Pharmacophore

ISSN-2229-5402

Journal home page: http://www.pharmacophorejournal.com

MULTIMODAL DEEP LEARNING FOR DRUG–TARGET INTERACTION PREDICTION USING MOLECULAR GRAPHS AND PROTEIN EMBEDDINGS

Ravi Kumar[1]*, Neha Sharma[1], Aniket Deshmukh[2], Arjun Nair[1], Meera Pillai[2]

1. ***Department of Computational Pharmacy and Drug Analytics, Faculty of Engineering, School of Pharmaceutical Technology, IIT Delhi, New Delhi, India.***
2. ***Department of Artificial Intelligence in Drug Systems, Faculty of Pharmacy, IIT Bombay, Mumbai, India.***

ARTICLE INFO

Received:
14 November 2024
Received in revised form:
16 February 2025
Accepted:
21 February 2025
Available online:
28 February 2025

Keywords: Drug–target interaction prediction, Multimodal deep learning, Molecular graphs, Protein embeddings, Graph attention networks, Protein language models

ABSTRACT

Identifying the proteins that a drug interacts with is essential for understanding its efficacy, selectivity, and safety, yet experimental profiling cannot feasibly screen all possible drug–target pairs across the proteome, highlighting the need for scalable computational prediction. Current in silico models often rely on either ligand-based or protein-based features in isolation, which can overlook complementary information arising from joint modeling of molecular structure and target biology. To address this, we propose a conceptual multimodal deep learning model that learns from molecular graphs on the ligand side and protein sequence embeddings on the target side, enabling prediction of both binding affinity and binary interaction status. The model employs a graph attention network to encode the molecular graph of each compound and a pre-trained protein language model to encode the target sequence, with a bilinear attention mechanism fusing ligand and protein representations into a joint embedding for downstream affinity regression and interaction classification. This approach is expected to deliver strong predictive performance on drug–target interaction benchmarks, generalize to unseen targets through protein language representations, and enhance interpretability via attention maps that highlight pharmacophoric substructures and relevant protein regions. By combining predictive modeling, biological generalization, and interpretable ligand–target reasoning, this multimodal framework has the potential to accelerate drug repurposing, selectivity profiling, and virtual screening.

To Cite This Article: Kumar R, Sharma N, Deshmukh A, Nair A, Pillai M. Multimodal Deep Learning for Drug–Target Interaction Prediction Using Molecular Graphs and Protein Embeddings. Pharmacophore. 2025;16(1):40-9. https://doi.org/10.51847/DexOGUwfOU

Introduction

Drug–target interactions define how small molecules influence biological systems, whether by binding orthosteric pockets, modulating allosteric sites, inhibiting catalytic activity, or perturbing signaling pathways. Because experimental screening cannot exhaustively evaluate all possible compound–protein pairs, computational models have become essential for prioritizing candidates before biochemical or cellular validation [1]. Deep learning approaches such as DeepDTA reframed drug–target affinity prediction as a representation learning problem, showing that sequence-derived compound and protein features could be mapped to continuous affinity estimates [1]. Subsequent models extended this direction by incorporating richer molecular and protein encoders to support drug repurposing, selectivity profiling, and virtual screening workflows [2, 3].

The field has evolved from ligand similarity and hand-crafted descriptor methods toward proteochemometric and deep learning models that encode both compound and target information. DeepAffinity introduced a unified neural architecture for compound–protein affinity prediction, emphasizing that both molecular and protein features should contribute to the learned interaction representation [2]. DeepConv-DTI further showed that convolutional modeling over protein sequences could support interaction prediction without requiring experimentally resolved structures [4]. Cross-domain architectures such as DeepCDA then strengthened the idea that ligand and protein encoders should be trained as complementary sources of information rather than as isolated predictors [5].

Corresponding Author: Ravi Kumar; Department of Computational Pharmacy and Drug Analytics, Faculty of Engineering, School of Pharmaceutical Technology, IIT Delhi, New Delhi, India. E-mail: ravi.kumar@outlook.com.

Graph neural networks and protein language models represent two major advances that motivate the proposed architecture. GraphDTA demonstrated that molecular graphs can improve drug–target binding affinity prediction by learning directly from atom–bond topology rather than relying only on linearized SMILES strings [6]. In parallel, ProtTrans showed that self-supervised protein language models can learn informative sequence representations from large protein corpora, offering a general-purpose alternative to hand-crafted protein descriptors [7]. The remaining architectural challenge is not simply to encode ligands and proteins separately, but to fuse their representations in a way that captures pairwise molecular substructures and protein regions relevant to binding [8, 9].

This manuscript proposes a multimodal deep learning model that combines a graph attention network for ligand encoding with a pre-trained protein transformer for target encoding, followed by cross-modal bilinear attention. Related multimodal models such as MONN, MolTrans, and CoaDTI illustrate the value of modeling local compound–protein interaction patterns rather than relying only on global concatenation [3, 8, 9]. The proposed architecture is therefore model-oriented: it is designed to produce affinity and interaction predictions while also exposing attention maps that can be inspected for medicinal chemistry interpretation. Its intended contribution is a coherent conceptual design for graph-based ligand learning, protein embedding-based target learning, and interpretable multimodal fusion.

Background

Drug–Target Interaction Prediction in Drug Discovery

Drug–target interaction prediction supports several core tasks in computational drug discovery, including target-based lead prioritization, off-target risk assessment, polypharmacology analysis, and drug repurposing. BindingDB, DAVIS, KIBA, DrugBank, BIOSNAP, and related resources are commonly used to train and evaluate models that predict either continuous affinity or binary interaction status [1, 6]. Benchmark-focused studies such as DeepDTA and GraphDTA helped standardize the use of affinity regression settings, while heterogeneous graph frameworks broadened DTI prediction toward network-based drug and target relationships [1, 6, 10]. In practical deployment, a DTI model should therefore be evaluated not only for random-pair prediction, but also for its ability to reason about unfamiliar compounds, under-characterized targets, and clinically relevant off-target interactions [10, 11].

Molecular Graph Representations and Graph Neural Networks

Molecular graph representations treat atoms as nodes and bonds as edges, allowing neural networks to learn from the chemical topology that constrains molecular recognition. Graph neural networks, including graph convolutional networks, graph attention networks, and message-passing neural networks, are well suited to this representation because they aggregate local neighborhood information into atom- and molecule-level embeddings [6]. MGraphDTA showed how multiscale graph modeling can support explainable drug–target affinity prediction, while GraphscoreDTA further emphasized optimized graph learning for protein–ligand binding affinity tasks [12, 13]. These graph-based approaches are especially relevant for ligand encoding because pharmacophoric patterns often emerge from connected substructures rather than from independent molecular descriptors [14].

Protein Sequence Embeddings from Language Models

Protein sequence embeddings derived from pre-trained language models provide contextualized residue representations that can encode evolutionary, structural, and functional regularities from primary sequence alone. ProtTrans demonstrated that large self-supervised protein models can learn representations useful across biological prediction tasks, making them attractive encoders for targets whose structures may be unavailable or incomplete [7]. Models that incorporate protein sequence learning, such as DeepConv-DTI and DeepMHADTA, show that sequence-based encoders can support drug–target prediction without requiring docking-ready protein conformations [4, 15]. More recent contrastive and multimodal approaches further suggest that protein language space can be aligned with compound representations to improve interaction modeling and generalization [11, 16].

Multimodal Fusion Strategies in Deep Learning

Multimodal fusion strategies determine how ligand and protein representations interact inside a predictive model. Early concatenation simply joins global embeddings, while late fusion combines separate prediction streams, but both strategies can underrepresent fine-grained atom–residue dependencies. Attention-based and bilinear mechanisms, as used in MolTrans, FusionDTA, CoaDTI, and MCL-DTI, allow the model to learn cross-modal relationships between molecular substructures and protein sequence regions [8, 9, 17, 18]. These mechanisms are especially important for DTI prediction because binding is not a property of a ligand or target alone, but of their chemically and biologically compatible interaction [19, 20].

Prior DTI Models and the Gap in Interpretable Multimodal Architectures

Prior DTI models provide a foundation for the proposed architecture, but they also reveal a gap between predictive fusion and interpretable biological reasoning. DeepDTA established a sequence-based deep learning baseline, DeepAffinity emphasized interpretability in compound–protein affinity learning, and GraphDTA introduced graph neural networks for ligand representation [1, 2, 6]. Transformer- and attention-based models such as MolTrans, MONN, CoaDTI, and MMDTA advanced multimodal interaction modeling, but many architectures still provide limited residue- and substructure-level interpretability

suitable for medicinal chemistry decision-making [3, 8, 9, 21]. The proposed model responds to this gap by centering its design on graph attention, protein language embeddings, and an explicit ligand–protein attention map.

Table 1 compares the major representation strategies used in DTI prediction and clarifies why the proposed model combines molecular graphs, protein embeddings, multimodal attention, and uncertainty-aware reporting.

Table 1. Comparison of Core Representation Strategies for Drug–Target Interaction Prediction

Representation strategy	Drug-side information captured	Target-side information captured	Strength	Main limitation
Sequence-only encoding	SMILES string patterns	Amino-acid sequence patterns	Simple and scalable	May miss molecular topology and local chemical structure
Descriptor-based encoding	Hand-crafted molecular descriptors	Hand-crafted protein descriptors	Interpretable and computationally light	Depends heavily on feature engineering
Molecular graph encoding	Atom–bond topology and substructure relationships	Usually paired with sequence or structural features	Captures chemical connectivity directly	May omit 3D conformation and solvent effects
Protein language-model encoding	Usually combined with ligand features	Contextual residue and sequence representations	Supports transfer to targets without known structures	Embedding relevance depends on pre-training coverage
Structure-aware encoding	Molecular geometry or docking-related features	Protein structure, contact maps, or binding pockets	Adds spatial interaction context	Requires reliable structural data and higher computation
Multimodal attention fusion	Ligand substructures linked to protein regions	Protein regions linked to ligand features	Improves interaction-specific reasoning	Attention must be interpreted cautiously
Uncertainty-aware multimodal encoding	Drug and target features with reliability estimates	Confidence-aware target-side predictions	Helps identify unsupported predictions	Requires careful calibration and validation

Model Development Overview

High-Level Model Architecture

The proposed model accepts a ligand represented by a SMILES string and a target represented by an amino-acid sequence, then transforms each input into a modality-specific neural representation. The ligand branch converts the compound into a molecular graph and applies a graph attention network, following the motivation of GraphDTA and multiscale graph affinity models that learn from chemical topology [6, 12]. The protein branch passes the target sequence through a pre-trained protein transformer, building on the premise that sequence embeddings can encode functional and structural regularities relevant to binding [7, 11]. A bilinear attention fusion module then produces a joint ligand–target embedding that feeds two output heads, one for affinity regression and one for binary interaction classification [17, 18].

Core Input Representations

Ligands are represented as molecular graphs with atom type, degree, formal charge, hybridization, aromaticity, stereochemical indicators, and other chemically meaningful node attributes, while bond type and conjugation define edge features. This design follows graph-based DTI models that treat molecular topology as central to learning ligand representations for affinity and interaction prediction [6, 13]. Proteins are represented as tokenized amino-acid sequences processed by a frozen or fine-tuned transformer encoder, consistent with the use of sequence-based protein representations in DeepConv-DTI, ProtTrans, and contrastive protein language-space modeling [4, 7, 11]. Long or multidomain proteins can be handled through windowing, residue pooling, or hierarchical aggregation so that protein embeddings remain compatible with ligand graph outputs.

Design Principles

The proposed model is multimodal by design, because it explicitly learns from chemical structure and target sequence rather than compressing either modality into hand-crafted descriptors. It is end-to-end trainable, but it can also use frozen protein embeddings when the available DTI labels are limited, a strategy motivated by transfer-oriented protein representation learning [7, 11]. Interpretability is built into the architecture through graph attention and bilinear ligand–protein attention, reflecting the broader movement toward models that expose local interaction signals rather than only final scores [2, 6, 9]. The model is also intended to support cold-start generalization, where unseen drugs or targets must be evaluated using transferable graph and protein representations [14, 22].

Table 2 summarizes the key design priorities that shape the proposed multimodal DTI architecture and explains how each priority contributes to practical drug discovery use.

Table 2. Literature-Informed Design Priorities for the Proposed Multimodal DTI Model

Design priority	Why it matters for DTI prediction	How the proposed model addresses it	Added value for drug discovery

Joint ligand–target learning	Binding depends on compatibility between a molecule and a protein, not either modality alone	Uses molecular graph encoding and protein sequence embeddings before cross-modal fusion	Improves prioritization of candidate drug–target pairs
Cold-target generalization	Many biologically important targets have limited known ligand annotations	Uses transferable protein language-model embeddings	Supports screening for under-characterized or orphan targets
Chemical substructure sensitivity	Pharmacophoric patterns often arise from atom neighborhoods and scaffold-level features	Uses graph attention over molecular topology	Helps identify ligand regions relevant to predicted binding
Protein-region sensitivity	Interaction predictions may depend on specific residues or sequence regions	Retains residue-level protein embeddings before pooling or attention	Supports biological interpretation of target-side relevance
Cross-modal interpretability	Users need more than a final prediction score	Produces ligand–protein attention maps	Helps generate medicinal chemistry hypotheses
Uncertainty-aware use	Predictions outside the training domain can be misleading	Adds confidence, calibration, or uncertainty reporting	Prevents overconfident use in experimental prioritization
Deployment readiness	Practical tools must be usable by non-computational researchers	Can be implemented as an API or web-service workflow	Supports scalable virtual screening and repurposing workflows

Data Sources and Input Representations

Curation of Drug–Target Interaction Data

A conceptual training corpus would combine curated drug–target interaction and affinity records from resources such as BindingDB, DAVIS, KIBA, DrugBank, and related benchmark collections. Models such as DeepDTA and GraphDTA used affinity-oriented benchmarks to frame DTI prediction as a supervised learning task, while heterogeneous graph models incorporated broader drug and target network information [1, 6, 10]. The curation process should standardize chemical identifiers, protein accessions, affinity units, assay types, and duplicate records before assigning pairs to random, cold-drug, cold-target, or cold-pair splits. Such splitting is essential because models may appear effective under random evaluation while struggling when either the ligand or target modality is unseen during training [14, 22].

Ligand Graph Construction

Ligand graph construction begins by converting each canonicalized SMILES string into a two-dimensional molecular graph using cheminformatics software. Atom-level features should include element identity, valence-related descriptors, aromaticity, formal charge, chirality, and hybridization, while bond-level features should encode bond order, conjugation, ring membership, and stereochemical status. This representation is aligned with graph-based DTI models such as GraphDTA, MGraphDTA, and GraphscoreDTA, which use molecular topology to learn affinity-relevant ligand embeddings [6, 12, 13]. Although two-dimensional graphs do not explicitly encode conformation, they provide a computationally efficient ligand representation that can be extended later with conformer-aware features.

Protein Sequence Embedding

Protein sequence embedding can be performed by passing target amino-acid sequences through a pre-trained protein transformer such as ProtBERT or a related sequence language model. ProtTrans established the value of self-supervised protein representation learning, while DTI models using protein sequence encoders showed that sequence information can support interaction prediction even when structures are unavailable [4, 7]. For long proteins, the model can use sequence windowing, residue-level pooling, or domain-aware aggregation to produce embeddings compatible with the ligand encoder. These protein embeddings may be frozen to preserve broad biological priors or fine-tuned when sufficient task-specific DTI data are available [11, 16].

Multimodal Deep Learning Architecture

Ligand Encoder – Graph Attention Network

The ligand encoder uses stacked graph attention layers to update atom embeddings by weighting neighboring atoms according to their learned chemical relevance. This design is consistent with graph-based DTI work showing that molecular graph encoders can capture ligand features more directly than sequence-only compound representations [6, 12]. A global attention readout then aggregates atom embeddings into a fixed-size ligand vector while preserving information about substructures that may act as pharmacophores. Because attention scores remain associated with atoms and neighborhoods, the ligand branch can later contribute to interpretable maps of predicted binding determinants [2, 13].

Protein Encoder – Pre-trained Transformer

The protein encoder tokenizes the target sequence and processes it with a pre-trained transformer to obtain contextual residue-level embeddings. Protein language models such as those described by ProtTrans learn sequence representations that can reflect biological constraints beyond simple amino-acid composition [7]. A lightweight pooling or residue-attention layer

converts residue embeddings into a protein-level vector suitable for multimodal fusion, while retaining optional residue-level outputs for interpretability. This approach complements prior DTI models that rely on protein sequence encoders and supports generalization to targets without experimentally resolved structures [4, 15, 11].

Multimodal Fusion and Prediction Heads

The fusion module applies bilinear or cross-attention operations between ligand atom embeddings and protein residue embeddings to construct an interaction-aware joint representation. This strategy follows the motivation of attention-based DTI models such as MolTrans, FusionDTA, CoaDTI, and MCL-DTI, where cross-modal dependencies are modeled more explicitly than in simple concatenation [8, 9, 17, 18]. The resulting ligand–protein interaction map is pooled into a joint embedding and passed to two task-specific heads: one for conceptual binding affinity regression and one for binary interaction classification. This dual-head formulation allows the model to support both affinity-oriented benchmarks and interaction-status prediction without requiring separate architectures [3, 19].

Figure 1 presents the proposed multimodal drug–target interaction architecture, showing how ligand molecular graphs and protein language-model embeddings are fused through bilinear attention to generate affinity predictions, interaction classifications, and interpretable ligand–protein relevance maps.

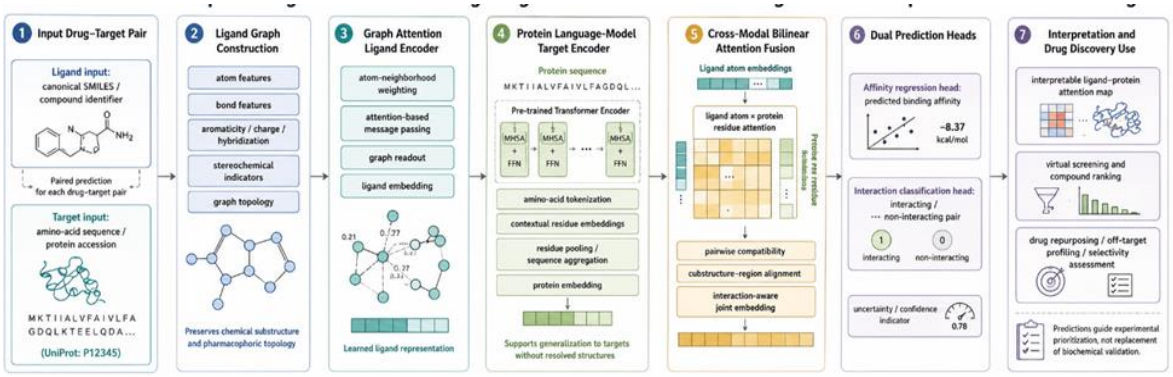

Figure 1. Multimodal Deep Learning Architecture for Drug–Target Interaction Prediction Using Molecular Graphs and Protein Embeddings

Table 3 decomposes the proposed multimodal architecture into its ligand, protein, fusion, prediction, and interpretation components, clarifying how each layer contributes distinct computational and drug discovery value.

Table 3. Architecture-Level Mapping of Multimodal Inputs, Encoders, Fusion Logic, and Drug Discovery Outputs

Model component	Primary information captured	Computational representation	Function within the proposed DTI model	Added interpretive value	Drug discovery relevance
Ligand identity	Chemical structure of the candidate compound	Canonical SMILES, compound identifier, standardized molecular record	Defines the small-molecule entity submitted for prediction	Ensures traceability between predicted interaction and chemical structure	Supports virtual screening, analogue prioritization, and compound library triage
Molecular graph construction	Atom–bond topology and local chemical environment	Nodes as atoms; edges as bonds; atom features including charge, aromaticity, hybridization, valence, chirality	Converts ligand input into a learnable graph suitable for message passing	Preserves chemically meaningful substructures rather than reducing the ligand to global descriptors	Helps identify pharmacophoric regions and scaffold-level patterns
Graph attention ligand encoder	Local and global molecular substructure relevance	Atom embeddings, neighborhood attention weights, graph-level ligand embedding	Learns ligand representations by weighting chemically informative atom neighborhoods	Allows atom- or substructure-level attention to be inspected after prediction	Supports medicinal chemistry reasoning about substituent importance and scaffold modification
Protein sequence input	Target identity and primary biological sequence	Amino-acid sequence, protein accession, tokenized residue sequence	Defines the biological target without requiring experimentally resolved protein structure	Enables target representation even when structural data are missing or incomplete	Expands use to under-characterized targets, orphan proteins, and broad proteome screening
Protein language-model encoder	Contextual residue patterns, evolutionary constraints, and sequence-derived biological regularities	Pre-trained transformer embeddings, residue-level vectors, pooled protein embedding	Generates transferable protein representations for interaction prediction	Allows residue regions or sequence windows to be linked to predicted binding relevance	Supports cold-target generalization and target-family extrapolation

Cross-modal bilinear attention	Pairwise compatibility between ligand substructures and protein residues	Ligand atom × protein residue attention matrix; interaction-aware joint embedding	Fuses ligand and protein modalities beyond simple concatenation	Produces an interpretable ligand–target relevance map	Helps explain why a compound is predicted to interact with a specific protein
Affinity regression head	Continuous binding-strength estimate	Predicted affinity score, ranking value, regression output	Supports quantitative prioritization of drug–target pairs	Separates relative binding strength from binary interaction presence	Useful for ranking compounds in target-based screening
Interaction classification head	Binary interaction status	Interacting / non-interacting probability or class label	Supports categorical DTI prediction when affinity labels are unavailable	Allows model use across heterogeneous datasets with different label types	Useful for drug repurposing, off-target detection, and interaction discovery
Uncertainty and calibration layer	Reliability of predicted score or classification	Confidence value, uncertainty estimate, calibration summary	Flags predictions requiring caution or experimental confirmation	Prevents overinterpretation of uncertain or out-of-distribution predictions	Supports responsible use in experimental prioritization

Training Strategies and Handling Data Imbalance

Cold-Target and Cold-Drug Split Evaluation

Cold-target and cold-drug evaluation should be treated as a central training and validation principle rather than as an optional robustness check. In these settings, the model is assessed on proteins or ligands that were not observed during training, which better reflects prospective use in orphan target discovery and early-stage screening [14]. Graph co-contrastive and heterogeneous graph learning approaches emphasize that transferable representations are needed when drug or target neighborhoods are sparse or missing [14, 22]. Protein language embeddings can strengthen cold-target evaluation because they provide biologically informed sequence representations even when a target has few annotated ligands [7, 11].

Addressing Highly Imbalanced Interaction Data

Drug–target interaction datasets are often imbalanced because experimentally confirmed interactions are much easier to define than true non-interactions. A conceptual training strategy should therefore combine class-weighted objectives, focal-style losses, calibrated negative sampling, and careful construction of inactive or unknown pairs to reduce bias toward majority labels [10]. Multimodal frameworks such as MGNDTI and MFCADTI support the broader idea that multiple feature sources and attention-based mechanisms can help stabilize learning when interaction evidence is sparse or unevenly distributed [23]. For affinity regression, the same concern appears as uneven coverage across assay types, target families, and chemical scaffolds, so loss functions and sampling procedures should prevent overrepresentation of well-studied kinases or drug classes [24].

Transfer and Few-Shot Learning for Orphan Targets

Transfer learning is particularly important for orphan targets, where only limited ligand annotations may be available. A protein transformer initialized from broad sequence pre-training can provide reusable biological priors, allowing the DTI model to adapt from related proteins rather than learning each target from scratch [7]. Contrastive learning in protein language space further suggests that compound and protein representations can be aligned in a way that may support few-shot interaction prediction for targets with sparse labels [11]. Few-shot DTI settings could also benefit from multimodal contrastive representation learning, where ligand graphs and protein embeddings are encouraged to form interaction-aware neighborhoods before supervised fine-tuning [16].

Model Interpretability and Biological Insights

Attention-Weighted Ligand–Protein Interaction Maps

The bilinear attention matrix can be visualized as a ligand–protein interaction heatmap, where rows correspond to ligand atoms or substructures and columns correspond to protein residues or sequence windows. Interpretable affinity models such as DeepAffinity and MONN motivate this design because they connect prediction with local compound–protein relevance rather than treating the model as an opaque scorer [2, 6]. Cross-attention architectures such as CAT-DTI and CoaDTI further support the idea that attention can reveal which parts of each modality participate in the predicted interaction [9, 20]. These maps should be interpreted cautiously, but they can guide inspection of predicted binding regions, especially when compared with known motifs, residue annotations, or available structural evidence.

Translating Attention into Drug Design Hypotheses

Attention-derived ligand substructures can be translated into medicinal chemistry hypotheses by identifying molecular regions that appear to contribute strongly to predicted binding. For example, an atom-level attention pattern could suggest which substituent should be preserved, modified, or avoided when designing analogues, while residue-level attention may indicate whether a prediction is driven by conserved sequence regions or target-specific features [12]. Evidential DTI modeling also highlights the importance of representing uncertainty, because a chemically plausible attention pattern may still require

confidence-aware interpretation before informing design decisions. When combined with counterfactual molecule generation or matched molecular pair analysis, attention maps could support hypothesis generation without replacing experimental validation [21].

Integration Into Drug Discovery Pipelines

Virtual Screening and Drug Repurposing

In a virtual screening workflow, the model could rank candidate compounds against a target of interest using a joint ligand–protein representation instead of relying only on ligand similarity or docking scores. Multimodal DTI models such as MMDTA, AttentionMGT-DTA, and MDNN-DTA illustrate how combining molecular and target information can support broader interaction prediction across drug discovery settings [21, 25]. The same architecture could be reversed for drug repurposing, where an approved or investigational compound is profiled against many proteins to identify plausible new indications or off-target liabilities. Such predictions should be used as prioritization signals for experimental follow-up rather than as definitive evidence of activity.

Deployment as a Web Service or API

Deployment as a web service or application programming interface would allow medicinal chemists and biologists to submit compound structures and protein sequences without directly interacting with model code. A containerized implementation could accept SMILES strings and amino-acid sequences, generate molecular graphs and protein embeddings, and return predicted interaction status, conceptual affinity estimates, uncertainty indicators, and interpretable attention maps [25]. Unified frameworks such as DTIAM suggest the value of serving multiple related prediction tasks, including interaction classification, affinity modeling, and functional mechanism annotation, through a shared representation pipeline. For responsible use, the service should expose model limitations, input validity checks, and documentation describing when predictions are most or least reliable.

Evaluation Strategy

Table 4 provides a deployment-oriented validation framework that separates predictive accuracy, cold-start generalization, interpretability, uncertainty, and practical readiness for multimodal DTI prediction.

Table 4. Validation, Interpretability, and Deployment Readiness Framework for Multimodal DTI Prediction

Evaluation domain	Recommended assessment strategy	What it tests conceptually	Key performance or evidence indicators	Failure mode detected	Practical implication for drug discovery deployment
Random-pair prediction	Train and test on randomly split drug–target pairs	Interpolation among compounds and proteins represented in the training space	High AUROC, AUPRC, concordance index, low RMSE	Inflated performance caused by similarity leakage	Useful as a baseline, but insufficient for prospective claims
Cold-drug evaluation	Hold out ligands unseen during training	Ability to generalize to novel chemical scaffolds	Stable classification and affinity performance on unseen compounds	Memorization of known ligand neighborhoods	Indicates whether the model can support new compound screening
Cold-target evaluation	Hold out proteins unseen during training	Ability to generalize through protein language-model representations	Performance retention across unseen targets and target families	Overfitting to well-studied proteins or kinase-heavy datasets	Critical for orphan target discovery and target expansion
Cold-pair evaluation	Hold out both ligand and target combinations	Prospective utility for unfamiliar compound–protein space	Robust ranking and classification under the hardest split	Dependence on known drug–target pair patterns	Provides the strongest evidence for real-world discovery use
Affinity regression validation	Compare predicted and measured affinity values	Quantitative ranking and binding-strength estimation	Concordance index, RMSE, Pearson/Spearman correlation	Poor calibration of continuous affinity estimates	Determines whether model outputs can guide compound prioritization
Interaction classification validation	Evaluate binary interacting versus non-interacting labels	Categorical DTI recognition under class imbalance	AUROC, AUPRC, sensitivity, specificity, calibrated probability	Majority-class bias or misleading AUROC under imbalance	Determines usefulness for interaction discovery and off-target screening
Attention-map interpretability	Compare ligand–protein attention patterns with known motifs, binding residues, pharmacophores, or structural evidence	Biological plausibility of model explanations	Alignment with annotated residues, conserved motifs, co-crystal evidence, or known SAR	Attention patterns that are unstable or chemically implausible	Supports medicinal chemistry interpretation but should not replace experimental validation

Explanation stability	Test attention consistency across analogues, homologous proteins, and repeated model initializations	Robustness of interpretability outputs	Similar attention patterns for close analogues or related targets	Explanations that change despite similar inputs	Determines whether attention maps are reliable enough for hypothesis generation
Dataset bias and imbalance analysis	Stratify performance by target family, assay type, chemical scaffold, and interaction label density	Whether the model performs equitably across data-rich and data-poor regions	Balanced subgroup performance, reduced scaffold or target-family bias	Dominance of well-studied targets or overrepresented chemical classes	Prevents misleading deployment in narrow or biased discovery spaces
Uncertainty and out-of-distribution testing	Estimate confidence for unusual ligands, rare targets, synthetic proteins, or chemically distant compounds	Reliability of predictions outside the training distribution	Calibrated uncertainty, abstention behavior, error-confidence alignment	Overconfident prediction on unsupported inputs	Enables responsible API or web-service deployment
Experimental prioritization readiness	Translate model outputs into ranked candidates for biochemical follow-up	Whether predictions are actionable for laboratory decision-making	Clear ranking, interpretable rationale, uncertainty flag, input-validity report	Predictions presented as definitive evidence rather than prioritization signals	Positions the model as a decision-support tool for screening, repurposing, and selectivity profiling
Deployment governance	Document model scope, input constraints, update policy, and validation boundaries	Transparency and reproducibility of model use	Model card, dataset documentation, version control, validation report	Unclear limitations or unsupported clinical/pharmaceutical claims	Supports responsible integration into computational drug discovery platforms

Predictive Performance Metrics

The model should be evaluated with task-appropriate metrics rather than a single aggregate score. For binding affinity regression, concordance index, root mean squared error, and Pearson or Spearman correlation are commonly used to assess ranking quality, calibration, and association between predicted and measured affinities [1]. For binary interaction classification, area under the receiver operating characteristic curve and area under the precision–recall curve are appropriate, especially because interaction labels are often imbalanced [10]. Evaluation should be repeated under random, cold-drug, cold-target, and cold-pair splits so that reported behavior reflects both interpolation among known entities and extrapolation to new chemical or protein space [14, 22].

Benchmark Comparisons

Benchmarking should compare the proposed model against traditional docking or descriptor-based baselines, single-modality neural models, and published multimodal DTI architectures. DeepDTA and DeepConv-DTI represent sequence-oriented baselines, while GraphDTA and MGraphDTA provide graph-based ligand comparisons [1, 4, 6, 12]. Multimodal and attention-based comparators should include models such as DeepCDA, MONN, MolTrans, FusionDTA, CoaDTI, MCL-DTI, MGNDTI, and MFCADTI because they represent different strategies for combining molecular and protein information [3, 5, 8, 9, 17, 18, 23]. Recent graph, sequence, and knowledge-integrated models such as SaeGraphDTI and multimodal affinity frameworks should also be included to test whether the proposed design remains competitive against contemporary architectures.

Interpretability Assessment

Interpretability should be evaluated separately from predictive accuracy because a model may predict well while producing unstable or biologically uninformative explanations. Qualitative assessment can compare ligand–protein attention maps with known binding-site residues, conserved motifs, pharmacophoric substructures, or co-crystal evidence when such information is available [2]. Quantitative assessment could examine whether attention patterns remain consistent across close analogues, homologous proteins, and repeated model initializations, following the interpretability motivation of attention-based and graph-based affinity models [6, 12]. Uncertainty-aware approaches such as evidential DTI prediction can further help distinguish confident explanations from attention maps produced for uncertain or out-of-distribution pairs.

Limitations

Dependence on Pre-trained Protein Models

The proposed model depends partly on the representational quality of the pre-trained protein encoder. Protein language models trained on broad sequence corpora can encode useful biological patterns, but unusual targets, engineered proteins, synthetic constructs, noncanonical residues, and poorly represented protein families may not be captured with equal fidelity [7]. Fine-tuning may improve task alignment, yet excessive fine-tuning on narrow DTI datasets could weaken generalization by overfitting to well-studied target families [11]. The model should therefore include uncertainty estimation, domain checks, and evaluation on cold-target splits before being applied to under-characterized proteins.

Structural Limitations of 2D Graph Representations

A two-dimensional molecular graph captures connectivity, atom identity, and bond structure, but it does not fully represent conformational ensembles, solvent effects, induced fit, metal coordination, or water-mediated contacts. These limitations are important because binding affinity and selectivity often depend on three-dimensional complementarity between ligand and protein. Structure-aware extensions, such as protein contact maps, molecular conformers, or protein–ligand geometric encoders, could address part of this gap at additional computational cost [26]. The proposed architecture should therefore be viewed as a scalable multimodal screening model rather than a replacement for structure-based modeling and experimental validation.

Conclusion

The proposed manuscript describes a multimodal deep learning model for drug–target interaction prediction using molecular graphs and protein embeddings. The model combines a graph attention network for ligand representation, a pre-trained protein transformer for target representation, and a bilinear attention module for interaction-aware fusion.

Its main strength is joint representation learning across chemical and biological modalities. By pairing molecular graph learning with protein language embeddings, the architecture could generalize beyond memorized drug–target pairs and provide interpretable attention signals for virtual screening, selectivity profiling, and drug repurposing.

Important challenges remain. The model depends on the quality and domain coverage of pre-trained protein encoders, uses ligand graphs that may omit three-dimensional binding effects, and would still require rigorous prospective validation before supporting experimental decisions.

Future work should prioritize transparent implementation, careful benchmarking, uncertainty-aware reporting, and open-source release. Integration into collaborative drug discovery platforms could help make interpretable multimodal DTI prediction more accessible for target-based drug design.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug–target binding affinity prediction. Bioinformatics. 2018;34(17):i821-9.
2. Karimi M, Wu D, Wang Z, Shen Y. DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. Bioinformatics. 2019;35(18):3329-38.
3. Li S, Wan F, Shu H, Jiang T, Zhao D, Zeng J. MONN: a multi-objective neural network for predicting compound-protein interactions and affinities. Cell Syst. 2020;10(4):308-22.
4. Lee I, Keum J, Nam H. DeepConv-DTI: prediction of drug-target interactions via deep learning with convolution on protein sequences. PLoS Comput Biol. 2019;15(6):e1007129.
5. Abbasi K, Razzaghi P, Poso A, Amanlou M, Ghasemi JB, Masoudi-Nejad A. DeepCDA: deep cross-domain compound–protein affinity prediction through LSTM and convolutional neural networks. Bioinformatics. 2020;36(17):4633-42.
6. Nguyen T, Le H, Quinn TP, Nguyen T, Le TD, Venkatesh S. GraphDTA: predicting drug–target binding affinity with graph neural networks. Bioinformatics. 2021;37(8):1140-7.
7. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell. 2021;44(10):7112-27.
8. Huang K, Xiao C, Glass LM, Sun J. MolTrans: molecular interaction transformer for drug–target interaction prediction. Bioinformatics. 2021;37(6):830-6.
9. Huang L, Lin J, Liu R, Zheng Z, Meng L, Chen X, et al. CoaDTI: multi-modal co-attention based framework for drug–target interaction annotation. Brief Bioinform. 2022;23(6):bbac446.
10. Peng J, Wang Y, Guan J, Li J, Han R, Hao J, et al. An end-to-end heterogeneous graph representation learning-based framework for drug–target interaction prediction. Brief Bioinform. 2021;22(5):bbaa430.
11. Singh R, Sledzieski S, Bryson B, Cowen L, Berger B. Contrastive learning in protein language space predicts interactions between drugs and protein targets. Proc Natl Acad Sci U S A. 2023;120(24):e2220778120.
12. Yang Z, Zhong W, Zhao L, Chen CY. MGraphDTA: deep multiscale graph neural network for explainable drug–target binding affinity prediction. Chem Sci. 2022;13(3):816-33.
13. Wang K, Zhou R, Tang J, Li M. GraphscoreDTA: optimized graph neural network for protein–ligand binding affinity prediction. Bioinformatics. 2023;39(6):btad340.

14. Li Y, Qiao G, Gao X, Wang G. Supervised graph co-contrastive learning for drug–target interaction prediction. Bioinformatics. 2022;38(10):2847-54.
15. Deng L, Zeng Y, Liu H, Liu Z, Liu X. DeepMHADTA: prediction of drug-target binding affinity using multi-head self-attention and convolutional neural network. Curr Issues Mol Biol. 2022;44(5):2287-99.
16. Zhang L, Ouyang C, Liu Y, Liao Y, Gao Z. Multimodal contrastive representation learning for drug-target binding affinity prediction. Methods. 2023;220:126-33.
17. Yuan W, Chen G, Chen CY. FusionDTA: attention-based feature polymerizer and knowledge distillation for drug-target binding affinity prediction. Brief Bioinform. 2022;23(1):bbab506.
18. Qian Y, Li X, Wu J, Zhang Q. MCL-DTI: using drug multimodal information and bi-directional cross-attention learning method for predicting drug–target interaction. BMC Bioinform. 2023;24(1):323.
19. Liu S, Wang Y, Deng Y, He L, Shao B, Yin J, et al. Improved drug–target interaction prediction with intermolecular graph transformer. Brief Bioinform. 2022;23(5):bbac162.
20. Zeng X, Chen W, Lei B. CAT-DTI: cross-attention and Transformer network with domain adaptation for drug-target interaction prediction. BMC Bioinform. 2024;25(1):141.
21. Zhong KY, Wen ML, Meng FF, Li X, Jiang B, Zeng X, et al. MMDTA: a multimodal deep model for drug-target affinity with a hybrid fusion strategy. J Chem Inf Model. 2023;64(7):2878-88.
22. Li M, Cai X, Xu S, Ji H. Metapath-aggregated heterogeneous graph neural network for drug–target interaction prediction. Brief Bioinform. 2023;24(1):bbac578.
23. Peng L, Liu X, Chen M, Liao W, Mao J, Zhou L. MGNDTI: a drug-target interaction prediction framework based on multimodal representation learning and the gating mechanism. J Chem Inf Model. 2024;64(16):6684-98.
24. Kalemati M, Zamani Emani M, Koohi S. BiComp-DTA: drug-target binding affinity prediction through complementary biological-related and compression-based featurization approach. PLoS Comput Biol. 2023;19(3):e1011036.
25. Wu H, Liu J, Jiang T, Zou Q, Qi S, Cui Z, et al. AttentionMGT-DTA: a multi-modal drug-target affinity prediction using graph transformer and attention mechanism. Neural Netw. 2024;169:623-36.
26. Jiang M, Li Z, Zhang S, Wang S, Wang X, Yuan Q, et al. Drug–target affinity prediction using graph neural network and contact maps. RSC Adv. 2020;10(35):20701-12.