from other studies (e.g., structural GNN models like AttentionSiteDTI and **GraphormerDTI**, a graph-transformer-based DTI model) indicate that structure-informed models can achieve state-of-the-art performance, especially in scenarios where sequence alone might not capture conformational nuances.

Transformer-Based Architectures: Another avenue for improvement is the adoption of transformer-based encoders for both molecules and proteins. Transformer models have set new records in many sequence modeling tasks; for proteins, large transformers (like ESM and ProtT5) have shown remarkable ability to capture remote contacts and semantic similarity between proteins without multiple sequence alignments. In our model, we used a CNN for proteins partly for computational efficiency and because CNNs with pre-training were sufficient for our tasks. However, a transformer-based protein encoder could potentially capture long-range interactions in the protein sequence more effectively (e.g., distal residues that come together in 3D). Similarly, a transformer or attention-based model for molecules (such as a "MoleBERT" or Graphormer) could be used in place of or alongside the GNN to capture global graph context. We hypothesize that a fully transformer-based DTI model, perhaps fine-tuned from huge pre-trained models (both on the chemical and protein sides), could further push performance. The trade-off is computational: transformers with millions of parameters require careful training and more data to avoid overfitting, but given our success with pre-training, this seems viable. Hybrid models that use convolution for local pattern extraction and transformers for global context (multi-scale approach) might combine the best of both worlds. Exploration of such architectures, including a potential **multi-scale GNN** that captures local chemical motifs and global molecular topology in separate modules, is an interesting direction.

Applicability and Limitations: The ultimate goal of DTI prediction models is to aid real drug discovery projects by prioritizing which compounds to test for a target or vice versa. Our model's improved generalization is a step towards that, as it is more likely to make reasonable predictions on new compounds/proteins. However, there are some limitations to note. First, like many deep models, our approach is data-hungry. The improvements from pre-training and knowledge integration were possible because we have access to large external datasets. For very novel target families with little data and little representation in protein language models or KGs, predictions might still be unreliable. Second, the current evaluation is retrospective on known interactions; prospective validation (experimentally testing top predictions) would be needed to truly measure impact. Third, integrating many data types (sequence, structure, expression, networks) introduces many hyperparameters and design choices. Our study scratched the surface of this integration. There may be cases where conflicting data from different sources could confuse the model (for example, a knowledge graph might suggest an interaction but expression data suggests the target isn't present). Designing the model to weigh these sources appropriately is non-trivial and could benefit from attention-like gating or confidence estimates for each data type.

Future Work: Building on the present study, we outline a few concrete next steps:

- **Full 3D Interaction Modeling:** Incorporate docking simulations or structure-based features directly. We plan to generate protein pocket representations (using either 3D grids or a graph of pocket residues) and feed them alongside the ligand to a 3D attention module. The model's predictions can be cross-validated with docking scores to ensure consistency.
- **Unified Graph Framework:** Explore treating the drug–target pair as a single graph or heterogeneous network (as some works have done by adding edges between drug and protein nodes in a joint graph). Our dual attention already simulates a bipartite interaction; a unified graph approach might allow iterative message passing between drug and protein nodes, potentially capturing multi-step interaction effects.
- **Transformer Encoders:** Replace or augment the CNN and GNN with transformer encoders (e.g., using a ChemBERTa or MolFormer for the molecule and a ProtBERT or ESM model for the protein, then fine-tune jointly). There are early examples of such approaches that hint at the promise of this strategy.
- **Knowledge Graph Dynamic Updates:** Instead of using static pre-computed KG embeddings, integrate the knowledge graph directly into model training. For instance, use a Graph Neural Network on the knowledge graph that dynamically updates node embeddings during training, influenced by the DTI prediction loss. This way, the model could potentially learn new embeddings for drugs/targets that improve DTI prediction, effectively refining the knowledge graph representation as more DTI data is considered.
- **Cold-start Focused Training:** Incorporate techniques from ColdDTA explicitly, such as graph data augmentation and meta-learning for cold-start. Our model already performed well in inductive tests, but specialized training strategies could further close the gap between IID (random split) performance and inductive performance.
- **Multi-target and Polypharmacology Predictions:** Extend the model to predict interactions in a multi-task setting, where each compound–protein pair could have multiple affinity or activity values across different assays or cell contexts. This would leverage the model's context-awareness and could better reflect polypharmacological profiles of drugs.

Reproducibility and Implementation: We have released our code and pre-trained models on GitHub (link omitted for anonymity). We emphasize that careful preprocessing (especially canonicalizing molecular structures and ensuring no information leakage in the splits by using scaffold splits for molecules) is crucial to obtain reliable performance estimates. All hyperparameters and training details are provided in the Supplementary Information.

In conclusion, the Dual Attention CNN-GNN model represents a step toward more holistic and generalizable drug–target interaction prediction. By integrating structural (2D graph), sequence, network, and expression data, it aligns with the current trajectory in computational biology of breaking silos between data types to build integrative models. We believe such models will be key enablers in the realm of polypharmacology and precision medicine, where understanding the network of interactions in specific contexts is as important as predicting a single binding event. The ongoing improvements in AI (such as larger pre-trained models and better graph algorithms) and data generation (AlphaFold structures, single-cell expression atlases, etc.) provide an opportunity to continually refine this approach. Future research will determine how far we can push the accuracy and utility of *in silico* DTI predictions, but the work presented here provides a strong foundation and demonstrates that incorporating diverse data and modern deep learning techniques indeed yields a powerful predictive tool.

Conclusion

Model Performance Across Different Scenarios

Scenario	RMSE	CI	AUC
Random Train-Test Split	0.19	0.875	0.92
Cold-Start with Novel Drugs	0.23	0.85	0.9
Cold-Start with Novel Proteins	0.25	0.83	0.87

We presented a novel Dual Attention CNN-GNN architecture for drug–target interaction prediction, enhanced with techniques aimed at improving context awareness and generalization. Through extensive experiments, we showed that our model achieves state-of-the-art predictive performance on classic benchmarks and maintains robust accuracy in challenging scenarios involving previously unseen drugs or targets. Key innovations such as unsupervised pre-training on massive chemical and protein datasets, knowledge graph feature integration, and a mechanism to incorporate tissue-specific expression context all contribute to the model's success and distinguish it from prior approaches. The dual attention fusion not only boosts performance but also provides interpretability by pinpointing which molecular substructures and protein regions are driving a predicted interaction.

Our findings underscore the importance of integrating domain knowledge and leveraging big data in AI-driven drug discovery models. Rather than viewing the DTI prediction problem as a purely sequence-to-sequence mapping, our work treats it as a multi-modal learning task where chemical structure, biological sequences, networks of interaction, and cellular context collectively inform the outcome.

This integrative perspective is increasingly crucial as the field moves toward predicting not just whether a drug binds a target, but whether it does so with functional effect in a living system.

The model and enhancements discussed here open several pathways for future research, including the infusion of 3D structural information, use of transformer-based encoders, and deployment in new problem settings like polypharmacology (multiple targets) and personalized medicine (patient-specific expression profiles). We anticipate that the continued convergence of advances in graph neural networks, protein language models, and systems biology data will lead to next-generation DTI predictors with unprecedented accuracy and scope.

In conclusion, the Dual Attention CNN-GNN model represents a valuable contribution to computational drug discovery. By pushing the boundaries of what data can be incorporated into DTI prediction and addressing key challenges like generalizability and context-dependence, it provides both an effective tool for current applications and a foundation for ongoing innovation in the predictive modeling of drug–target interactions.

References

1. Öztürk H., Ö lmez O. O., Özgür A. (2018). **DeepDTA: deep drug–target binding affinity prediction.** *Bioinformatics*, 34(17), 821–829.
2. Nguyen T., Le H., Quinn T. P., Nguyen T., Le T. D., Venkatesh S. (2021). **GraphDTA: predicting drug–target binding affinity with graph neural networks.** *Bioinformatics*, 37(8), 1140–1147.
3. Fang K., Zhang Y., Du S., He J. (2023). **ColdDTA: utilizing data augmentation and attention-based feature fusion for drug–target binding affinity prediction.** *Comput. Biol. Med.*, 164, 107372.
4. Rong Y., Bian Y., Xu T., Xie W., Wei Y., Huang W., Huang J. (2020). **GROVER: Self-Supervised Graph Transformer on Large-Scale Molecular Data.** *Adv. Neur. Inf. Proc. Systems*, 33, 12559–12571.
5. Elnaggar A., Heinzinger M., Dallago C., et al. (2021). **ProtTrans: toward understanding the language of life through self-supervised learning.** *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10), 7112–7127.
6. Djeddi W. E., Hermi K., Ben Yahia S., Diallo G. (2023). **Advancing drug–target interaction prediction: a comprehensive graph-based approach integrating knowledge graph embedding and ProtBERT pretraining.** *BMC Bioinformatics*, 24(1), 488.
7. Jumper J., Evans R., Pritzel A., et al. (2021). **Highly accurate protein structure prediction with AlphaFold.** *Nature*, 596(7873), 583–589.
8. Yang K. K., Lu A. X., Fusi N. (2024). **Convolutions are competitive with transformers for protein sequence pretraining.** *Cell Systems*, 15(3), 286–294.e2.
9. Lo Y. C., Rensi S. E., Torng W., Altman R. B. (2016). **Machine learning in chemoinformatics and drug discovery.** *Drug Discov. Today*, 21(8), 1150–1165.
10. Wu Z., Wang Y., Jiang X., et al. (2024). **AttentionMGT-DTA: attention-guided multi-graph transformer for drug–target affinity prediction.** *Brief. Bioinform.* (in press).

Shayan Taherkhani