scientific reports

Check for updates

OPEN

MSCMF-DTB: a multi-scale cross-modal fusion framework for drug–target binding prediction

Juan Huang[1,2], Yuxue Pan[1,2] & Qu Chen[1]✉

Predicting drug–target binding remains a central challenge in computational drug discovery, particularly due to the need for models that jointly capture molecular topology, chemical substructures, and protein sequence dependencies. We propose MSCMF-DTB, an end-to-end deep learning framework supporting both drug–target interaction (DTI) classification and drug–target affinity (DTA) regression. On the drug side, molecular graphs generated with RDKit are encoded using a DenseGCN module, while a parallel fingerprint channel captures fragment-level and compositional features. On the protein side, contextualized embeddings from TAPE-BERT are processed through a multi-scale 1D CNN to extract local sequence patterns. Cross-modal drug–protein relationships are modeled using cross-attention mechanism coupled with a tensor network for higher-order feature interaction. The fused representations are fed into an MLP for final prediction. Extensive experiments demonstrate that MSCMF-DTB achieves competitive and consistent performance across small- and large-scale datasets (Human, *C. elegans*, GPCR, BioSNAP, and DrugBank for DTI, and DAVIS and KIBA for DTA). Notably, on the large-scale DrugBank dataset for DTI prediction, MSCMF-DTB improved AUC and Recall by up to 3.2% and 6.1%, respectively, compared with the second-best model (DrugBAN). For DTA prediction, the model achieved stable performance on the large and heterogeneous KIBA dataset, with an MSE of 0.146, a Concordance Index of 0.886, and an r_m^2 of 0.765. Attention-based interpretability further shows that the model learns biologically meaningful interaction regions. Finally, a cold-start case study indicates that MSCMF-DTB successfully identifies experimentally validated inhibitors to AKT1, illustrating its practical utility in virtual screening and drug repurposing.

Keywords Drug–Target Interaction, Drug–Target Affinity, Cross-Attention Mechanism, Cross-Modal Fusion, Tensor Networks, Molecular Fingerprints

Ever since the discovery of the double-helical structure of DNA, modern biology has approached biological processes through a molecular-level perspective[1]. As experimental techniques become increasingly costly, labor-intensive, and sophisticated, in silico methods have gradually emerged as indispensable tools for providing mechanistic insights and complementing laboratory studies[2]. Broadly speaking, in silico methods can be categorized into physics-based and data-driven approaches, as well as their hybrid variants. Physics-based approaches, such as molecular dynamics (MD), can elucidate molecular mechanisms underlying biological phenomena that are difficult to observe experimentally[3–5]. Meanwhile, the advent of high-throughput technologies including next-generation sequencing and cryo-electron microscopy has generated vast multi-omics datasets including genomics and proteomics. Utilizing these data, machine learning (ML) has become a powerful data-driven paradigm capable of uncovering complex, non-linear relationships within biological systems. Recent advances in ML, particularly in deep learning (DL) and generative modeling, have enabled the rapid screening of potential compound candidates and the *de novo* design of bioactive molecules[6,7]. Drug discovery is a typical example that benefits from these developments, as ML models can efficiently predict drug–target bindings (DTBs), a crucial step that reveals how small molecules interact with biological macromolecules such as enzymes, receptors, and ion channels[8]. The ability to accurately predict DTBs can substantially accelerate drug development, reduce experimental costs, and improve success rates in pharmaceutical research pipeline[9].

In recent years, a wide range of ML models, from classical ML to DL, have been developed to improve DTB predictions, as summarized in some excellent surveys and reviews[7,8,10]. These predictive models are generally

[1]School of Biological and Chemical Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, People's Republic of China. [2]These authors contributed equally: Juan Huang and Yuxue Pan. ✉email: chenqu@zust.edu.cn

categorized into two groups based on their tasks: classification models predicting whether a drug binds to its target protein (drug–target interaction, DTI), and regression models predicting how strong a drug binds to its target protein (drug–target affinity, DTA). The performance of such models largely depends on two key factors: the learning algorithm and the quality of the data. Traditional ML models based on feature engineering, such as support vector machines (SVMs) and random forests (RFs), rely on handcrafted molecular descriptors and fingerprints. In contrast, DL models represent a new generation of algorithms capable of automatically learning hierarchical representations directly from raw biological and chemical data. Proteins are typically represented by amino acid sequences using one-hot coding, while drugs are often represented by SMILES strings. These representations are trained by various neural architectures, such as graph neural networks (GNNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), with significantly enhanced predictive accuracy and generalization in DTB prediction. Among these, DeepDTI[11] and DeepDTA[12] are considered pioneering DL frameworks for DTI and DTA prediction, respectively. Based on these foundational models, subsequent studies have proposed various improvements to further enhance prediction accuracy. For instance, implementing 2D topological representations of drug molecules (where atoms are represented as graph nodes and bonds as edges) has proven effective, as reflected by the application of RDkit in the GraphDTA model[13]. The incorporation of attention mechanisms in models such as HyperAttentionDTI[14] and IIFDTI[15] has promoted DTB prediction by adaptively weighting different regions of the input, allowing models to focus on the most informative molecular and sequence features, and hence mitigating the difficulty of capturing intricate interaction patterns between drugs and target proteins. More recently, the development of DTB prediction models has increasingly concentrated on multi-model integration[16,17] while improving model interpretability[18,19] and introducing uncertainty analysis[20,21], marking a transition towards more robust and transparent predictive frameworks for data-driven drug discovery.

A powerful algorithm is of limited value without high-quality datasets derived from well-established databases. In the field of DTB prediction, several benchmark datasets are widely used in the literature, including Human, *C. elegans*, DrugBank, BioSNAP, KIBA, DAVIS, BindingDB, DTINet, DUD-E, Yamanishi, and PDBbind. Among them, the Yamanishi datasets, first introduced by Yamanishi et al. in 2008, are one of the earliest benchmarks for DTI prediction. They contain only experimentally validated positive interactions sourced from KEGG BRITE[22], BRENDA[23], SuperTarget[24], and DrugBank[25], and are organized into four domain-specific networks: enzymes, ion channels (ICs), G-protein–coupled Receptors (GPCRs), and nuclear receptors (NCs). Later, the Human and *C. elegans* datasets, two balanced DTI benchmarks containing equal numbers of positive and negative samples, were curated by Liu et al. and are specific to human proteins and the model organism *Caenorhabditis elegans*, respectively[26]. Positive interactions in these datasets were generated from DrugBank[25], Matador[24], and STITCH[27] databases, while highly reliable negative interactions, which are typically unavailable experimentally, were built upon the similarity-based assumption that compounds with similar chemical structures are likely to interact with similar protein targets. The DrugBank dataset is also a balanced DTI benchmark constructed by Zhao et al.[14], who extracted positive drug–target interactions from DrugBank 5.1.5[28]. During preprocessing, drugs that were inorganic compounds, extremely small molecules, and drugs with SMILES strings incompatible with RDKit were removed. Negative interactions were then generated by sampling from the set of unlabeled drug–protein pairs, resulting in a balanced dataset with equal numbers of positive and negative samples. BioSNAP, released by the Stanford Network Analysis Platform (SNAP) group, is another balanced dataset sourced from DrugBank and aimed for DTI binary classification[29]. DTINet is a more complex DTI benchmark constructed as a heterogeneous network integrating multiple biological data sources, including drug–protein associations from DrugBank[28], protein–protein interactions from HPRD[30], disease–protein associations from the Comparative Toxicogenomics Database[31], and drug–side-effect links from SIDER[32]. Finally, the DUD-E (Directory of Useful Decoys, Enhanced) dataset, originally designed to evaluate structure-based virtual screening, consists of 102 protein targets, each associated with an average of 224 experimentally confirmed active ligands and 50 physicochemically matched but topologically dissimilar decoys[33].

In addition to these datasets for DTI, several datasets including DAVIS, KIBA, and BindingDB are benchmarks with both DTI and DTA labels. The DAVIS dataset, introduced by Davis et al. in 2011, is a kinase-focused benchmark that compiles interaction data from selectivity assays conducted on a panel of human kinases and their small-molecule inhibitors[34]. For DTA prediction tasks, each drug–target pair is labeled with experimentally measured dissociation constants (K_d). Similarly, the KIBA dataset, proposed by Tang et al. in 2014, is also a kinase-focused benchmark that integrates drug–target interaction data from multiple bioactivity resources, including K_i, K_d, and IC_{50} measurements[35]. To make these heterogeneous affinity metrics comparable, the authors introduced the "KIBA score", a unified affinity measure derived through a statistical integration scheme. BindingDB further expands the landscape as a large repository of experimentally measured protein-ligand affinities, such as K_i, K_d, IC_{50} and EC_{50} curated from databases including ChEMBL[36], Protein Data Bank (PDB)[37], PubChem[38], and UniProt[39]. Because the dataset is extensive and heterogeneous, researchers rarely use it directly. Instead, subsets of BindingDB are typically extracted, filtered, and standardized to construct task-specific DTI or DTA benchmark datasets. Finally, PDBbind[40] is a widely used benchmark only for DTA prediction, providing experimentally measured binding affinities together with the 3D structures of protein–ligand complexes collected from the PDB. The dataset is organized into three subsets: the general set, containing all curated complexes; the refined set, filtered for higher structural and experimental quality; and the core set, a small, high-confidence benchmark subset commonly used for model evaluation. Testing models across such diverse datasets helps ensure robustness and generalizability.

Despite the above-mentioned advances in algorithms and datasets, several limitations remain in current DTB prediction research. Some models still rely on single-scale representations or limited modalities, making it challenging to jointly capture molecular topology, chemical substructures, and protein sequence dependencies within a unified framework. Moreover, many existing methods are evaluated exclusively on either DTI or DTA,

which may leave their generalization across task types insufficiently explored. To address these gaps, we propose MSCMF-DTB, an end-to-end multi-scale cross-modal fusion model designed to integrate heterogeneous drug and protein information more comprehensively. The multi-scale design is motivated by the fact that drug and protein molecules exhibit meaningful patterns from local sequence and chemical fragments to global topological structures. Meanwhile, cross-modal interaction modeling is essential because drug graphs, molecular fingerprints, and protein embeddings capture distinct yet complementary aspects of binding behavior. By combining multi-scale feature extraction with cross-modal fusion, MSCMF-DTB achieves a more biologically coherent representation of drug–target relationships. Indeed, similar cross-modal and integrative representation learning strategies have already demonstrated effectiveness in related drug discovery problems[41–43]. We extensively evaluated MSCMF-DTB on multiple DTI datasets, including Human, *C. elegans*, BioSNAP, GPCR, and DrugBank, as well as DTA datasets such as DAVIS and KIBA. Notably, to ensure transparency and comparability with state-of-the-art (SOTA) methods, we reran competing models under unified settings and re-benchmarked all DTI metrics.

This paper is organized as follows. Sections "Performance on DTI tasks" and "Performance on DTA tasks" present performance comparisons of our models against SOTA methods across multiple benchmark datasets for both DTI and DTA tasks, respectively. Section "Ablation Experiments" reports the results of ablation studies evaluating the contributions of some key model components. Section "Case Studies" provides two case studies illustrating the interpretability of the attention mechanism and the model's application in screening AKT1 allosteric inhibitors. Section "Materials and Methods" describes the experimental environment and hyperparameter settings, the statistics of the datasets used, all baseline SOTA models, evaluation metrics, and the components of the proposed framework. Finally, our study is concluded in Section "Conclusion". The code for our model is provided in the "Data availability" statement.

Results and Discussion

The workflow of our proposed model MSCMF-DTB is outlined in Fig. 1. The model takes drug SMILES and protein sequences as raw inputs, converting them into feature vectors using RDKit and TAPE-BERT, respectively, while also constructing a fingerprint-like embedding channel derived from atomic symbols. On the drug side, a DenseGCN module extracts multi-level molecular topological features, complemented by a parallel fingerprint branch that encodes statistical substructure patterns. On the protein side, contextualized TAPE-BERT embeddings are further refined by a multi-scale CNN with varying kernel sizes to capture local and long-range sequence patterns. At the interaction stage, a Cross Attention mechanism, coupled with a Tensor Network, directly models interactions between drug and protein features and models higher-order nonlinear interactions. The outputs from the molecular graph, fingerprint, protein sequence, and interaction branches are then concatenated and fed into a multilayer perceptron (MLP), followed by a task-specific output layer for final prediction. For a detailed description of each module, please refer to the "Materials and Methods" section.

Currently, the benchmarking of newly developed methods often varies widely in both hyperparameters and datasets, making it difficult to ensure fair comparison and to accurately assess real algorithmic progress. To address this issue for the DTI task, we standardized the evaluation protocol by unifying all hyperparameters across experiments (see "Materials and Methods" for details). Competing models were selected based on their use of diverse computational modules to process drug and protein information, ensuring a comprehensive assessment of strategies for representing molecular features and modeling drug–protein interactions. However, for the DTA tasks, the baseline results were taken directly from previously published studies that followed the

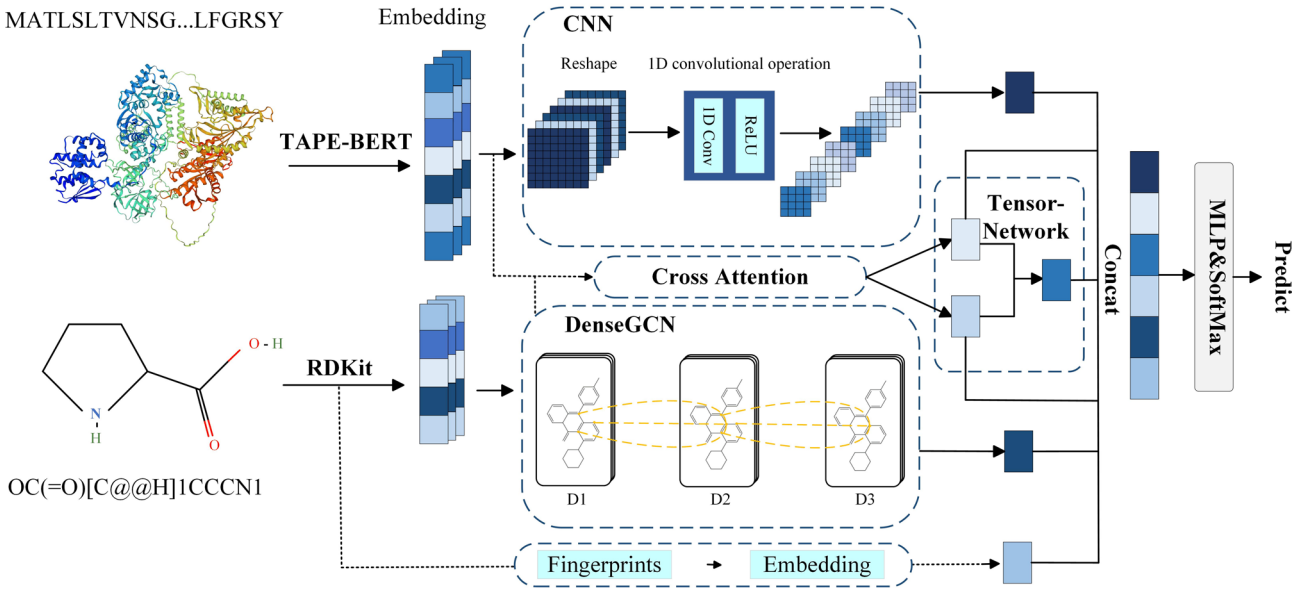

Fig. 1. Workflow of our proposed DL framework MSCMF-DTB for DTI and DTA predictions.

Model	AUC	AUPR	Accuracy	Precision	Recall	F1
MolTrans	0.981 (0.002)	0.983 (0.005)	0.936 (0.015)	0.932 (0.012)	0.946 (0.016)	0.939 (0.015)
Mutual-DTI	0.983 (0.001)	0.987 (0.001)	0.936 (0.002)	0.943 (0.013)	0.938 (0.013)	0.940 (0.001)
AMMVF-DTI	0.984 (0.004)	0.983 (0.004)	0.937 (0.009)	*0.948 (0.021)*	*0.947 (0.023)*	**0.948 (0.008)**
FMCA-DTI	**0.991 (0.001)**	**0.991 (0.001)**	**0.949 (0.002)**	**0.964 (0.006)**	0.929 (0.006)	*0.946 (0.010)*
DrugBAN	0.981 (0.005)	0.975 (0.008)	*0.944 (0.004)*	0.921 (0.011)	0.934 (0.010)	0.927 (0.006)
MSCMF-DTB	*0.985 (0.002)*	*0.988 (0.003)*	*0.944 (0.003)*	0.939 (0.007)	**0.958 (0.014)**	**0.948 (0.005)**

Table 1. Performance comparison of MSCMF-DTB and competing models on the Human dataset. Values in parentheses indicate standard deviations. Best results are highlighted in bold, and second-best results are underlined.

Model	AUC	AUPR	Accuracy	Precision	Recall	F1
MolTrans	0.986 (0.003)	0.980 (0.010)	0.963 (0.004)	**0.978 (0.016)**	0.938 (0.012)	0.958 (0.013)
Mutual-DTI	0.986 (0.002)	0.989 (0.001)	0.943 (0.011)	0.966 (0.015)	0.935 (0.028)	0.950 (0.011)
AMMVF-DTI	0.990 (0.002)	0.991 (0.002)	0.952 (0.012)	0.973 (0.023)	0.948 (0.010)	*0.960 (0.011)*
FMCA-DTI	**0.995 (0.001)**	**0.995 (0.001)**	*0.967 (0.003)*	*0.975 (0.009)*	*0.958 (0.008)*	**0.966 (0.004)**
DrugBAN	0.989 (0.002)	*0.991 (0.001)*	**0.968 (0.003)**	0.959 (0.009)	**0.960 (0.008)**	0.959 (0.005)
MSCMF-DTB	*0.991 (0.001)*	0.990 (0.004)	0.952 (0.007)	0.950 (0.004)	0.957 (0.012)	0.954 (0.007)

Table 2. Performance comparison of MSCMF-DTB and competing models on the *C.elegans* dataset. Values in parentheses indicate standard deviations. Best results are highlighted in bold, and second-best results are underlined.

widely adopted benchmark datasets and evaluation protocols. We report these results as originally published rather than re-tuning their hyperparameters, in order to preserve consistency with the existing literature. For the DTI task, MolTrans (2021)[44], Mutual-DTI (2023)[45], AMMVF-DTI (2023)[46], FMCA-DTI (2024)[47], and DrugBAN (2023)[48] were included, while for the DTA task, DeepGS (2020)[49], DeepCDA (2020)[50], DeepFusionDTA (2022)[51], MATT_DTI (2021)[12], and AttentionMGT-DTA (2024)[52] were selected.

For DTI evaluation, we selected five most popular benchmark datasets: Human, *C. elegans*, GPCR, BioSNAP, and DrugBank. Notably, GPCR is a subset of the original Yamanishi datasets. We focused on these datasets because they provide well-curated drug–target interaction pairs and sufficient coverage of both drugs and protein targets. For DTA evaluation, we employed the DAVIS and KIBA datasets, which are widely adopted in the literature and provide high-quality binding affinity measurements suitable for regression-based tasks. DTI metrics for all competing models were re-evaluated under standardized settings, whereas DTA metrics were obtained from the literature. Results for the DTI task are presented in Tables 1, 2, 3, 4 and 5, and those for the DTA task are shown in Tables 6 and 7.

Performance on DTI tasks

For the DTI task, we conducted classic five-fold cross-validation on the above-mentioned benchmark datasets to ensure the reliability and generalizability of our model. As shown in Tables 1 and 2, across the two small datasets, Human and *C. elegans*, MSCMF-DTB achieved competitive overall performance. On Human, it attained an area under the receiver operating characteristic curve (AUC) of 0.985 and an area under the precision–recall curve (AUPR) of 0.988, slightly below FMCA-DTI's AUC (0.991) and AUPR (0.991), but it outperformed all models in Recall (0.958) and F1-score (0.948), demonstrating improved coverage of true positive interactions. Its Accuracy (0.944) and Precision (0.939) were slightly lower than those performed by FMCA-DTI, reflecting a modest trade-off between overall correctness and completeness. On *C. elegans*, MSCMF-DTB reached an AUC of 0.991 and AUPR of 0.990, with Recall of 0.957 and F1-score of 0.954, maintaining a balanced performance close to the best models, while Accuracy (0.952) and Precision (0.950) remained competitive. FMCA-DTI often achieved the highest AUC and Accuracy but showed lower Recall, indicating some true positives were missed. Other models, such as AMMVF-DTI and Mutual-DTI, maintained high AUC and AUPR scores but showed trade-offs between Precision and Recall. Overall, MSCMF-DTB demonstrates balanced predictive capability, particularly adept at identifying true interactions, which is critical for reliable DTI prediction in small-sample datasets.

Model	AUC	AUPR	Accuracy	Precision	Recall	F1
MolTrans	0.831 (0.014)	0.811 (0.023)	0.724 (0.012)	0.672 (0.009)	0.749 (0.014)	0.708 (0.008)
Mutual-DTI	0.838 (0.009)	0.855 (0.009)	0.755 (0.007)	0.751 (0.010)	*0.796 (0.046)*	*0.780 (0.010)*
AMMVF-DTI	0.862 (0.002)	0.857 (0.002)	0.779 (0.004)	0.762 (0.006)	0.782 (0.009)	0.773 (0.003)
FMCA-DTI	0.782 (0.005)	0.777 (0.009)	0.717 (0.007)	0.734 (0.013)	0.703 (0.013)	0.718 (0.012)
DrugBAN	*0.869 (0.002)*	*0.866 (0.004)*	*0.783 (0.005)*	**0.792 (0.013)**	0.758 (0.035)	0.774 (0.012)
MSCMF-DTB	**0.876 (0.003)**	**0.873 (0.006)**	**0.790 (0.003)**	*0.786 (0.020)*	**0.800 (0.039)**	**0.792 (0.010)**

Table 3. Performance comparison of MSCMF-DTB and competing models on the GPCR dataset. Values in parentheses indicate standard deviations. Best results are highlighted in bold, and second-best results are underlined.

Model	AUC	AUPR	Accuracy	Precision	Recall	F1
MolTrans	0.885 (0.004)	0.892 (0.006)	0.814 (0.022)	0.813 (0.025)	0.816 (0.016)	0.814 (0.016)
Mutual-DTI	0.869 (0.003)	0.873 (0.004)	0.794 (0.004)	0.793 (0.022)	0.786 (0.030)	0.789 (0.004)
AMMVF-DTI	0.858 (0.003)	0.869 (0.004)	0.779 (0.005)	0.793 (0.026)	0.760 (0.037)	0.775 (0.007)
FMCA-DTI	0.855 (0.003)	0.869 (0.002)	0.787 (0.003)	0.822 (0.009)	0.730 (0.005)	0.773 (0.005)
DrugBAN	*0.905 (0.003)*	*0.908 (0.004)*	*0.838 (0.006)*	*0.831 (0.012)*	*0.843 (0.011)*	*0.837 (0.009)*
MSCMF-DTB	**0.921 (0.006)**	**0.923 (0.003)**	**0.846 (0.011)**	**0.835 (0.025)**	**0.865 (0.025)**	**0.849 (0.009)**

Table 4. Performance comparison of MSCMF-DTB and competing models on the BioSNAP dataset. Values in parentheses indicate standard deviations. Best results are highlighted in bold, and second-best results are underlined.

Model	AUC	AUPR	Accuracy	Precision	Recall	F1
MolTrans	0.860 (0.006)	0.856 (0.014)	0.765 (0.023)	0.720 (0.006)	0.767 (0.015)	0.743 (0.004)
Mutual-DTI	0.849 (0.003)	0.857 (0.003)	0.763 (0.006)	0.772 (0.015)	0.750 (0.029)	0.759 (0.009)
AMMVF-DTI	0.822 (0.004)	0.831 (0.003)	0.744 (0.009)	0.734 (0.018)	0.758 (0.012)	0.746 (0.005)
CrossAtt-DTI	0.822 (0.004)	0.831 (0.003)	0.744 (0.009)	0.734 (0.018)	0.758 (0.012)	0.746 (0.005)
FMCA-DTI	0.821 (0.002)	0.836 (0.011)	0.757 (0.004)	0.795 (0.028)	0.691 (0.036)	0.738 (0.007)
DrugBAN	*0.883 (0.003)*	*0.882 (0.002)*	*0.812 (0.009)*	*0.808 (0.012)*	*0.806 (0.017)*	*0.813 (0.010)*
MSCMF-DTB	**0.915 (0.004)**	**0.916 (0.003)**	**0.837 (0.006)**	**0.818 (0.017)**	**0.867 (0.016)**	**0.841 (0.004)**

Table 5. Performance comparison of MSCMF-DTB and competing models on the DrugBank dataset. Values in parentheses indicate standard deviations. Best results are highlighted in bold, and second-best results are underlined.

On the structurally complex GPCR dataset, MSCMF-DTB demonstrated strong predictive performance across multiple evaluation metrics. As shown in Table 3, it achieved the highest AUC (0.876), surpassing CrossAtt-DTI (0.862), Mutual-DTI (0.838), MolTrans (0.831), and DrugBAN (0.869). The model also attained the best Recall (0.800) among all competing models, while its Precision (0.786) was close to the top value of 0.792 observed for DrugBAN. The combination of top scores across AUC, AUPR (0.873), Accuracy (0.790), Recall, and F1 (0.792), along with competitive Precision, indicates that MSCMF-DTB effectively captures the

Model	MSE	CI	r_m^2
DeepGS	0.252	0.882	0.686
DeepCDA	0.248	*0.891*	0.649
DeepFusionDTA	0.253	0.887	n/a
MATT_DTI	0.227	*0.891*	0.683
AttentionMGT-DTA	**0.193**	*0.891*	*0.699*
MSCMF-DTB	*0.203*	**0.898**	**0.715**

Table 6. Performance comparison of MSCMF-DTB and competing models on the DAVIS dataset. Best results are highlighted in bold, and second-best results are underlined.

Model	MSE	CI	r_m^2
DeepGS	0.193	0.860	0.684
DeepCDA	0.176	*0.889*	0.682
DeepFusionDTA	0.176	0.876	n/a
MATT_DTI	0.150	*0.889*	0.756
AttentionMGT-DTA	**0.140**	**0.893**	**0.786**
MSCMF-DTB	*0.146*	0.886	*0.765*

Table 7. Performance comparison of MSCMF-DTB and competing models on the KIBA dataset. Best results are highlighted in bold, and second-best results are underlined.

complex structural features of GPCR targets and maintains robust discrimination capability even in a challenging prediction scenario.

On the large-scale BioSNAP and DrugBank datasets, MSCMF-DTB achieved stable and significant improvements over the second-best models across all evaluation metrics. As shown in Table 4, on BioSNAP, MSCMF-DTB attained an AUC of 0.921, AUPR of 0.923, Accuracy of 0.846, Precision of 0.835, Recall of 0.865, and F1-score of 0.849, representing improvements of 1.6%, 1.5%, 0.8%, 0.4%, 2.2%, and 1.2%, respectively, compared with the second-best model, DrugBAN. On the DrugBank dataset (see Table 5), the improvements of AUC, AUPR, Accuracy, Precision, Recall and F1-score reached 3.2%, 3.4%, 2.5%, 1.0%, 6.1%, and 2.8%, respectively, with a particularly notable increase in Recall to 0.867, highlighting MSCMF-DTB's enhanced ability to detect potential DTIs in large, heterogeneous datasets. These results indicate that the model not only generalizes well to diverse molecular spaces but also demonstrates improved sensitivity in identifying true interactions in large-scale datasets.

In summary, MSCMF-DTB demonstrates competitive or strong performance across datasets of varying size and complexity. It achieves high coverage of true positive interactions in small-sample scenarios while maintaining leading performance across multiple metrics on complex targets and large-scale datasets. Combined with the low standard deviations observed under five-fold cross-validation, these results highlight the model's stability, strong generalization ability, and practical applicability in real-world DTI prediction.

Performance on DTA tasks

We further performed standard five-fold cross-validation on the DAVIS and KIBA datasets for DTA tasks, with the predicted results summarized in Tables 6 and 7. Unlike the DTI datasets, the performance metrics for DTA were obtained directly from previously reported results[52,53]. Overall, MSCMF-DTB consistently achieved competitive performance across both the DAVIS and KIBA kinase-specific datasets. Specifically, on the DAVIS dataset (see Table 6), MSCMF-DTB achieved a Concordance Index (CI) of 0.898 and a r_m^2 of 0.715, both the highest among the compared models, with a mean squared error (MSE) of 0.203, slightly higher than the best reported result of 0.193 from AttentionMGT-DTA. These results indicate that MSCMF-DTB effectively maintains the correct ranking of DTAs while controlling prediction errors, which is particularly important for kinase-targeted drug discovery where correct prioritization of candidates is critical.

On the larger and more heterogeneous KIBA dataset (see Table 7), MSCMF-DTB achieved an MSE of 0.146, a CI of 0.886, and a r_m^2 of 0.765, placing it close to the best-performing model AttentionMGT-DTA with MSE (0.140), CI (0.893), and r_m^2 (0.786). The relatively smaller performance gap between KIBA and DAVIS reflects the increased complexity and heterogeneity of the kinase–drug interactions in this dataset, which pose challenges such as dense drug–target pairings and diverse chemical and structural distributions that can lead to oversmoothing in graph-based representations. Nevertheless, MSCMF-DTB maintained robust correlation and error control, highlighting its generalizability across large-scale and complex kinase-specific datasets.

To further illustrate regression behavior, Fig. 2 presents scatter plots of true versus predicted binding affinities for both datasets. In these plots, the black dashed line denotes the ideal fit ($y = x$), while the colored dashed line represents a first-order linear regression fitted to the predicted–true pairs, whose slope and intercept reflect potential systematic prediction bias. For the DAVIS dataset, most points are distributed around the ideal line,

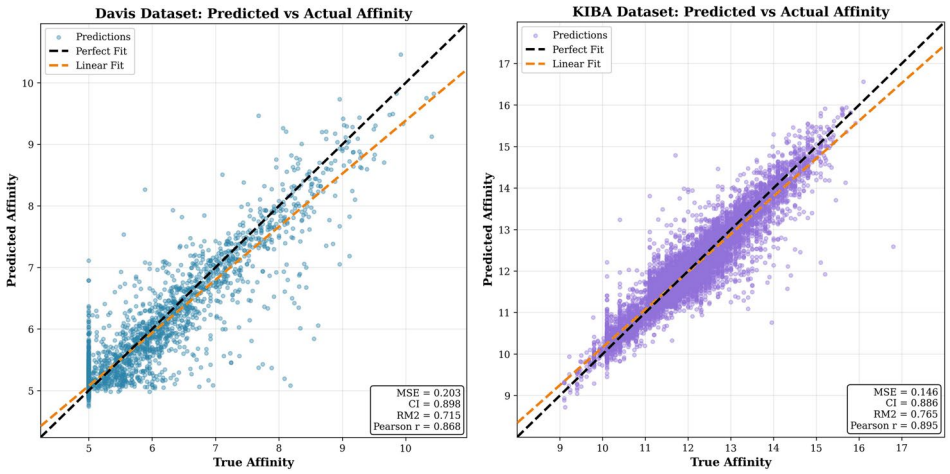

Fig. 2. Scatter plot comparing experimental binding affinity values (x-axis) with model-predicted affinities (y-axis) for the DTA regression task.

indicating general agreement between predicted and true values. A noticeable vertical clustering appears in the lower-affinity region (true values around 5–7), which is primarily due to the high sample density and limited variation of true values in this interval, resulting in relatively dispersed predictions. In contrast, the KIBA dataset covers a wider affinity range, producing a denser and more evenly distributed scatter pattern. The fitted regression line lies closer to the ideal line, suggesting reduced systematic bias. We further present the Pearson correlation coefficient in the plots ($r = 0.868$ for DAVIS and $r = 0.895$ for KIBA), which directly quantifies the linear association between predicted and true values. These correlations are consistent with the MSE, CI, and r_m^2 metrics reported in the tables, which provides evidence of the model's regression reliability.

These results suggest that MSCMF-DTB shows competitive predictive performance in large, heterogeneous kinase-focused datasets. The model's performance can be attributed to the complementary encoding of drugs via DenseGCN and fingerprint channels, multi-scale 1D convolution capturing local protein motifs, and fine-grained cross-modal interaction modeling using cross-attention and tensor networks. Collectively, these features enable MSCMF-DTB to balance accuracy, robustness, and generalizability across datasets of varying size, complexity, and kinase specificity. It should be noted that although traditional five-fold cross-validation is widely adopted in DTI and DTA predictions, it may result in overestimation of model performance since structurally similar compounds can appear in both training and test sets. This scenario does not fully reflect the practical challenge of predicting interactions for structurally novel molecules. A more stringent evaluation strategy could be scaffold-based dataset partitioning, in which compounds having the same Bemis–Murcko scaffold are assigned exclusively to either the training or test set[54]. This approach reduces chemical similarity between data splits and provides a more rigorous assessment of a model's ability to generalize previously unseen molecular frameworks. However, in the present study, five-fold cross-validation was retained to enable direct comparison with previously published models.

Ablation experiments

To evaluate the effectiveness of each model component and understand their contributions to overall performance in DTI and DTA tasks, we conducted an ablation study on MSCMF-DTB using the following configurations: (1) MSCMF-DTB without the TAPE-BERT modules (wo_1), where the module was removed and protein features were instead encoded using Word2Vec; (2) MSCMF-DTB without molecular fingerprint features (wo_2), in which the fingerprint-based secondary drug channel was excluded from the final feature fusion; (3) MSCMF-DTB without residual/dense connections in the drug encoder (wo_3), replacing DenseGCN with a standard graph convolutional network (GCN) to assess the contribution of dense connectivity; (4) MSCMF-DTB without the Tensor-Network interaction module (wo_4), where the Tensor-Network was removed and the model relied solely on remaining fusion mechanisms; and (5) the complete MSCMF-DTB model (full), including all designed modules and serving as the reference baseline.

The results are presented in Figs. 3, 4, 5, 6, 7 and 8 as boxplots, which illustrate the performance distributions and statistical stability of MSCMF-DTB and its ablated models on the binary classification datasets (Human, *C. elegans*, GPCR, BioSNAP, and DrugBank) and regression datasets (DAVIS and KIBA). On the classification datasets, removing TAPE-BERT and substituting Word2Vec (wo_1) caused the most significant performance drop on large and heterogeneous datasets (BioSNAP and DrugBank), highlighting the critical role of pretrained semantic embeddings in capturing protein sequence and structural features. Excluding the fingerprint-based drug channel (wo_2) led to noticeable declines in AUC, AUPR, Accuracy, and F1, indicating that fingerprint features provide complementary substructure- and fragment-level information beyond atomic-level graph features. Replacing DenseGCN with a standard GCN (wo_3) reduced Precision and model stability, validating the value of dense connectivity in enhancing drug molecular representations. Finally, removing the Tensor-

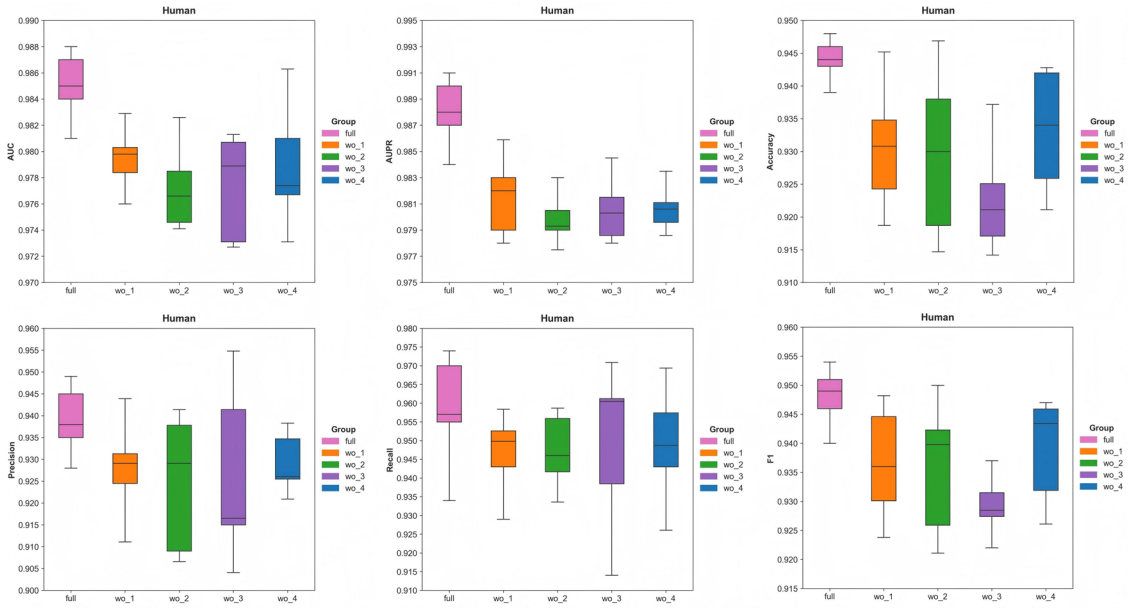

Fig. 3. Performance comparison of MSCMF-DTB and its ablation models on the Human dataset in terms of AUC, AUPR, Accuracy, Precision, Recall, and F1. wo_1: MSCMF-DTB without the TAPE-BERT modules; wo_2: MSCMF-DTB without molecular fingerprint features; wo_3: MSCMF-DTB without residual/dense connections in the drug encoder; wo_4: MSCMF-DTB without the Tensor-Network interaction module; full: the complete MSCMF-DTB model.

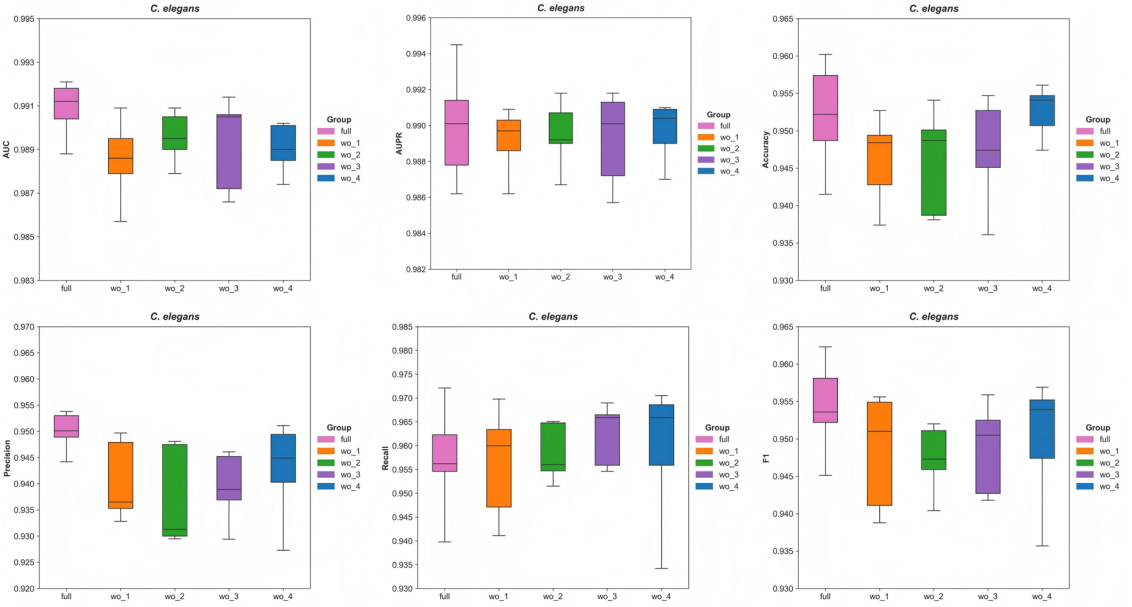

Fig. 4. Performance comparison of MSCMF-DTB and its ablation models on the *C. elegans* dataset in terms of AUC, AUPR, Accuracy, Precision, Recall, and F1. wo_1: MSCMF-DTB without the TAPE-BERT modules; wo_2: MSCMF-DTB without molecular fingerprint features; wo_3: MSCMF-DTB without residual/dense connections in the drug encoder; wo_4: MSCMF-DTB without the Tensor-Network interaction module; full: the complete MSCMF-DTB model.

Network (wo_4) caused a pronounced decrease in overall performance, indicative of the importance of high-order interaction modeling for accurate DTI prediction.

For the regression datasets, the boxplots similarly show that the full MSCMF-DTB model exhibits competitive statistical performance: the median MSE is lower, and the median r_m^2 values are higher with tighter distributions compared to all ablated settings. Notably, removing TAPE-BERT (wo_1) or the Tensor-Network (wo_4) had the

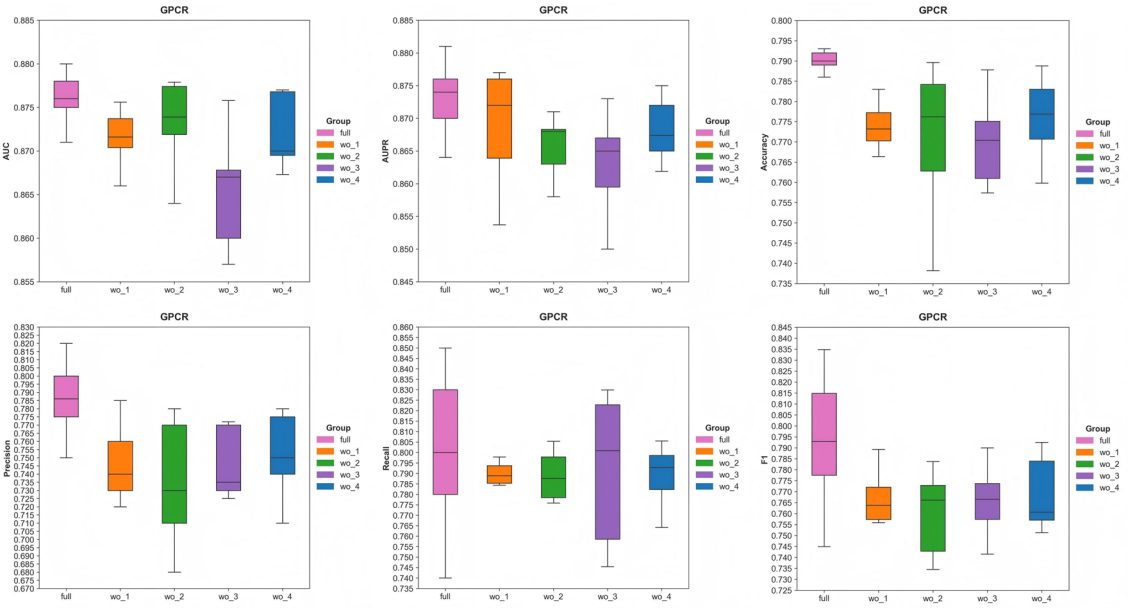

Fig. 5. Performance comparison of MSCMF-DTB and its ablation models on the GPCR dataset in terms of AUC, AUPR, Accuracy, Precision, Recall, and F1. wo_1: MSCMF-DTB without the TAPE-BERT modules; wo_2: MSCMF-DTB without molecular fingerprint features; wo_3: MSCMF-DTB without residual/dense connections in the drug encoder; wo_4: MSCMF-DTB without the Tensor-Network interaction module; full: the complete MSCMF-DTB model.

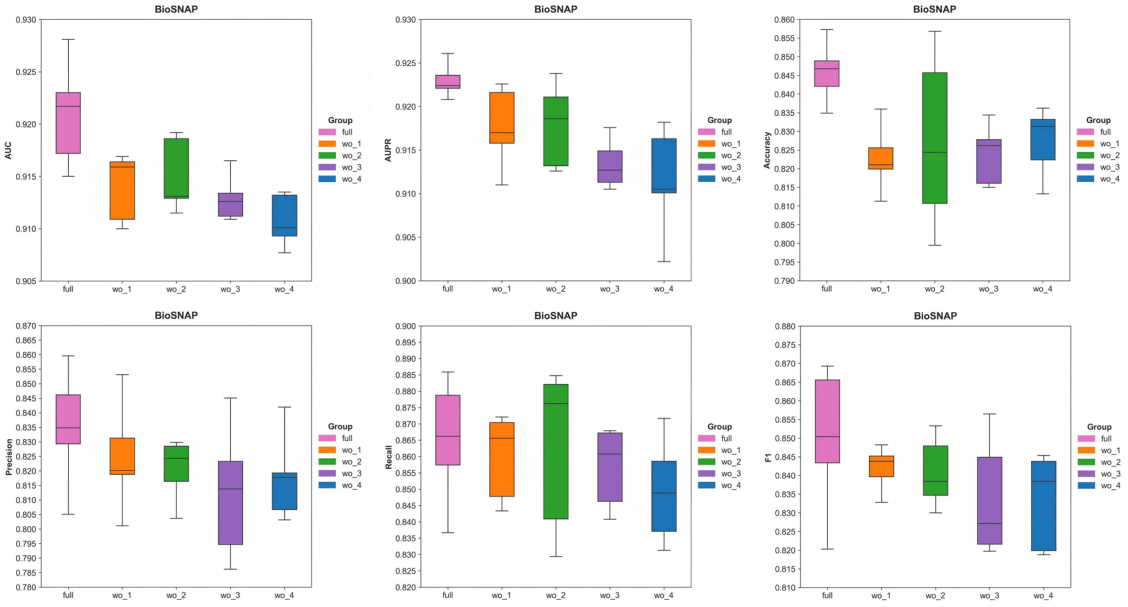

Fig. 6. Performance comparison of MSCMF-DTB and its ablation models on the BioSNAP dataset in terms of AUC, AUPR, Accuracy, Precision, Recall, and F1. wo_1: MSCMF-DTB without the TAPE-BERT modules; wo_2: MSCMF-DTB without molecular fingerprint features; wo_3: MSCMF-DTB without residual/dense connections in the drug encoder; wo_4: MSCMF-DTB without the Tensor-Network interaction module; full: the complete MSCMF-DTB model.

most pronounced effect on regression performance, emphasizing the critical roles of pretrained embeddings and high-order interaction modeling in DTA prediction. These observations indicate that the performance differences are primarily driven by variations in input representations, drug-side branches, and the interaction layer. It should be noted that the protein sequence encoder remained unchanged across all ablation settings, as

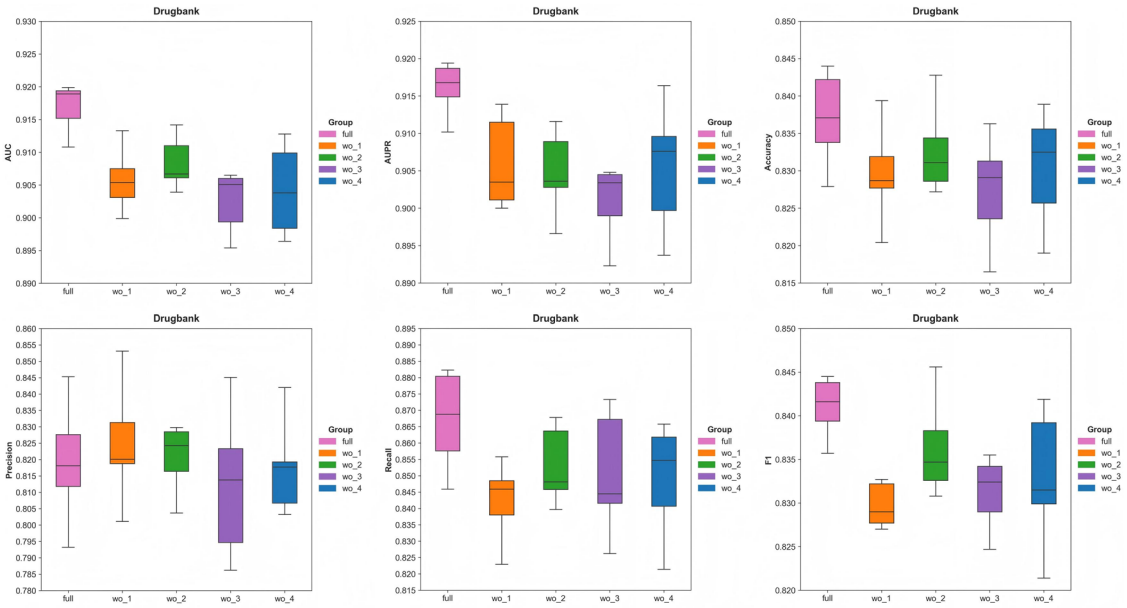

Fig. 7. Performance comparison of MSCMF-DTB and its ablation models on the DrugBank dataset in terms of AUC, AUPR, Accuracy, Precision, Recall, and F1. wo_1: MSCMF-DTB without the TAPE-BERT modules; wo_2: MSCMF-DTB without molecular fingerprint features; wo_3: MSCMF-DTB without residual/dense connections in the drug encoder; wo_4: MSCMF-DTB without the Tensor-Network interaction module; full: the complete MSCMF-DTB model.

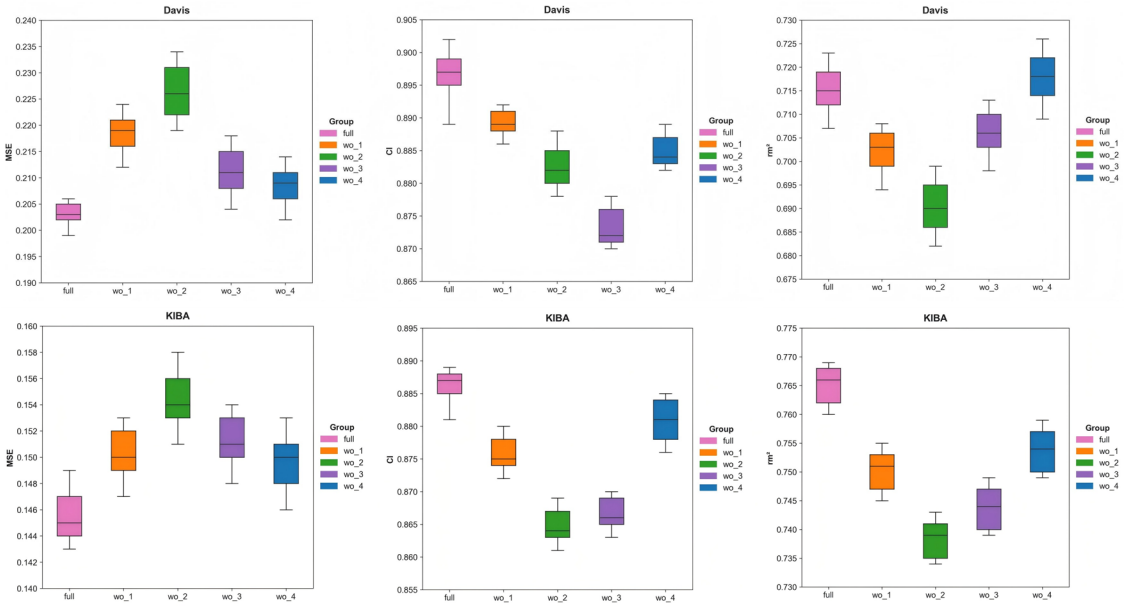

Fig. 8. Performance comparison of MSCMF-DTB and its ablation models on the DAVIS and KIBA datasets in terms of MSE, CI, and r_m^2. wo_1: MSCMF-DTB without the TAPE-BERT modules; wo_2: MSCMF-DTB without molecular fingerprint features; wo_3: MSCMF-DTB without residual/dense connections in the drug encoder; wo_4: MSCMF-DTB without the Tensor-Network interaction module; full: the complete MSCMF-DTB model.

it is responsible for multi-scale local pattern extraction and sequence-level aggregation on top of pretrained or alternative embeddings.

Collectively, MSCMF-DTB's improved performance arises from the complementary and synergistic contributions of its modules: the representation layer ensures rich input features; the molecular fingerprint channel and DenseGCN supply fragment-level and multi-level graph information; the protein-side encoder

captures local sequence patterns; and the Tensor-Network effectively models complex cross-modal high-order interactions. These components enable the full model to achieve a more comprehensive and accurate characterization of drug–target relationships, with demonstrated effectiveness and stability across both classification and regression tasks. It is worth noting that while most performance metrics decrease when individual modules are removed, some metrics remain almost unchanged. This phenomenon is very common in ablation studies due to trade-offs between complementary features, metric sensitivity, and dataset-specific effects, and does not undermine the overall contribution of each module to model performance.

Case studies

While MSCMF-DTB exhibits competitive predictive performance on standard benchmark datasets, it faces two key challenges: model interpretability and practical applicability. First, enhancing the interpretability of DL-based models is crucial for understanding which molecular or sequence features drive predictions and for assessing whether the model focuses on biologically meaningful binding sites. Second, benchmark datasets often include previously observed drug–target pairs, which may not fully reflect a model's ability to generalize to novel targets. By conducting "cold-start" experiments, we can simulate a more realistic scenario in which the target is entirely unseen during training. Therefore, two case studies were conducted to complement benchmark evaluations and guide improvements in both accuracy and explainability.

Case study 1: Attention weight visualization

To assess whether the cross-attention mechanism in MSCMF-DTB effectively identifies biologically relevant regions of target proteins, we conducted a case study using two drug–target complexes with experimentally determined binding sites: epidermal growth factor receptor (EGFR) complexed with Neratinib (PDB ID: 2JIV)[55] and cyclin-dependent kinase 6 (CDK6) complexed with Palbociclib (PDB ID: 2EUF)[56]. Protein residues assigned high attention weights (≥ 0.7) by the model were mapped onto the corresponding three-dimensional protein structures. As illustrated in Fig. 9, experimentally confirmed binding residues are highlighted in red, residues assigned high attention weight from the model are highlighted in yellow, and residues identified by both experiment and DL model are highlighted in green. While only a subset of true binding residues is captured by the attention mechanism, these regions correspond to key interaction sites and provide a meaningful visualization of potential binding regions.

To provide quantitative support, we report the overlap between experimentally determined residues and attention-identified residues in Table 8. Specifically, for the Neratinib–EGFR complex, 5 out of 9 experimentally confirmed binding residues were identified by the attention mechanism (overlap rate = 0.556), while for the Palbociclib–CDK6 complex, 5 out of 8 true binding residues were captured (overlap rate = 0.625). In total, the model assigned high attention weights to 77 and 128 residues in EGFR and CDK6, respectively, indicating

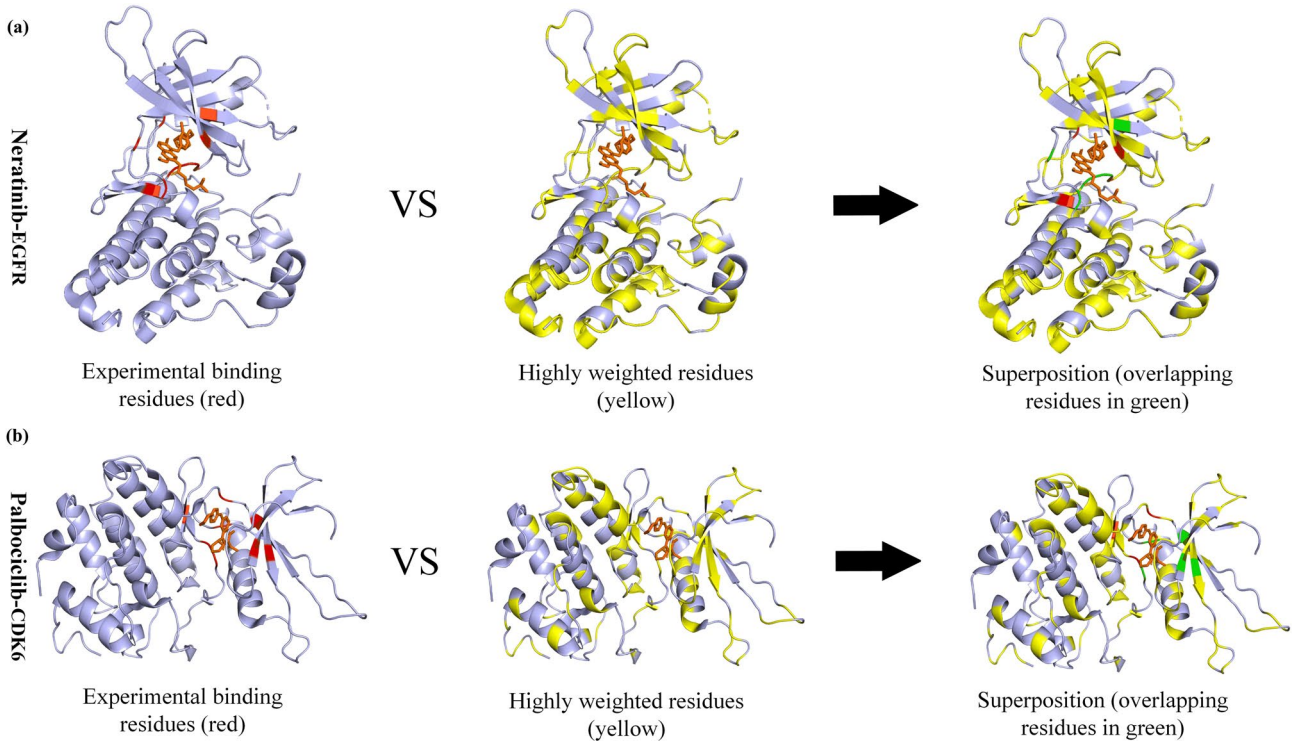

Fig. 9. Visualization of attention weights. (a) Mapping of interaction attention onto the Neratinib–epidermal growth factor receptor (EGFR) complex (PDB ID: 2JIV); (b) mapping of interaction attention onto the Palbociclib–cyclin-dependent kinase 6 (CDK6) complex (PDB ID: 2EUF).

Drug-Target Complex (PDB ID)	TBR	AIR	TP	FN	FP	Binding Overlap Rate(TP/TBR)
Neratinib-EGFR (2JIV)	9	77	5	4	72	0.556
Palbociclib-CDK6 (2EUF)	8	128	5	3	123	0.625

Table 8. Statistics of residues identified by the cross-attention mechanism in MSCMF-DTB. The table summarizes, for each drug–target complex, the total number of experimentally confirmed binding residues (True Binding Residues, TBR), the total number of residues identified by the attention mechanism (Attention-Identified Residues, AIR), the number of overlapping residues (True Positives, TP), false negatives (FN, experimentally confirmed residues not identified), false positives (FP, residues identified by attention but not experimentally confirmed), and the overlap ratio (TP/TBR).

Rank	Compounds	Experiments
1	1-methyl-8-(phenylamino)−4,5-dihydro-1 H-pyrazolo[4,3-h]quinazoline-3-carboxylic acid	Unconfirmed
2	(2 S,3 S)−3-[3-[2-chloro-4-(methylsulfonyl)phenyl]−1,2,4-oxadiazol-5-yl]−1-cyclopentylidene-4-cyclopropyl-1-fluorobutan-2-amine	Unconfirmed
3	**MK-2206**	**Confirmed**
4	2-Amino-4-fluoro-5-[(1-methyl-1 H-imidazol-2-yl)sulfanyl]-N-(1,3-thiazol-2-yl)benzamide	Unconfirmed
5	2-(2,6-Difluorophenoxy)-N-(2-fluorophenyl)−9-isopropyl-9 H-purin-8-amine	Unconfirmed
6	**Pifusertib**	**Confirmed**
7	3-bromo-5-phenyl-N-(pyrimidin-5-ylmethyl)pyrazolo[1,5-a]pyridin-7-amine	Unconfirmed
8	8-(Methylsulfonylamino)quinoline	Unconfirmed
9	MK-1421 (A-443654)	Unconfirmed
10	**Rizavasertib**	**Confirmed**
11	**BAY1125976**	**Confirmed**
12	(5-Phenyl-7-(pyridin-3-ylmethylamino)pyrazolo[1,5-a]pyrimidin-3-yl)methanol	Unconfirmed
13	**GSK690693**	**Confirmed**
14	Candesartan	Unconfirmed
15	Lesopitron	Unconfirmed

Table 9. Top 15 candidate inhibitors predicted by MSCMF-DTB for the target AKT1 in an unseen-target cold-start experiment. Experimentally validated inhibitors are highlighted in bold.

that attention is broadly distributed but still enriched around experimentally validated binding regions. These results show that the cross-attention mechanism may be capable of highlighting core binding residues, even though some true sites are missed (false negatives) and additional non-binding residues receive high attention (false positives). Overall, the attention analysis provides biologically meaningful insights into protein–drug interactions, and helps reduce the search space for potential binding regions while enhancing the interpretability of the model's predictions. The Python code for visualizing residues with varying attention weights is provided in the Supplementary Materials.

Case study 2: Unseen-target cold-start experiment

To further assess whether the model's predictions truly reflect generalization to a previously unseen target rather than memorization, we conducted another case study on the AKT1 target (UniProt ID: P31749)[57]. Specifically, approximately 20,000 DTI data points were randomly sampled from the aforementioned databases (Human, *C. elegans*, GPCR, BioSNAP, and DrugBank) and supplemented with information on relevant therapeutic compounds. Among these compounds, 13 have been previously validated through in vitro or in vivo experiments as the inhibitors to AKT1, both allosteric and orthosteric. To simulate a true unseen-target scenario, all DTI entries involving AKT1 were removed from the training and validation sets, including its sequence-derived features and all associated interaction labels. However, drugs known to interact with AKT1 were retained in the training data through their interactions with other targets. This data split follows a strict target-disjoint principle, ensuring that no targets overlap between the training/validation data and the cold-start evaluation stage, and hence preventing target-level information leakage.

The model was subsequently applied to evaluate and rank the interaction probability between AKT1 and all 8138 candidate drugs in the screening library. As shown in Tables 8 and 9 the top 15 predicted compounds are listed, and 5 of them are included in the 13 AKT1 inhibitors. Notably, three of the top ten predictions (MK-2206, Pifusertib, and Rizavasertib) are known allosteric inhibitors of AKT1, while BAY1125976 (an allosteric inhibitor) and GSK690693 (an orthosteric inhibitor) were also identified among the top 15. These results indicate that our model is capable of accurately predicting interactions between drugs and previously unseen targets from large datasets, indicative of its potential value for drug screening and drug repurposing applications. It should be noted that this cold-start case study is intended as an application-oriented demonstration, but incorporating baseline comparisons under the same cold-start protocol, statistical significance analysis, and predictive uncertainty estimation would allow for a more rigorous quantitative assessment.

Hyperparameter	Value
Number of training epochs	60
Dropout	0.1
Learning rate	5×10^{-4}
Weight decay	1×10^{-3}
Drug atom feature dimension	34
Protein sequence embedding dimension	768
Fingerprint vocabulary size	10,000
Hidden layer dimension	64
Number of GCN layers	2
Number of multi-head attention heads	8
Activation function of CNN layers	ReLU
Activation function of DenseGCN layers	ReLU

Table 10. Hyperparameter settings for MSCMF-DTB and all competing models.

	Human	C.elegans	GPCR	BioSNAP	DrugBank
Number of drugs	**2726**	**1767**	**5098**	**4504**	**6707**
Number of proteins	2001	1876	351	2180	4791
Total number of samples	6728	7786	13,796	27,200	36,930
Number of positive samples	3364	3893	6898	13,600	18,465
Number of negative samples	3364	3893	6898	13,600	18,465

Table 11. Statistics of all five datasets for the DTI task.

Materials and methods

Hyperparameter Settings

The training, validation, and testing of our model were conducted on a vGPU (32 GB) environment using Python 3.10, CUDA 12.1, and PyTorch 2.1.0. We employed the AdamW optimizer[58], an adaptive first-order optimization method that decouples weight decay from the gradient update, offering faster convergence and more stable training compared with standard stochastic gradient descent (SGD). All experiments were performed using five-fold cross-validation. Model selection and early stopping were based on validation AUC for the DTI task and MSE for the DTA task, with the learning rate adaptively reduced when validation performance plateaued. Key hyperparameter settings are provided in Table 10. Hyperparameter values were determined through a systematic grid search guided by quantitative validation. Candidate ranges were explored for learning rate, dropout, hidden layer dimensions, and attention heads. The final settings were chosen to balance predictive performance, model stability, and computational efficiency across datasets of varying size and complexity. Specifically, the number of attention heads was tuned within the candidate set [4, 8, 16, 32], while the candidate set for the hidden layer dimensions was [16, 32, 64, 128, 256]. The learning rate was set to 5×10^{-4}, as GNN and attention-based architectures for DTI/DTA tasks are prone to training instability when the learning rate is too large; this value provided stable convergence across datasets. The dropout rate was fixed at 0.1, which effectively reduced overfitting on larger datasets (e.g., DrugBank and BioSNAP) while maintaining sufficient representation capacity for smaller datasets.

Datasets

For classification tasks, we employed five widely used benchmark datasets. The Human dataset includes 2726 drugs and 2001 protein targets; the *C. elegans* dataset contains 1767 drugs and 1876 protein targets; the GPCR dataset comprises 5098 drugs and 351 protein targets; the BioSNAP dataset includes 4504 drugs and 2180 protein targets; and the DrugBank dataset contains 6707 drugs and 4791 protein targets. After sampling and balancing, these datasets consist of 6728, 7786, 13,796, 27,200, and 36,930 samples, respectively, each with a 1:1 ratio of positive to negative interactions. All datasets were processed using a uniform preprocessing and data-splitting strategy to reduce class imbalance and ensure fair comparison. It should be noted that positive and negative interaction labels were assigned using a threshold of 6.0 based on the negative logarithm of binding affinity values (e.g., pIC_{50}, pEC_{50}, and pK_i), which is consistent with the data processing protocol described by the work of Chen et al.[59] Table 11 summarizes the detailed statistics of the five datasets.

For regression tasks, the DAVIS dataset contains 68 drugs, 442 protein targets, and 30,056 samples, with binding affinities ranging from 5.0 to 10.8, providing a moderate-scale benchmark for evaluating prediction accuracy in mid-range affinities. The KIBA dataset includes 2111 drugs, 229 protein targets, and 118,254 samples, covering a broader affinity range and larger sample size, which allows assessment of regression stability under large-scale and wide-affinity scenarios. Detailed statistics for all two datasets are summarized in Table 12.

Dataset	Number of Drugs	Number of Proteins	Total Samples	Affinity Range
DAVIS	68	442	30,056	5.0–10.8.0.8
KIBA	2111	229	118,254	0.0–17.2.0.2

Table 12. Statistics of all two datasets for the DTA task.

SOTA models

The following five models represent state-of-the-art approaches selected for comparison with our work for the DTI task. MolTrans[44] decomposes drug and protein sequences into explicit substructure sequences using a fragment pattern mining algorithm (FCS module) and enhances contextual embeddings with a Transformer encoder to capture long-range dependencies. It constructs pairwise interaction score maps of drug–protein substructures and employs CNNs to extract higher-order interaction features for probabilistic DTI prediction. Mutual-DTI[45] applies a GNN to capture spatial features of drug molecules and a gated convolutional network to extract local patterns from protein sequences; the features are fed into an enhanced Transformer decoder, augmented with a mutual-feature module to model complex interactions, with predictions generated via fully connected layers. AMMVF-DTI[46] extracts drug and protein features via GAT and BERT modules, generates graph-level representations through an ATT module, and then models drug–protein interactions using ITM and NTN modules. All resulting features are finally integrated and fed into an MLP for DTI classification. FMCA-DTI[47] uses branch-centric mining (BCM) and class-fragment mining (CFM) to extract multiple functional fragment types from drugs and proteins, which are first processed via CNNs to produce feature matrices and then passed through a shared-weight multi-head cross-attention module. Interaction features are ultimately used by the predictor to score drug–target interactions. DrugBAN[48] employs GCN and CNN to extract drug and protein features, explicitly models local interactions via a bilinear attention network (BAN), integrates a conditional domain adversarial network (CDAN) for cross-domain scenarios, and generates final predictions through an MLP based on combined interaction features.

To comprehensively benchmark the performance of our model on the DTA task, five state-of-the-art regression models were selected for comparison. DeepGS[49] encodes drug SMILES and protein amino acid sequences using Smi2Vec and Prot2Vec, respectively, extracts localized chemical context features through deep neural networks, and incorporates a novel molecular structure modeling module before performing end-to-end affinity prediction. DeepCDA[50] jointly applies CNNs, LSTMs, and a bidirectional attention mechanism to learn compound and protein representations, followed by adversarial domain adaptation to derive a domain-invariant encoder for affinity scoring. DeepFusionDTA[51] integrates multiple analysis modules to fuse sequence- and structure-based features of drugs and targets into a unified feature map, and employs a Bagging ensemble strategy for regression. MATT_DTI[12] enhances atomic representations using relation-aware self-attention to explicitly encode inter-atomic positional and bonding information, extracts drug and protein features via CNNs, and models cross-modal interactions using multi-head attention before producing affinity predictions. AttentionMGT-DTA[52] represents drugs as molecular graphs and targets as binding-pocket graphs, fuses multimodal information through a graph Transformer combined with intra- and inter-graph attention mechanisms, and outputs final affinity scores.

Evaluation metrics

The performance of a classification model is commonly characterized using four fundamental components: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). TP represents the number of instances for which the model correctly predicts a positive label, whereas TN denotes the number of correctly predicted negative instances. FP refers to cases in which the actual label is negative but the model incorrectly predicts it as positive, and FN corresponds instances where the actual label is positive but the model fails to identify it and predicts it as negative. Based on these four components, two additional metrics are derived: True Positive Rate (TPR) and False Positive Rate (FPR). TPR, also called recall or sensitivity, measures the proportion of actual positive samples that a model correctly identifies, reflecting how effectively true positives are captured. In contrast, FPR quantifies the proportion of actual negative samples that are incorrectly predicted as positive, indicating how frequently the model mistakenly flags negative cases as positive.

We selected six metrics for the DTI task: area under the receiver operating characteristic curve (AUC), the area under the precision–recall curve (AUPR), Accuracy, Precision, Recall, and F1-score. AUC provides an overall measure of the model's ability to distinguish positive from negative samples by integrating the TPR and FPR. It is defined as the area under the receiver operating characteristic (ROC) curve, where the x-axis represents the FPR and the y-axis represents the TPR, as shown in Eq. (1).

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR} \cdot d(\mathrm{FPR}) \tag{1}$$

AUPR measures a model's ability to detect positive samples, especially useful for imbalanced datasets, as it integrates Precision and Recall across thresholds, shown in Eq. (2).

$$\mathrm{AUPR} = \int_0^1 \mathrm{Precision} \cdot d(\mathrm{Recall}) \tag{2}$$

Accuracy reflects the overall proportion of correctly classified samples, shown in Eq. (3). Precision evaluates the correctness of positive predictions as shown in Eq. (4), while Recall measures the model's ability to identify all true positive instances as shown in Eq. (5). The F1-score, defined as the harmonic mean of Precision and Recall, balances predictive accuracy and completeness shown in Eq. (6).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

For regression evaluation, mean squared error (MSE), concordance index (CI), and r_m^2 were used to systematically assess the predictive performance of the models. Among them, MSE measures the average of the squared differences between predicted and true values, as defined as Eq. (7), and smaller values indicate higher quantitative accuracy in affinity prediction:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \tag{7}$$

where n denotes the total number of samples, y_i represents the actual value of the i-th sample, and $\hat{y}_i$ represents the model's predicted value for the i-th sample. CI, defined in Eq. (8), is used to evaluate the consistency between the predicted values and the true values in terms of their relative ranking:

$$\text{CI} = \frac{1}{Z}\sum_{y_i > y_j} h(\widehat{y_i} - \widehat{y_j}) \tag{8}$$

where $h(x)$ is a step function defined as 0 when $x<0$, 0.5 when $x=0$, and 1 when $x>0$; Z is the normalization constant that represents the total number of comparable sample pairs in the dataset (i.e., all drug–target pairs with different affinity values). The CI value ranges from 0 to 1, with higher values indicating stronger predictive accuracy. r_m^2 score is commonly used to assess a model's goodness-of-fit and the robustness of a predictive model. It penalizes models where the regression line does not pass close to the origin and emphasizes predictive reliability, often defined as Eq. (9):

$$r_m^2 = r^2(1 - \sqrt{r^2 - r_0^2}) \tag{9}$$

where r^2 is the squared correlation coefficient between true values y and predicted values $\hat{y}$, and r_0^2 is the squared correlation coefficient when the regression line is forced through the origin. The closer r_m^2 is to 1, the better the model's predictive performance. The Pearson correlation coefficient r, defined in Eq. (10), quantifies linear relationship between predicted and true values, with $r=1$ representing a perfect positive linear correlation, $r=-1$ representing a perfect negative linear correlation, and $r=0$ representing no linear correlation:

$$r = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(\widehat{y_i} - \overline{\widehat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}\sqrt{\sum_{i=1}^{n}(\widehat{y_i} - \overline{\widehat{y}})^2}} \tag{10}$$

where $\bar{y}$ denotes the mean of true values.

Components of our framework

The MSCMF-DTB framework integrates multiple components to process heterogeneous drug and protein information, including RDKit-based molecular graph construction, TAPE-BERT protein embeddings, fingerprint channel, multi-scale CNNs, DenseGCN, Cross Attention, and a Tensor Network. The following section provides a brief explanation of the underlying mechanisms of these key components.

RDKit

To process raw drug SMILES strings, the model first parses the chemical inputs into machine-readable representations. SMILES (Simplified Molecular Input Line Entry System) is a compact textual format for encoding chemical structures, capturing both atomic composition and bonding relationships. RDKit[60] is employed to convert SMILES into structured molecular representations. It is a widely used open-source cheminformatics toolkit capable of generating diverse molecular descriptors, including topological indices, 3D conformational information, and physicochemical properties. Then it produces both atom-level feature vectors and the molecular adjacency matrix. Each atom is encoded with features such as atomic symbol, valence, formal

charge, number of radical electrons, hybridization state, and aromaticity, while the adjacency matrix specifies bonding relationships, where entries of "1" denote covalent bonds and "0" denote the absence of connections.

TAPE-BERT

To process protein amino-acid sequences, the model also needs to parse the biological inputs into machine-readable representations. As a result, on the protein side, the model employs the ProteinBERT encoder from the TAPE (Tasks Assessing Protein Embeddings) framework, which is commonly referred to as TAPE-BERT, to generate contextualized representations of amino acid sequences[61]. TAPE-BERT is built upon the BERT architecture and is pretrained in a self-supervised manner on millions of unlabeled protein sequences using masked language modeling. By tokenizing sequences with the IUPAC amino acid alphabet, the model learns residue embeddings that capture both local biochemical patterns and long-range contextual dependencies that are essential for understanding protein structure and function. Owing to its large-scale pretraining and multi-head self-attention mechanism, TAPE-BERT provides rich and transferable representations that have demonstrated strong performance across diverse protein-related tasks, including function prediction, secondary and tertiary structure modeling, and the analysis of protein–protein or protein–small molecule interactions. These advantages make it more informative and robust than traditional handcrafted descriptors or shallow sequence features. In this work, we directly use the per-residue embeddings output by TAPE-BERT as the protein feature input, which are subsequently refined by downstream modules for feature extraction and drug–target interaction modeling.

Fingerprint channel

To complement the atom-level graph representation generated by RDKit, the model integrates a fingerprint channel that captures fragment-level and composition-level structural information. Molecular fingerprints are standardized fixed-length feature vectors (typically binary or integer arrays) that encode the presence or absence of predefined structural motifs or substructures. One representative example includes Extended Connectivity Fingerprints (ECFP)[62], which iteratively expand the chemical environment surrounding each atom and hash the resulting substructure patterns. Another example is MACCS keys[63], which assess the existence of 166 predefined functional groups or structural features. In the proposed fingerprint channel, hydrogen-completed atomic symbols are first mapped to discrete indices, then embedded and aggregated via average pooling to produce a molecule-level descriptor. This representation provides a compact summary of compositional statistics and complements the graph-based molecular encoding by adding fragment-level information that may not be fully captured through topology alone.

DenseGCN

Graph Convolutional Networks (GCNs), as a representative framework for processing graph-structured data, propagate messages and aggregate features based on adjacency relationships, enabling nodes to integrate multi-order neighborhood information and learn discriminative representations[64]. In this study, we adopt an enhanced graph convolutional architecture, DenseGCN, which incorporates densely stacked graph convolutional layers together with residual connections to further strengthen its ability to capture structural features from complex molecular graphs[65]. Compared with conventional GCNs, DenseGCN uses deep feature aggregation and residual preservation to more effectively extract multi-level neighborhood information. Specifically, an atom list is first constructed from each molecular structure, and an adjacency matrix is generated where an entry of 1 indicates a bond between two atoms and 0 indicates no connection. The length of the atom list corresponds to the total number of atoms in the molecule, and the resulting adjacency matrix has dimensions $n_{\text{atom}} \times n_{\text{atom}}$. After processing through the embedding layer, we obtain the molecular feature matrix $X \in R^{n\text{atom} \times dim}$, where *dim* denotes the dimensionality of the atom embedding vectors. In DenseGCN, each layer first applies a linear transformation to the node features and then performs neighborhood aggregation through the adjacency matrix through Eq. (11):

$$H^{(l)} = \sigma(AH^{(l-1)}W^{(l)} + b^{(l)}) \tag{11}$$

where σ is a nonlinear activation function (ReLU); A is the adjacency matrix; $H^{(0)}$ equals X; $W^{(l)}$ and $b^{(l)}$ are learnable parameters. Unlike classical GCNs, a residual connection is introduced in each convolutional layer; when the input and output dimensions are consistent, the input features are directly added to the output using Eq. (12):

$$H^{(l)} = H^{(l)} + H^{(l-1)} \tag{12}$$

This design mitigates the over-smoothing problem in deep GCNs and improves both training stability and feature expressiveness. DenseGCN is constructed by stacking multiple residual GCN layers through Eq. (13):

$$H^{(L)} = \text{DenseGCN}(H^{(0)}) \tag{13}$$

This equation allows the final molecular representation $H^{(L)}$ to integrate information from multiple levels of atomic neighborhoods, thereby enhancing semantic understanding of the nodes and supporting downstream tasks such as graph classification or interaction prediction. Finally, mean pooling is applied across all node features to obtain a molecule-level representation vector $h_{\text{mol}} \in R^{dim}$, which is used for subsequent protein interaction modeling and prediction through Eq. (14):

$$h_{mol} = \frac{1}{n_{atom}} \sum_{i=1}^{n_{atom}} H_i^{(L)} \tag{14}$$

CNN

Protein sequences contain rich local context and short-range dependencies. One-dimensional convolution, which is well suited for capturing such neighborhood patterns along the sequence axis, offers efficient parameter sharing and strong parallelism, facilitating stable training on large datasets. Our model uses the continuous sequence embeddings generated by TAPE-BERT as input, without introducing any additional handcrafted features. Multi-scale 1D convolutional kernels K = [3,5,7] are applied in parallel along the sequence dimension to capture local contextual patterns of varying lengths. Global max-pooling is then performed over the temporal (sequence) dimension to produce a fixed-length representation independent of sequence length. Finally, dropout regularization and a linear projection are applied to obtain a compact hidden vector, which serves as the input for downstream classification or regression tasks.

Cross-Attention

Cross-attention is a widely used mechanism in deep learning for processing sequential data. Its core function is to establish dependencies between different input sequences, enabling the model to more effectively capture cross-sequence information. It is commonly employed in sequence-to-sequence tasks such as machine translation and text generation, and has been broadly adopted in both natural language processing (NLP) and computer vision (CV). In contrast to self-attention that models dependencies within a single sequence, cross-attention is inherently asymmetric: one sequence provides the queries (Q), while the other supplies the keys and values (K/V). The model computes relevance scores between Q and K and uses them to produce a weighted aggregation of V, thereby conditionally integrating information from one sequence into the other. In practice, cross-attention is typically implemented with multi-head attention to capture diverse semantic subspaces in parallel, and can be combined with masking strategies to control the flow of information.

Tensor-Network

Effectively modeling the binding between drugs and targets is crucial for achieving high predictive performance. Conventional feature fusion strategies such as vector concatenation or dot product typically capture only shallow interactions and are insufficient for representing complex nonlinear relationships. To address this limitation, the present study incorporates a Tensor Network module to enhance the modeling capacity for drug–target feature interactions. Specifically, the module operates on the interaction-aware vectors generated by the Cross-Attention mechanism, which already integrate contextual information from both drug and target representations. The Tensor Network first performs an element-wise multiplication between the drug feature vector e_1 and the target feature vector e_2, explicitly capturing their dimensional-level correspondences through Eq. (15):

$$h = e_1 \odot e_2 \tag{15}$$

where $\odot$ represents the element-wise product. The resulting interaction vector h is then passed through two layers of nonlinear fully connected transformations to extract higher-order and more complex DTI patterns; Compared with traditional fusion methods such as concatenation or dot product, this design not only enhances the expressive power of feature fusion while maintaining computational efficiency but also amplifies informative cross-modal signals and suppresses irrelevant noise.

Conclusion

This work presents MSCMF-DTB, an end-to-end deep learning framework for drug–target binding prediction that supports both DTI classification and DTA regression. On the drug side, molecular graphs constructed via RDKit are encoded using a DenseGCN architecture, complemented by a parallel fingerprint channel that captures fragment-level and compositional features. On the protein side, contextual embeddings generated by TAPE-BERT are processed through a multi-scale 1D CNN module to extract local sequence patterns. Cross-modal DTIs are modeled using a decoder-style cross-attention mechanism integrated with a tensor network to capture higher-order feature interactions. The learned representations from the drug graph, drug fingerprint channel, protein sequence, and their cross-modal interaction are concatenated and fed into an MLP for final prediction.

Comprehensive classification experiments demonstrate that MSCMF-DTB achieves strong generalization across large and diverse datasets, maintaining stable and improved performance as dataset size and complexity increase. For instance, MSCMF-DTB achieved up to 3.2% improvement in AUC and 6.1% improvement in Recall over the second-best model (DrugBAN) on the large-scale DrugBank dataset. Regression results further show advantages in error control, ranking consistency, and cross-dataset robustness compared with some SOTA models. In particular, on the larger and more heterogeneous KIBA dataset, the model maintained competitive performance with an MSE of 0.146, a CI of 0.886, and a r_m^2 of 0.765. Interpretability analyses visualize attention distributions, revealing biologically meaningful interaction regions and supporting the model's mechanistic reliability. Finally, a cold-start case study on AKT1 confirms the practical applicability of MSCMF-DTB in virtual screening, enabling effective drug repurposing and the discovery of novel candidate inhibitors.

Although MSCMF-DTB demonstrates robust and transferable performance in predicting DTIs and DTAs, further improvements in DTB predictions should emphasize not only architectural advances but also multi-level validation. To generate results that are truly feasible for downstream in vitro and in vivo studies in drug discovery and repurposing, computational pipelines should integrate complementary methodologies that provide different

mechanistic perspectives. Physics-based approaches, including molecular docking[66], molecular dynamics simulations[67], and enhanced free-energy calculations[68], can work synergistically with DL models to establish a more reliable and cost-effective platform for early-stage compound prioritization, reducing the experimental burden traditionally required at this stage.

Data availability

The datasets generated during and/or analysed during the current study are available in the github repository: https://github.com/frankchenqu/MSCMF-DTB.

Received: 17 December 2025; Accepted: 9 March 2026
Published online: 12 March 2026

References

1. Klug, A. The discovery of the DNA double helix. *J. Mol. Biol.* **335**, 3–26 (2004).
2. Kollmann, M. & Sourjik, V. In silico biology: from simulation to understanding. *Curr. Biol.* **17**, R132–134 (2007).
3. Chen, Q. Molecular insights into the impact of nanotube confinement on protein α-helical structures. *Mol. Phys.* **123**, e2556847 (2025).
4. Lu, X. & Huang, J. A thermodynamic investigation of amyloid precursor protein processing by human γ-secretase. *Commun. Biol.* **5**, 837 (2022).
5. Shan, Y. et al. How does a small molecule bind at a cryptic binding site? *PLoS Comput. Biol.* **18**, e1009817 (2022).
6. Jimenez-Luna, J., Grisoni, F., Weskamp, N. & Schneider, G. Artificial intelligence in drug discovery: recent advances and future perspectives. *Expert Opin. Drug Discov.* **16**, 949–959 (2021).
7. Debnath, K., Rana, P. & Ghosh, P. A survey on deep learning for drug-target binding prediction: models, benchmarks, evaluation, and case studies. *Brief. Bioinform.* **26**, bbaf491 (2025).
8. Deng, J., Yang, Z., Ojima, I., Samaras, D. & Wang, F. Artificial intelligence in drug discovery: applications and techniques. *Brief. Bioinform.* **23**, bbab430 (2022).
9. Cavasotto, C. N. & Di Filippo J. I. Artificial intelligence in the early stages of drug discovery. *Arch. Biochem. Biophys.* **698**, 108730 (2021).
10. Vefghi, A., Rahmati, Z. & Akbari, M. Drug-target interaction/affinity prediction: deep learning models and advances review. *Comput. Biol. Med.* **196**, 110438 (2025).
11. Wen, M. et al. Deep-learning-based drug-target interaction prediction. *J. Proteome Res.* **16**, 1401–1409 (2017).
12. Zeng, Y., Chen, X., Luo, Y., Li, X. & Peng, D. Deep drug-target binding affinity prediction with multiple attention blocks. *Brief. Bioinform.* **22**, bbab117 (2021).
13. Nguyen, T. et al. GraphDTA: predicting drug-target binding affinity with graph neural networks. *Bioinformatics* **37**, 1140–1147 (2021).
14. Zhao, Q., Zhao, H., Zheng, K. & Wang, J. HyperAttentionDTI: improving drug-protein interaction prediction by sequence-based deep learning with attention mechanism. *Bioinformatics* **38**, 655–662 (2022).
15. Cheng, Z., Zhao, Q., Li, Y. & Wang, J. IIFDTI: predicting drug-target interactions through interactive and independent features based on attention mechanism. *Bioinformatics* **38**, 4153–4161 (2022).
16. Yang, Z., Zhong, W., Zhao, L. & Chen, C. Y. ML-DTI: mutual learning mechanism for interpretable drug-target interaction prediction. *J. Phys. Chem. Lett.* **12**, 4247–4261 (2021).
17. Yuan, W., Chen, G. & Chen, C. Y. FusionDTA: attention-based feature polymerizer and knowledge distillation for drug-target binding affinity prediction. *Brief. Bioinform.* **23**, bbab506 (2022).
18. Huang, J., Gao, J. & Chen, Q. An interpretable deep learning and molecular docking framework for repurposing existing drugs as inhibitors of SARS-CoV-2 main protease. *Molecules* **30**, 3409 (2025).
19. Pang, W., Chen, M. & Qin, Y. Prediction of anticancer drug sensitivity using an interpretable model guided by deep learning. *BMC Bioinform.* **25**, 182 (2024).
20. Kyro, G. W., Smaldone, A. M., Shee, Y., Xu, C. & Batista, V. S. T-ALPHA: a hierarchical transformer-based deep neural network for protein-ligand binding affinity prediction with uncertainty-aware self-learning for protein-specific alignment. *J. Chem. Inf. Model.* **65**, 2395–2415 (2025).
21. Zhao, Y. et al. Evidential deep learning-based drug-target interaction prediction. *Nat. Commun.* **16**, 6915 (2025).
22. Kanehisa, M. et al. KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36**, D480–484 (2008).
23. Schomburg, I., Chang, A. & Schomburg, D. BRENDA, enzyme data and metabolic information. *Nucleic Acids Res.* **30**, 47–49 (2002).
24. Günther, S. et al. SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res.* **36**, D919–922 (2008).
25. Wishart, D. S. et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34**, D668–672 (2006).
26. Liu, H., Sun, J., Guan, J., Zheng, J. & Zhou, S. Improving compound-protein interaction prediction by building up highly credible negative samples. *Bioinformatics* **31**, i221–229 (2015).
27. Kuhn, M. et al. STITCH 4: integration of protein-chemical interactions with user data. *Nucleic Acids Res.* **42**, D401–407 (2014).
28. Wishart, D. S. et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).
29. Zitnik, M., Sosic, R. & Leskovec, J. BioSNAP datasets: Stanford biomedical network dataset collection. (2018). http://snap.stanford.edu/biodata
30. Keshava Prasad, T. S. et al. Human protein reference database—2009 update. *Nucleic Acids Res.* **37**, D767–772 (2009).
31. Davis, A. P. et al. Comparative toxicogenomics database's 20th anniversary: update 2025. *Nucleic Acids Res.* **53**, D1328–D1334 (2025).
32. Kuhn, M., Letunic, I., Jensen, L. J. & Bork, P. The SIDER database of drugs and side effects. *Nucleic Acids Res.* **44**, D1075–1079 (2016).
33. Mysinger, M. M., Carchia, M., Irwin, J. J. & Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* **55**, 6582–6594 (2012).
34. Davis, M. I. et al. Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **29**, 1046–1051 (2011).
35. Tang, J. et al. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *J. Chem. Inf. Model.* **54**, 735–743 (2014).
36. Zdrazil, B. et al. The ChEMBL database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res.* **52**, D1180–D1192 (2024).
37. Burley, S. K. et al. Updated resources for exploring experimentally-determined PDB structures and computed structure models at the RCSB Protein Data Bank. *Nucleic Acids Res.* **53**, D564–D574 (2025).
38. Kim, S. et al. PubChem 2025 update. *Nucleic Acids Res.* **53**, D1516–D1525 (2025).

39. UniProt, C. UniProt: the universal protein knowledgebase in 2025. *Nucleic Acids Res.* **53**, D609–D617 (2025).
40. Liu, Z. et al. Forging the basis for developing protein-ligand interaction scoring functions. *Acc. Chem. Res.* **50**, 302–309 (2017).
41. Kang, X., Liu, X., Zou, Q. & Li, T. Luo X. CDFA: calibrated deep feature aggregation for screening synergistic drug combinations. *Front. Pharmacol.* **16**, 1608832 (2025).
42. Wang, Q. et al. BridgeSyn: a bridging fusion framework for drug combination synergy prediction. *Brief. Bioinform.* **26**, bbaf624 (2025).
43. Wang, Y. et al. Integrative graph-based framework for predicting circRNA drug resistance using disease contextualization and deep learning. *IEEE J. Biomed. Health Inf.* **29**, 7932–7944 (2025).
44. Huang, K., Xiao, C., Glass, L. M. & Sun, J. MolTrans: molecular interaction transformer for drug-target interaction prediction. *Bioinformatics* **37**, 830–836 (2021).
45. Wen, J. et al. Mutual-DTI: a mutual interaction feature-based neural network for drug-target protein interaction prediction. *Math. Biosci. Eng.* **20**, 10610–10625 (2023).
46. Wang, L., Zhou, Y. & Chen, Q. AMMVF-DTI: a novel model predicting drug-target interactions based on attention mechanism and multi-view fusion. *Int. J. Mol. Sci.* **24**, 14142 (2023).
47. Zhang, Q. et al. FMCA-DTI: a fragment-oriented method based on a multihead cross attention mechanism to improve drug-target interaction prediction. *Bioinformatics* **40**, btae347 (2024).
48. Bai, P., Miljković, F., John, B. & Lu, H. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. *Nat. Mach. Intell.* **5**, 126–136 (2023).
49. Lin, X. DeepGS: deep representation learning of graphs and sequences for drug-target binding affinity prediction. *Preprint at.* https://doi.org/10.48550/arXiv.2003.13902 (2020).
50. Abbasi, K. et al. DeepCDA: deep cross-domain compound-protein affinity prediction through LSTM and convolutional neural networks. *Bioinformatics* **36**, 4633–4642 (2020).
51. Pu, Y., Li, J., Tang, J. & Guo, F. DeepFusionDTA: drug-target binding affinity prediction with information fusion and hybrid deep-learning ensemble model. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **19**, 2760–2769 (2022).
52. Wu, H. et al. AttentionMGT-DTA: a multi-modal drug-target affinity prediction using graph transformer and attention mechanism. *Neural Netw.* **169**, 623–636 (2024).
53. Zhang, L., Wang, C. & Chen, X. Predicting drug-target binding affinity through molecule representation block based on multi-head attention and skip connection. *Brief. Bioinform.* **23**, bbac468 (2022).
54. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. molecular frameworks. *J. Med. Chem.* **39**, 2887–2893 (1996).
55. Yun, C. H. et al. The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. *Proc. Natl. Acad. Sci. USA* 105, 2070–2075 (2008).
56. Lu, H. & Schulze-Gahmen, U. Toward understanding the structural basis of cyclin-dependent kinase 6 specific inhibition. *J. Med. Chem.* **49**, 3826–3831 (2006).
57. Uko, N. E., Guner, O. F., Matesic, D. F. & Bowen, J. P. Akt pathway inhibitors. *Curr. Top. Med. Chem.* **20**, 883–900 (2020).
58. Zhou, P., Xie, X., Lin, Z. & Yan, S. Towards understanding convergence and generalization of AdamW. *IEEE Trans. Pattern Anal. Mach. Intell.* **46**, 6486–6493 (2024).
59. Chen, L. et al. TransformerCPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics* **36**, 4406–4414 (2020).
60. Bento, A. P. et al. An open source chemical structure curation pipeline using RDKit. *J. Cheminform.* **12**, 51 (2020).
61. Rao, R. B. et al. Evaluating protein transfer learning with TAPE. *Adv. Neural Inf. Process. Syst.* **32**, 9689 (2019).
62. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
63. Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **42**, 1273–1280 (2002).
64. Duan, S. et al. Semi-supervised classification of fundus images combined with CNN and GCN. *J. Appl. Clin. Med. Phys.* **23**, e13746 (2022).
65. Li, G. et al. DeepGCNs: making GCNs go as deep as CNNs. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 6923–6939 (2023).
66. Zhao, H. et al. Comprehensive evaluation of 10 docking programs on a diverse set of protein-cyclic peptide complexes. *J. Chem. Inf. Model.* **64**, 2112–2124 (2024).
67. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **9**, 646–652 (2002).
68. Fu, H., Zhu, Y. & Chen, Q. Free energy calculations in biomolecule-nanomaterial interactions. *Front. Phys.* **12**, 1469515 (2024).

Acknowledgements

Q. C. wishes to extend his appreciation to the Natural Science Foundation of Zhejiang Province for financial support (grant no. LY19B060002).

Author contributions

Conceptualization, Y.P. and Q.C.; investigation, Y.P.; methodology, Y.P.; writing—original draft, Y.P. and Q.C.; writing—review and editing, Q.C.; supervision, J.H.; funding acquisition, Q.C. All authors have read and agreed to the published version of the manuscript.

Funding

This work was financially supported by the Natural Science Foundation of Zhejiang Province (grant nos. LY19B060002).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Q.C.