FusionDTI: Fine-grained Binding Discovery with Token-level Fusion for Drug-Target Interaction

Anonymous ACL submission

Abstract

Predicting drug-target interaction (DTI) is critical in the drug discovery process. Despite remarkable advances in recent DTI models through the integration of representations from diverse drug and target encoders, such models often struggle to capture the fine-grained interactions between drugs and protein, i.e. the binding of specific drug atoms (or substructures) and key amino acids of proteins, which is crucial for understanding the binding mechanisms and optimising drug design. To address this issue, this paper introduces a novel model, called FusionDTI, which uses a token-level **Fusion** module to effectively learn fine-grained information for **D**rug-**T**arget **I**nteraction. In particular, our FusionDTI model uses the SELFIES representation of drugs to mitigate sequence fragment invalidation and incorporates the structure-aware (SA) vocabulary of target proteins to address the limitation of amino acid sequences in structural information, additionally leveraging pre-trained language models extensively trained on large-scale biomedical datasets as encoders to capture the complex information of drugs and targets. Experiments on three well-known benchmark datasets show that our proposed FusionDTI model achieves the best performance in DTI prediction compared with eight existing state-of-the-art baselines. Furthermore, our case study indicates that FusionDTI could highlight the potential binding sites, enhancing the explainability of the DTI prediction.[1]

1 Introduction

The task of predicting drug-target interactions (DTI) plays a pivotal role in the drug discovery progress, as it helps identify potential therapeutic effects of drugs on biological targets facilitating the development of effective treatments (Askr et al., 2023). DTI fundamentally relies on the binding of specific drug atoms (or substructures) and key amino acids of proteins (Schenone et al., 2013). In particular, each binding site is an interaction between a single amino acid and a single drug atom, which we refer to as a fine-grained interaction. For instance, Figure 1 B demonstrates the interaction between *HIV-1 protease* and the drug *lopinavir*. A critical component of this interaction is the formation of a hydrogen bond between a ketone group in lopinavir (represented in the SELFIES (Krenn et al., 2022) notation as [C][=O]) and the side chain of an aspartate residue Asp25 (i.e. Dd) within the protease (Brik and Wong, 2003; Chandwani and Shuter, 2008). Therefore, capturing such fine-grained interaction information during the fusion of drug and target representations is crucial for building effective DTI prediction models (Wu et al., 2022; Peng et al., 2024; Zeng et al., 2024).

To obtain representations of drugs and targets for the DTI task, some previous studies (Lee et al., 2019; Nguyen et al., 2021) have used graph neural networks (GNNs) or convolutional neural networks (CNNs) using a fixed-size window, potentially leading to a loss of contextual information, especially when drugs and targets are in a long-term sequence. These models directly concatenate the representations together to make predictions without considering fine-grained interactions. More recently, some computational models (Huang et al., 2021; Bai et al., 2023) employed the fusion module (e.g. Deep Interactive Inference Network (DIIN) (Gong et al., 2018) and Bilinear Attention Network (BAN) (Kim et al., 2018)) to obtain fine-grained interaction information and the 3-mer approach that binds three amino acids together as a target binding site to address the lack of structural information in the amino acid sequence. While useful for highlighting possible regions of interaction, these models do not offer the sufficient granularity needed to gauge the specifics of binding sites, as each binding site only contains one

[1] The complete code and datasets are available in the software section of the submission.

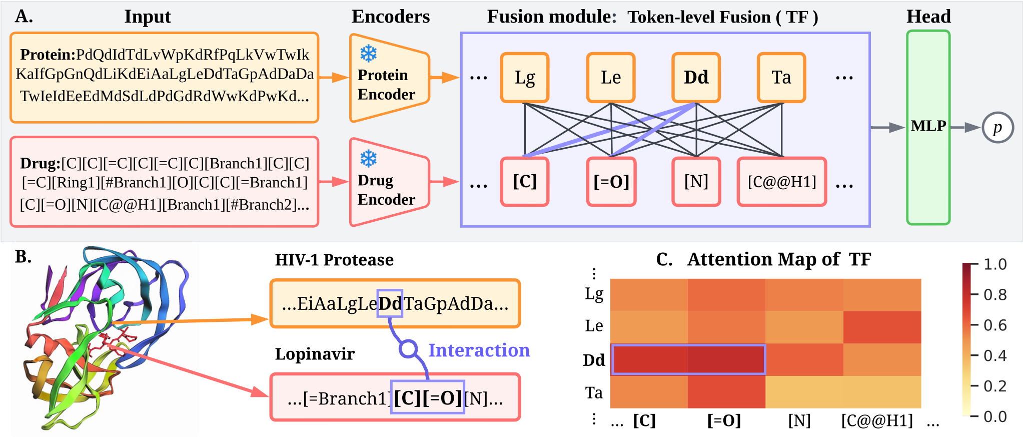
A.
Input
Encoders
Fusion module: Token-level Fusion (TF)
Head
Protein:PdQdIdTdLvWpKdRfPqLkVwTwIk KaIfGpGnQdLiKdEiAaLgLeDdTaGpAdDa TwIeIdEeEdMdSdLdPdGdRdWwKdPwKd...
Protein Encoder
Drug:[C][C][=C][C][=C][C][Branch1][C][C] [=C][Ring1][#Branch1][O][C][C][=Branch1] [C][=O][N][C@@H1][Branch1][#Branch2]...
Drug Encoder
Lg
Le
Dd
Ta
[C]
[=O]
[N]
[C@@H1]
MLP
p
B.
HIV-1 Protease
...EiAaLgLeDdTaGpAdDa...
Lopinavir
Interaction
...[=Branch1][C][=O][N]...
C. Attention Map of TF
Lg
Le
Dd
Ta
[C]
[=O]
[N]
[C@@H1]
1.0
0.8
0.6
0.4
0.2
0.0

Figure 1: **A**. Illustration of the FusionDTI model: frozen encoder, fusion module and classifier. The token-level fusion (TF) focuses on fine-grained interactions between tokens within and across sequences. **B**. This is a token-level interaction instance of HIV-1 protease and lopinavir. Lopinavir forms a hydrogen bond with residue Dd (Asp25) in the active site of the protease via its ketone molecule ([C][=O]). **C**. The attention map of TF visualises the weight between tokens, indicating the contribution of each drug atom and residue to the final prediction result.

residue (Schenone et al., 2013). Therefore, obtaining contextual representations of drugs and targets and capturing fine-grained interaction information for DTI remains challenging.

To address these challenges, we propose a novel model (called FusionDTI) with a Token-level Fusion (TF) module for an effective learning of fine-grained interactions between drugs and targets. In particular, our FusionDTI model utilises two pre-trained language models (PLMs), namely Saport (Su et al., 2023) as the protein encoder that is able to integrate both residue tokens with structure token; and SELFormer (Yüksel et al., 2023) as the drug encoder to ensure that each drug is valid and contains structural information. To effectively learn fine-grained information from these contextual representations of drugs and targets, we explore two strategies for the TF module, i.e. Bilinear Attention Network (BAN) (Kim et al., 2018) and Cross Attention Network (CAN) (Li et al., 2021; Vaswani et al., 2017), to find the best approach for integrating the rich contextual embeddings derived from Saport and SELFormer. We conduct a comprehensive performance comparison against eight existing state-of-the-art DTI prediction models. The results show that our proposed model achieves about 6% accuracy improvement over the best baseline on the BindingDB dataset. The main contributions of our study are as follows:

- We propose FusionDTI, a novel model that leverages PLMs to encode drug SELFIES, as well as protein residues and structures for rich semantic representations and uses the token-level fusion to capture fine-grained interaction between drugs and targets effectively.
- We compare two TF modules: CAN and BAN and analyse the influence of fusion scales based on FusionDTI, demonstrating that CAN is superior for DTI prediction both in terms of effectiveness and efficiency.
- We conduct a case study of three drug-target pairs by FusionDTI to evaluate whether potential binding sites would be highlighted for the DTI prediction explainability.

2 Related Work

2.1 Drug and Protein Representation

For drug molecules, most existing methods represent the input by the Simplified Molecular Input Line Entry System (SMILES) (Weininger, 1988; Weininger et al., 1989). However, SMILES suffers from numerous problems in terms of validity and robustness, and some valuable information about the drug structure may be lost which may prevent the model from efficiently mining the knowledge hidden in the data (Krenn et al., 2022). To address the limitations of SMILES, we apply SELFIES, a string-based representation that circumvents the issue of robustness and that always generates valid molecular graphs for each character.

Regarding proteins, the conventional approach uses amino acid sequences as model inputs (Huang et al., 2021; Bai et al., 2023), overlooking the crucial structural information of the protein. Inspired by the SA vocabulary of SaProt (Su et al., 2023), the SaProt enhances inputs by amalgamating each residue of the amino acid sequence with a 3D geometric feature that is obtained by encoding protein structure information using Foldseek (Van Kempen et al., 2024). This innovative combination offers richer protein representations through the SA vocabulary, contributing to the discovery of fine-grained interactions.

2.2 Molecular and Protein Language Models

Molecular language models trained on the large-scale molecular corpus capture the subtleties of chemical structures and their biological activities, setting new standards in the encoding of chemical compounds achieving meaningful representations (Ying et al., 2021; Rong et al., 2020). For example, MoLFormer (Ross et al., 2022) focused on leveraging the self-attention mechanism to interpret the complex, non-linear interactions within molecules, while SELFormer (Yüksel et al., 2023) employed SELFIES, ensuring valid and interpretable chemical structures.

Protein language models have revolutionized the way we understand and represent protein sequences, learning intricate patterns and features that define the protein functionality and interactions. ProtBERT (Elnaggar et al., 2021) and ESM (Lin et al., 2023) applied a transformer architecture to protein sequences, capturing the complex relationships between amino acids. Saport (Su et al., 2023, 2024) further enhanced this approach by integrating SA vocabularies to provide protein structure information.

3 Methodology

3.1 Model Architecture

Given a sequence-based input drug-target pair, the DTI prediction task aims to predict an interaction probability score $p \in [0, 1]$ between the given drug-target pair, which is typically achieved through learning a joint representation $\mathbf{F}$ space from the given sequence-based inputs. To address the DTI task and effectively capture fine-grained interaction, we proposed a novel model, called FusionDTI, which is a bi-encoder model (Liu et al., 2021) with a fusion module that fuses the representations of drugs and targets. The overall framework of FusionDTI is illustrated in Figure 1 A. In general, FusionDTI takes sequence-based inputs of drugs and targets, which are encoded into token-level representation vectors by two frozen encoders. Then, a fusion module fuses the representations to capture fine-grained binding information for a final prediction through a prediction head.

Input: The initial inputs of drugs and targets are string-based representations. For protein $\mathcal{P}$, the SA vocabulary (Su et al., 2023; Van Kempen et al., 2024) is employed, where each residue is replaced by one of 441 SA vocabularies that bind an amino acid to a 3D geometric feature to address the lack of structural information in amino acid sequences. For drug $\mathcal{D}$, as mentioned in the previous section, we use the SELFIES, which is a formal syntax that always generates valid molecular graphs (Krenn et al., 2022). We provide the steps and code to obtain SA and SELFIES in Appendix A.3.

Encoder: The proposed model contains two frozen encoders: Saport (Su et al., 2023) and SELFormer (Yüksel et al., 2023), which generate a drug representation $\mathbf{D}$ and a protein representation $\mathbf{P}$ separately. It is of note that FusionDTI is flexible enough to easily replace encoders with other PLMs or address SELFIES or SA representations that are unavailable. Furthermore, $\mathbf{D}$ and $\mathbf{P}$ are stored in memory for later-stage online training.

Fusion module: In developing FusionDTI, we have investigated two options for the fusion module: BAN and CAN to fuse representations, as indicated in Figure 2. The CAN is utilised to fuse each pair as $\mathbf{D}^*$ and $\mathbf{P}^*$, and then concatenate them into one $\mathbf{F}$ for fine-grained binding information. For BAN, we need to obtain bilinear attention maps and generate $\mathbf{F}$ through the bilinear pooling layer.

Prediction head: Finally, we obtain the probability score p of the DTI prediction by a multilayer perceptron (MLP) classifier trained with the binary cross-entropy loss, i.e. $p = \mathrm{MLP}(\mathbf{F})$.

Since the encoders and the fusion module constitute the key components of our FusionDTI model, we will describe them in detail in the following.

3.2 Drug and Protein Encoders

Employing sequences with detailed biological functions and structures is a critical step in exploring the fine-grained binding of drugs and targets. For drugs, SMILES is the most commonly used input sequence but suffers from invalid sequence segments and potential loss of structural information (Krenn et al., 2022). To address the limitations, we transform SMILES into SELFIES, a formal grammar that generates a valid molecular graph for each element (Krenn et al., 2022). Besides, to address the lack of structural information in the

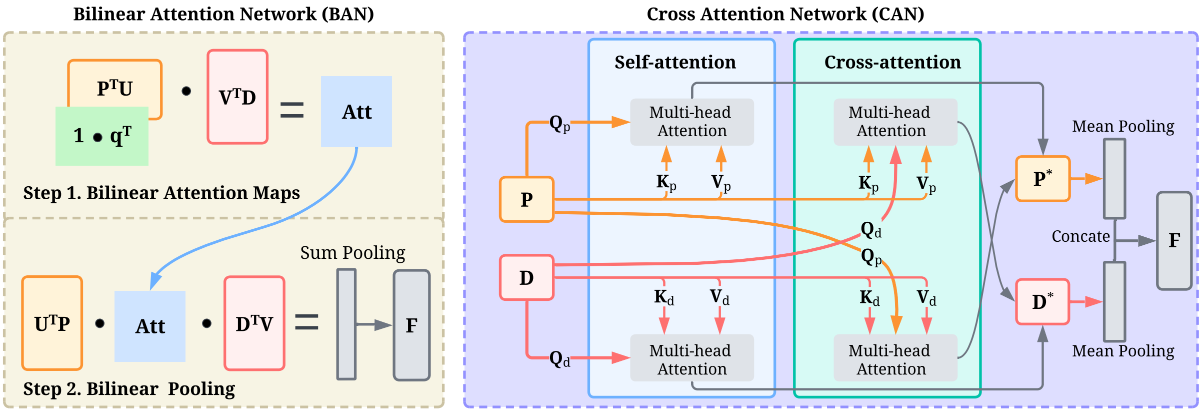

Figure 2: **BAN:** In step 1, the bilinear attention map is obtained by a bilinear interaction modelling via transformation matrices. In step 2, the joint representation $\mathbf{F}$ is generated using the attention map by bilinear pooling via the shared transformation matrices $\mathbf{U}$ and $\mathbf{V}$. **CAN:** It fuses protein and drug representations through multi-head, self-attention and cross-attention. Then fused representations $\mathbf{P}^*$ and $\mathbf{D}^*$ are concatenated into $\mathbf{F}$ after mean pooling.

amino acid sequences, we utilise the SA sequence of targets to combine each amino acid with an SA vocabulary by Foldseek (Van Kempen et al., 2024).

PLMs have shown promising achievements in the biomedical domain leveraging transformers since they pay attention to contextual information and are pre-trained on large-scale biomedical databases. Therefore, we utilise Saport (Su et al., 2023) as a protein encoder to encode protein input $\mathcal{P}$ of both the SA sequence and amino acid sequence. Meanwhile, SELFormer (Yüksel et al., 2023) is used as our drug encoder to encode the drug SELFIES input $\mathcal{D}$. Then these encoded protein representation $\mathbf{P}$ and drug representation $\mathbf{D}$ are further used as inputs for the later fusion module (Subsection 3.3). These rich contextual representations ensure that we can explore the fine-grained binding information effectively. To further justify this, we also compare our encoders with other existing protein language models (such as ESM-2 (Lin et al., 2023)) and molecular language models (such as MoLFormer (Ross et al., 2022) and ChemBERTa-2 (Ahmad et al., 2022)), and the results can be found in Appendix A.6.

3.3 Fusion Module

In order to capture the fine-grained binding information between a drug and a target, our FusionDTI model applies a fusion module to learn token-level interactions between the token representations of drugs and targets encoded by their respective encoders. As shown in Figure 2, two fusion modules are investigated to fuse representations: the Bilinear Attention Network (Kim et al., 2018) and the Cross Attention Network (Vaswani et al., 2017).

3.3.1 Bilinear Attention Network (BAN)

Motivated by DrugBAN (Bai et al., 2023), our model considers BAN (Kim et al., 2018) as an option to learn pairwise fine-grained interactions between drug $\mathbf{D} \in \mathbb{R}^{M\times\phi}$ and target $\mathbf{P} \in \mathbb{R}^{N\times\rho}$, denoted as FusionDTI-BAN. For BAN as indicated in Figure 2, bilinear attention maps are obtained by a bilinear interaction modelling to capture pairwise weights in step 1, and then the bilinear pooling layer to extract a joint representation $\mathbf{F}$. The equation of BAN is shown below:

$$\begin{aligned} \mathbf{F} &= \mathrm{BAN}(\mathbf{P}, \mathbf{D}; Att) \\ &= \mathrm{SumPool}(\sigma(\mathbf{P}^\top\mathbf{U}) \cdot Att \cdot \sigma(\mathbf{D}^\top\mathbf{V}), s), \end{aligned} \tag{1}$$

where $\mathbf{U} \in \mathbb{R}^{N\times K}$ and $\mathbf{V} \in \mathbb{R}^{M\times K}$ are transformation matrices for representations. SumPool is an operation that performs a one-dimensional and non-overlapped sum pooling operation with stride s and $\sigma(\cdot)$ denotes a non-linear activation function with $\mathrm{ReLU}(\cdot)$. $Att \in \mathbb{R}^{\rho\times\phi}$ represents the bilinear attention maps using the Hadamard product and matrix-matrix multiplication and is defined as:

$$Att = ((\mathbf{1}\cdot\mathbf{q}^\top) \circ \sigma(\mathbf{P}^\top\mathbf{U})) \cdot \sigma(\mathbf{V}^\top\mathbf{D}), \tag{2}$$

Here, $\mathbf{1} \in \mathbb{R}^\rho$ is a fixed all-ones vector, $\mathbf{q} \in \mathbb{R}^K$ is a learnable weight vector and $\circ$ denotes the Hadamard product. In this way, pairwise interactions contribute sub-structural pairs to predictions.

BAN captures the token-level interactions between the protein and drug representations without considering the relationships within each sequence itself, which may limit its ability to understand deeper contextual dependencies.

3.3.2 Cross Attention Network (CAN)

Inspired by ProST (Xu et al., 2023), we also consider CAN as our fusion module to learn fine-grained interaction information of drugs and targets. We denote our FusionDTI model that uses a CAN fusion module as FusionDTI-CAN. By processing $\mathbf{D} \in \mathbb{R}^{m \times h}$ and $\mathbf{P} \in \mathbb{R}^{n \times h}$ separately, the fused drug $\mathbf{D}^* \in \mathbb{R}^{m \times h}$ and target $\mathbf{P}^* \in \mathbb{R}^{n \times h}$ representations are obtained. To synthesise the fine-grained joint representation $\mathbf{F}$, we employ a pooling aggregation strategy for both $\mathbf{D}^*$ and $\mathbf{P}^*$independently and then concatenate them as shown in Figure 2. The process is described by the following equation:

$$\mathbf{F} = \mathrm{Concat}[\mathrm{MeanPool}(\mathbf{D}^*), \mathrm{MeanPool}(\mathbf{P}^*)], \tag{3}$$

where $\mathrm{MeanPool}$ calculates the element-wise mean of all tokens across the sequence dimension, and Concat denotes the concatenation of the resulting mean vectors. In this context, the multi-head, self-attention and cross-attention mechanisms are used to refine the representations of each residue and atom as below:

$$\mathbf{D}^* = \frac{1}{2}\left[\mathit{MHA}(\mathbf{Q}_d, \mathbf{K}_d, \mathbf{V}_d) + \mathit{MHA}(\mathbf{Q}_p, \mathbf{K}_d, \mathbf{V}_d)\right], \tag{4}$$

$$\mathbf{P}^* = \frac{1}{2}\left[\mathit{MHA}(\mathbf{Q}_p, \mathbf{K}_p, \mathbf{V}_p) + \mathit{MHA}(\mathbf{Q}_d, \mathbf{K}_p, \mathbf{V}_p)\right], \tag{5}$$

where $\mathbf{Q}_d, \mathbf{K}_d, \mathbf{V}_d \in \mathbb{R}^{m \times h}$ and $\mathbf{Q}_p, \mathbf{K}_p, \mathbf{V}_p \in \mathbb{R}^{n \times h}$ are the queries, keys and values for drug and target protein, respectively. And *MHA* denotes the Multi-head Attention mechanism. To guide this process, two distinct sets of projection matrices guide the attention mechanism as follows:

$$\mathbf{Q}_d = \mathbf{D}\mathbf{W}_q^d, \quad \mathbf{K}_d = \mathbf{D}\mathbf{W}_k^d, \quad \mathbf{V}_d = \mathbf{D}\mathbf{W}_v^d, \tag{6}$$

$$\mathbf{Q}_p = \mathbf{P}\mathbf{W}_q^p, \quad \mathbf{K}_p = \mathbf{P}\mathbf{W}_k^p, \quad \mathbf{V}_p = \mathbf{P}\mathbf{W}_v^p, \tag{7}$$

Here, the projection matrices $\mathbf{W}_q^d, \mathbf{W}_k^d, \mathbf{W}_v^d \in \mathbb{R}^{h \times h}$ and $\mathbf{W}_q^p, \mathbf{W}_k^p, \mathbf{W}_v^p \in \mathbb{R}^{h \times h}$ are used to derive the queries, keys and values, respectively.

In summary, our CAN module combines multi-head, self-attention and cross-attention mechanisms to capture dependencies within individual sequences and between different sequences for a more nuanced understanding of interactions. In the results of Sections 4.3 and 4.5, we analyse and compare these two fusion strategies and different fusion scales in detail.

4 Experimental Setup and Results

4.1 Datasets and Baselines

Three public DTI datasets, namely BindingDB (Gilson et al., 2016), BioSNAP (Zitnik et al., 2018) and Human (Liu et al., 2015; Chen et al., 2020), are used for evaluation, where each dataset is split into training, validation, and test sets with a 7:1:2 ratio using two different splitting strategies: in-domain and cross-domain. For the in-domain split, the datasets are randomly divided. For the cross-domain setting, the datasets are split such that the drugs and targets in the test set do not overlap with those in the training set, making it a more challenging scenario where models must generalise to novel drug-target interactions. Since DTI is a binary classification task, we use AUROC (area under the receiver operating characteristic curve) (Bai et al., 2023; Huang et al., 2021) and AUPRC (area under the precision-call curve) (Nguyen et al., 2021) as the major metrics to evaluate models' performance. In Appendix A.10, we report other evaluation metrics, including F1-score, Sensitivity, Specificity, and Matthews Correlation Coefficient (MCC) to provide a more comprehensive assessment.

We compare FusionDTI with eight baseline models in the DTI prediction task. These models include two traditional machine learning methods such as SVM (Cortes and Vapnik, 1995) and Random Forest (RF) (Ho, 1995), as well as five deep learning methods including DeepConv-DTI (Lee et al., 2019), GraphDTA (Nguyen et al., 2021), MolTrans (Huang et al., 2021), DrugBAN (Bai et al., 2023) and SiamDTI (Zhang et al., 2024). The latter five models employ the same two-stage process whereby the drug and target features are initially extracted by specialised encoders before being integrated for prediction. In addition, we also include the BioT5 (Pei et al., 2023) model, which is a biomedical pre-trained language model that could directly predict the DTI. Further details regarding the datasets, baseline models, and the methodology for generating drug SELFIES and protein SA sequences are provided in Appendix A.3.

4.2 Evaluation of DTI Prediction

We start by comparing our FusionDTI model (FusionDTI-CAN and FusionDTI-BAN) with eight existing state-of-the-art baselines for DTI prediction on three widely used datasets. Table 1 reports the in-domain comparative results. In general, our FusionDTI-CAN model performs the best on all metrics across all three datasets. A key highlight from these results is the exceptional performance of FusionDTI-CAN on the BindingDB dataset, where FusionDTI-CAN demonstrates superior metrics

	BindingDB			Human		BioSNAP		
Method	AUROC	AUPRC	Accuracy	AUROC	AUPRC	AUROC	AUPRC	Accuracy
SVM	.939±.001	.928±.002	.825±.004	.940±.006	.920±.009	.862±.007	.864±.004	.777±.011
RF	.942±.011	.921±.016	.880±.012	.952±.011	.953±.010	.860±.005	.886±.005	.804±.005
DeepConv-DTI	.945±.002	.925±.005	.882±.007	.980±.002	.981±.002	.886±.006	.890±.006	.805±.009
GraphDTA	.951±.002	.934±.002	.888±.005	.981±.001	.982±.002	.887±.008	.890±.007	.800±.007
MolTrans	.952±.002	.936±.001	.887±.006	.980±.002	.978±.003	.895±.004	.897±.005	.825±.010
DrugBAN	.960±.001	.948±.002	.904±.004	.982±.002	.980±.003	.903±.005	.902±.004	.834±.008
SiamDTI	.961±.002	.945±.002	.890±.006	.970±.002	.969±.003	.912±.005	.910±.003	.855±.004
BioT5	.963±.001	.952±.001	.907±.003	.989±.001	.985±.002	.937±.001	.937±.004	.874±.001
FusionDTI-BAN	.975±.002	.976±.002	.933±.003	.984±.002	.984±.003	.923±.002	.921±.002	.856±.001
FusionDTI-CAN	**.989±.002**	**.990±.002**	**.961±.002**	**.991±.002**	**.989±.002**	**.951±.002**	**.952±.002**	**.889±.002**

Table 1: In-domain performance comparison of FusionDTI and the baselines on the BindingDB, Human and BioSNAP datasets (**Best**, Second Best).

	BindingDB			Human		BioSNAP		
Method	AUROC	AUPRC	Accuracy	AUROC	AUPRC	AUROC	AUPRC	Accuracy
SVM	.490±.015	.460±.001	.531±.009	.621±.036	.637±.009	.602±.005	.528±.005	.513±.011
RF	.493±.021	.468±.023	.535±.012	.642±.011	.663±.050	.590±.015	.568±.018	.499±.004
GraphDTA	.536±.015	.496±.029	.472±.009	.822±.009	.759±.006	.618±.005	.618±.008	.535±.024
DeepConv-DTI	.527±.038	.499±.035	.490±.027	.761±.016	.628±.022	.645±.022	.642±.032	.558±.025
MolTrans	.554±.024	.511±.025	.470±.004	.810±.021	.745±.034	.621±.015	.608±.022	.546±.032
DrugBAN	.604±.027	.570±.047	.509±.021	.833±.020	.760±.031	.685±.044	.713±.041	.565±.056
SiamDTI	.627±.027	.571±.024	.563±.033	**.863±.019**	.807±.040	.718±.055	.725±.054	.623±.070
BioT5	.651±.002	.653±.003	.621±.005	.856±.003	**.853±.003**	.720±.008	.718±.004	.715±.009
FusionDTI-BAN	.659±.002	.663±.002	.633±.003	.784±.002	.790±.003	.723±.002	.721±.002	.756±.001
FusionDTI-CAN	**.681±.005**	**.680±.012**	**.652±.005**	.801±.037	.803±.032	**.748±.021**	**.766±.017**	**.734±.012**

Table 2: Cross-domain performance comparison of FusionDTI and the baselines on the BindingDB, Human and BioSNAP datasets (**Best**, Second Best).

across the board: an AUROC of 0.989, an AUPRC of 0.990, and an accuracy of 96.1%. Note that the main difference between the FusionDTI-CAN model and others is the fusion strategy. Furthermore, despite FusionDTI-BAN and DrugBAN both utilising the same BAN module, FusionDTI-BAN consistently outperforms DrugBAN on all datasets.

However, in-domain classification using random splits holds limited practical significance. Thus, we also evaluate the more challenging cross-domain DTI prediction, where the training data and the test data contain distinct drugs and targets. This setting precludes the use of known drug or target features when making predictions on the test data. As shown in Table 2, the performance of all models is diminished compared to the in-domain setting due to the reduced availability of information. Nevertheless, the FusionDTI-CAN model demonstrates outstanding performance in cross-domain DTI prediction on the BindingDB and BioSNAP datasets, highlighting its robustness in predicting novel drug-target interactions. For instance, on the BindingDB dataset, FusionDTI-CAN achieves the highest metrics with an AUROC of 0.675 and an AUPRC of 0.676. This underscores the effectiveness of the model's fusion strategy in diverse and challenging scenarios. Similarly, despite sharing the BAN module, FusionDTI-BAN continues to outperform DrugBAN, further confirming the effectiveness of the FusionDTI framework in addressing cross-domain prediction challenges.

These findings highlight not only the substantial improvements of FusionDTI over existing approaches but also its effectiveness in capturing fine-grained information on DTI. The key to this success lies in FusionDTI's token-level fusion module, which enables the model to consider fine-grained interactions for each drug-target pair. This fine-grained interaction information aligns closely with biomedical pathways, where binding events often depend on the specific atoms or substructures involved in interactions with residues. Therefore, the model's ability to capture such fine-grained interactions significantly enhances its predictive performance for DTI.

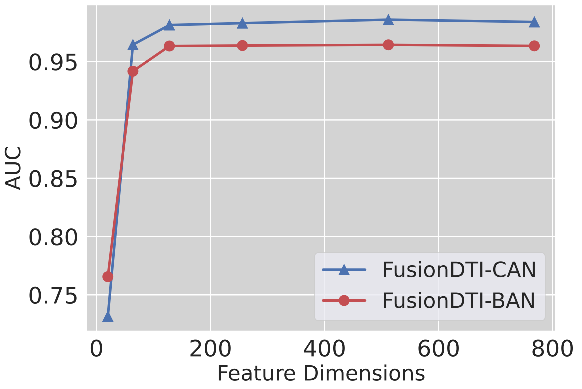

Figure 3: Performance comparison of two fusion strategies: BAN and CAN on the BindingDB.

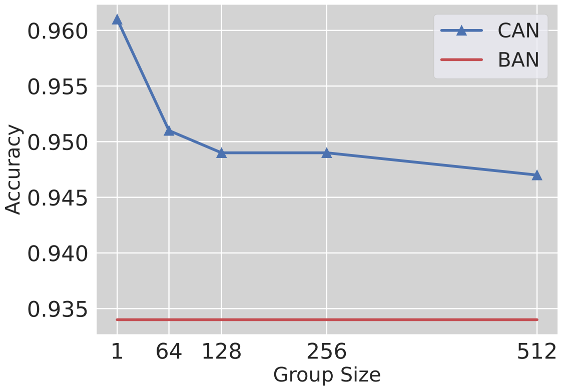

Figure 4: Performance evaluation of fusion scales on the BindingDB dataset.

CAN	AUC	AUPRC	Accuracy
×	0.954	0.963	0.894
✓	0.989	0.990	0.961

Table 3: Ablation study of the CAN module on the BindingDB dataset.

4.3 Comparison of the BAN and CAN

There are two fusion strategies available: BAN and CAN, thus determining which one works better is a key step for establishing FusionDTI's prediction effectiveness. We perform a fair comparison involving the same encoders, classifier and dataset. As shown in Figure 3, we compare BAN and CAN by employing two linear layers to adjust the feature dimensions of the drug and target representations. With the feature dimension increasing, the performance of FusionDTI-CAN continues to rise, while that of FusionDTI-BAN reaches a plateau. When the feature dimension is 512, both of the variants attain their peak positions with an AUC of 0.989 and 0.967, respectively. These results indicate that the CAN module seems to be better suited to the DTI prediction tasks and in capturing fine-grained interaction information. In contrast, BAN may not be able to fully capture fine-grained binding information between proteins and drugs, such as the specific interactions between the drug atoms and residues. Therefore, these findings suggest that the CAN strategy is more effective and adaptable to the complexities involved in DTI prediction, providing superior performance, especially as the feature dimension scales.

4.4 Ablation Study

The fine-grained interaction of drug and target representations is critical in DTI as it directly impacts the model's ability to infer potential binding sites. For FusionDTI, this interaction is facilitated by the CAN module, which markedly enhances the predictive accuracy by capturing the fine-grained interaction information between the drugs and targets. Table 3 demonstrates the impact of the CAN module on the prediction performance. When the fusion module is omitted, the model achieves an AUC of 0.954 and an accuracy of 0.894. Conversely, using the CAN module, there is a significant improvement, with the AUC increasing to 0.989 and the accuracy reaching 0.961. This highlights the effectiveness of the CAN module in improving the inference ability of FusionDTI. In Appendix A.7 and A.8, we further compare time-consuming and time complexity with baselines.

4.5 Analysis of Fusion Scales

In assessing fusion representations, it is critical to determine whether more fine-grained modelling enhances the predictive performance. Thus, we define a grouping function with the parameter **g** (Group size) for averaging tokens within each group before the CAN fusion module. The parameter **g**, representing the number of tokens per group, controls the granularity of the attention mechanism. Specifically, when **g** is set to 1, the fusion operates at the token level, where each token is considered independently. In contrast, when **g** is set to 512, the fusion occurs at a global level, considering the entire embedding as a single unit. We have the flexibility to control the fusion scale for the drug and protein representations, but the token length must be divisible by the group size. As shown in Figure 4, as the number of tokens per group increases from 1 to 512 (Maximum Token Length), the performance of the FusionDTI model declines

Drug-Target Interactions
EZL - 6QL2: **1**. sulfonamide oxygen - Leu198, Thr199 and Trp209; **2**. amino group - His94, His96, **His119** and Thr199; **3**. benzothiazole ring - Leu198, Thr200, **Tyr131**, Pro201 and **Gln92**; **4**. ethoxy group - **Gln135**;
9YA - 5W8L: **1**. amino group of sulfonamide - Asp140, Glu191; **2**. sulfonamide oxygen - Asp140, Ile141 and **Val139**; **3**. carboxylic acid oxygens - Arg168, His192, **Asp194** and Thr247; **4**. biphenyl rings - Arg105, Asn137 and **Pro138**; **5**. hydrophobic contact - **Ala237**, **Tyr238** and **Leu322**;
EJ4 - 4N6H: **1**. basic nitrogen of ligand - Asp128; **2**. hydrophobic pocket - Tyr308, Ile304 and **Tyr129**; **3**. water molecules - **Tyr129**, Met132, **Trp274**, Tyr308 and Lys214;

Table 4: FusionDTI predictions: **Bold** represents new predictions versus DrugBAN.

accordingly. This also aligns with the biomedical rules governing drug-protein interactions, where the principal factor influencing the binding is the interplay between the key atoms or substructures in the drug and primary residues in the protein. Furthermore, the CAN module outperforms BAN consistently at various scale settings, indicating that CAN better accesses the information between the drug and target. Consequently, this supports that the more detailed the interaction information obtained between the drugs and targets by the fusion module, the more beneficial it is for the enhancement of the model's prediction performance.

4.6 Case Study

A further strength of FusionDTI to enable explainability, which is critical for drug design efforts, is the visualisation of each token's contribution to the final prediction through cross-attention maps. To compare with the DrugBAN model, we examine three identical pairs of DTI from the Protein Data Bank (PDB) (Berman et al., 2007): (EZL - 6QL2 (Kazokaitė et al., 2019), 9YA - 5W8L (Rai et al., 2017) and EJ4 - 4N6H (Fenalti et al., 2014)), which are excluded from the training data. As shown in Table 4, our proposed model predicts more binding sites existing in the PDB (Berman et al., 2007) (in bold) by ranking the binding sites shown in the attention map. For instance, to predict the interaction of the drug EZL with the target 6QL2, our proposed model using BertViz (Vig, 2019) highlights potential binding sites as illustrated in Figure 5. Specifically, our CAN module is effective in capturing fine-grained binding information at the token level, as we have successfully predicted the novel binding between Gln92 and the benzothiazole ring (Di Fiore et al., 2008). In particular, we address the lack of structural information on protein sequences by employing the SA vocabulary, which matches each residue to a corresponding 3D feature via Foldseek (Van Kempen et al., 2024). This study highlights the effectiveness of FusionDTI in enhancing performance on the DTI task, thereby supporting more targeted and efficient drug development efforts. In Appendix A.9, we further investigate ten DTI pairs in non-small cell lung cancer (NSCLC) from PDB (Waliany et al., 2025), highlighting predicted binding residues.

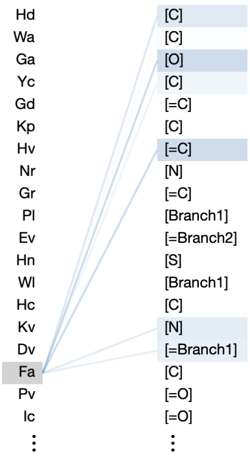

Figure 5: EZL - 6QL2: Fine-grained interactions via attention visualization.

5 Conclusions

With the rapid increase of new diseases and the urgent need for innovative drugs, it is critical to capture fine-grained interactions, since the binding of specific drug atoms to the main amino acids is key to the DTI task. Despite some achievements, fine-grained interaction information is not effectively captured. To address this challenge, we introduce FusionDTI uses token-level fusion to effectively obtain fine-grained interaction information. Through experiments on three well-known datasets, we demonstrate that our proposed FusionDTI model outperforms eight state-of-the-art baselines, particularly in the more realistic cross-domain scenario. Additionally, we show that the attention weights of the token-level fusion module can highlight potential binding sites, providing a certain level of explainability.

Limitations

Even if our proposed model identifies potentially useful DTI, these predictions need to be validated by wet experiments, a time-consuming and expensive process. We have shown that FusionDTI is effective and efficient in screening for possible DTI in large-scale data as well as in locating potential binding sites in the process of drug design. However, it is not directly applicable to human medical therapy and other biomedical interactions because it lacks clinical validation and regulatory approval for medical use.

References

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2022. Chemberta-2: Towards chemical foundation models. arXiv preprint arXiv:2209.01712.

Heba Askr, Enas Elgeldawi, Heba Aboul Ella, Yaseen AMM Elshaier, Mamdouh M Gomaa, and Aboul Ella Hassanien. 2023. Deep learning in drug discovery: an integrative review and future challenges. Artificial Intelligence Review, 56(7):5975–6037.

Peizhen Bai, Filip Miljković, Bino John, and Haiping Lu. 2023. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. Nature Machine Intelligence, 5(2):126–136.

Helen Berman, Kim Henrick, Haruki Nakamura, and John L Markley. 2007. The worldwide protein data bank (wwpdb): ensuring a single, uniform archive of pdb data. Nucleic acids research, 35(suppl_1):D301–D303.

Ashraf Brik and Chi-Huey Wong. 2003. Hiv-1 protease: mechanism and drug discovery. Organic & biomolecular chemistry, 1(1):5–14.

Dong-Sheng Cao, Qing-Song Xu, and Yi-Zeng Liang. 2013. propy: a tool to generate various modes of chou's pseaac. Bioinformatics, 29(7):960–962.

Ashish Chandwani and Jonathan Shuter. 2008. Lopinavir/ritonavir in the treatment of hiv-1 infection: a review. Therapeutics and clinical risk management, 4(5):1023–1033.

Lifan Chen, Xiaoqin Tan, Dingyan Wang, Feisheng Zhong, Xiaohong Liu, Tianbiao Yang, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. 2020. Transformercpi: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. Bioinformatics, 36(16):4406–4414.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. Machine learning, 20:273–297.

Anna Di Fiore, Carlo Pedone, Jochen Antel, Harald Waldeck, Andreas Witte, Michael Wurl, Andrea Scozzafava, Claudiu T Supuran, and Giuseppina De Simone. 2008. Carbonic anhydrase inhibitors: the x-ray crystal structure of ethoxzolamide complexed to human isoform ii reveals the importance of thr200 and gln92 for obtaining tight-binding inhibitors. Bioorganic & medicinal chemistry letters, 18(8):2669–2674.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Wang Yu, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. 2021. Prottrans: Towards cracking the language of lifes code through self-supervised deep learning and high performance computing. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1–1.

Gustavo Fenalti, Patrick M Giguere, Vsevolod Katritch, Xi-Ping Huang, Aaron A Thompson, Vadim Cherezov, Bryan L Roth, and Raymond C Stevens. 2014. Molecular control of δ-opioid receptor signalling. Nature, 506(7487):191–196.

Michael K Gilson, Tiqing Liu, Michael Baitaluk, George Nicola, Linda Hwang, and Jenny Chong. 2016. Bindingdb in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic acids research, 44(D1):D1045–D1053.

Yichen Gong, Heng Luo, and Jian Zhang. 2018. Natural language inference over interaction space. International Conference on Learning Representations.

Mercedes Herrera-Juárez, Cristina Serrano-Gómez, Helena Bote-de Cabo, and Luis Paz-Ares. 2023. Targeted therapy for lung cancer: Beyond egfr and alk. Cancer, 129(12):1803–1820.

Tin Kam Ho. 1995. Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition, volume 1, pages 278–282. IEEE.

Kexin Huang, Cao Xiao, Lucas M Glass, and Jimeng Sun. 2021. Moltrans: molecular interaction transformer for drug–target interaction prediction. Bioinformatics, 37(6):830–836.

Justina Kazokaitė, Visvaldas Kairys, Joana Smirnovienė, Alexey Smirnov, Elena Manakova, Martti Tolvanen, Seppo Parkkila, and Daumantas Matulis. 2019. Engineered carbonic anhydrase vi-mimic enzyme switched the structure and affinities of inhibitors. Scientific reports, 9(1):12710.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. Advances in neural information processing systems, 31.

Mario Krenn, Qianxiang Ai, Senja Barthel, Nessa Carson, Angelo Frei, Nathan C Frey, Pascal Friederich, Théophile Gaudin, Alberto Alexander Gayle, Kevin Maik Jablonka, and 1 others. 2022. Selfies and the future of molecular string representations. Patterns, 3(10).

Ingoo Lee, Jongsoo Keum, and Hojung Nam. 2019. Deepconv-dti: Prediction of drug-target interactions via deep learning with convolution on protein sequences. PLoS computational biology, 15(6):e1007129.

Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. 2021. Selfdoc: Self-supervised document representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5652–5660.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, and 1 others. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637):1123–1130.

Fangyu Liu, Yunlong Jiao, Jordan Massiah, Emine Yilmaz, and Serhii Havrylov. 2021. Trans-encoder: Unsupervised sentence-pair modelling through self-and mutual-distillations. In International Conference on Learning Representations.

Hui Liu, Jianjiang Sun, Jihong Guan, Jie Zheng, and Shuigeng Zhou. 2015. Improving compound–protein interaction prediction by building up highly credible negative samples. Bioinformatics, 31(12):i221–i229.

Thin Nguyen, Hang Le, Thomas P Quinn, Tri Nguyen, Thuc Duy Le, and Svetha Venkatesh. 2021. Graphdta: predicting drug–target binding affinity with graph neural networks. Bioinformatics, 37(8):1140–1147.

Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. 2023. BioT5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 1102–1123, Singapore. Association for Computational Linguistics.

Lihong Peng, Xin Liu, Long Yang, Longlong Liu, Zongzheng Bai, Min Chen, Xu Lu, and Libo Nie. 2024. Bindti: A bi-directional intention network for drug-target interaction identification based on attention mechanisms. IEEE Journal of Biomedical and Health Informatics.

Ganesha Rai, Kyle R Brimacombe, Bryan T Mott, Daniel J Urban, Xin Hu, Shyh-Ming Yang, Tobie D Lee, Dorian M Cheff, Jennifer Kouznetsova, Gloria A Benavides, and 1 others. 2017. Discovery and optimization of potent, cell-active pyrazole-based inhibitors of lactate dehydrogenase (ldh). Journal of medicinal chemistry, 60(22):9184–9204.

David Rogers and Mathew Hahn. 2010. Extended-connectivity fingerprints. Journal of chemical information and modeling, 50(5):742–754.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. 2020. Self-supervised graph transformer on large-scale molecular data. Advances in neural information processing systems, 33:12559–12571.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel

Das. 2022. Large-scale chemical language representations capture molecular structure and properties. Nature Machine Intelligence, 4(12):1256–1264.

Monica Schenone, Vlado Dančík, Bridget K Wagner, and Paul A Clemons. 2013. Target identification and mechanism of action in chemical biology and drug discovery. Nature chemical biology, 9(4):232–240.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. 2023. Saprot: protein language modeling with structure-aware vocabulary. Advances in neural information processing systems, pages 2023–10.

Jin Su, Zhikai Li, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, Dacheng Ma, The OPMC, Sergey Ovchinnikov, and Fajie Yuan. 2024. Saprothub: Making protein modeling accessible to all biologists. bioRxiv, pages 2024–05.

Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. 2024. Fast and accurate protein structure search with foldseek. Nature Biotechnology, 42(2):243–246.

Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, and 1 others. 2022. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic acids research, 50(D1):D439–D444.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems, 30.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Sarah Waliany, Jessica J Lin, and Justin F Gainor. 2025. Evolution of first versus next-line targeted therapies for metastatic non-small cell lung cancer. Trends in Cancer.

David Weininger. 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. Journal of chemical information and computer sciences, 28(1):31–36.

David Weininger, Arthur Weininger, and Joseph L Weininger. 1989. Smiles. 2. algorithm for generation of unique smiles notation. Journal of chemical information and computer sciences, 29(2):97–101.

David S Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. 2008. Drugbank: a knowledgebase for drugs, drug actions and drug targets. Nucleic acids research, 36(suppl_1):D901–D906.

Yifan Wu, Min Gao, Min Zeng, Jie Zhang, and Min Li. 2022. Bridgedpi: a novel graph neural network for predicting drug–protein interactions. Bioinformatics, 38(9):2571–2578.

Minghao Xu, Xinyu Yuan, Santiago Miret, and Jian Tang. 2023. Protst: Multi-modality learning of protein sequences and biomedical texts. In International Conference on Machine Learning, pages 38749–38767. PMLR.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. 2021. Do transformers really perform badly for graph representation? Advances in neural information processing systems, 34:28877–28888.

Atakan Yüksel, Erva Ulusoy, Atabey Ünlü, and Tunca Doğan. 2023. Selformer: molecular representation learning via selfies language models. Machine Learning: Science and Technology, 4(2):025035.

Xiaoting Zeng, Weilin Chen, and Baiying Lei. 2024. Cat-dti: cross-attention and transformer network with domain adaptation for drug-target interaction prediction. BMC bioinformatics, 25(1):141.

Dataset	Drugs	Proteins	Interactions
BindingDB	14,643	2,623	49,199
BioSNAP	4,510	2,181	27,464
Human	2,726	2,001	6,728

Table 5: Dataset Statistics.

Hongzhi Zhang, Xiuwen Gong, Shirui Pan, Jia Wu, Bo Du, and Wenbin Hu. 2024. A cross-field fusion strategy for drug-target interaction prediction. arXiv preprint arXiv:2405.14545.

Marinka Zitnik, Rok Sosic, and Jure Leskovec. 2018. Biosnap datasets: Stanford biomedical network dataset collection. Note: http://snap. stanford. edu/biodata Cited by, 5(1).

A Appendix

A.1 Hyperparameter of FusionDTI

FusionDTI is implemented in Python 3.8 and the PyTorch framework (1.12.1)[2]. The computing device we use is the NVIDIA GeForce RTX 3090. In the "Experimental Setup and Results" section, we only present experiment results based on the BindingDB dataset, as the performance trends are identical to the BioSNAP dataset and the Human dataset. Table 6 shows the parameters of the FusionDTI model and Table 7 lists the notations used in this paper with descriptions.

A.2 Dataset Sources

All the data used in this paper are from public sources. The statistics of the experimental datasets are presented in Table 5.

1. The BindingDB (Gilson et al., 2016) dataset is a web-accessible database of experimentally validated binding affinities, focusing primarily on the interactions of small drug-like molecules and proteins. The BindingDB source is found at `https://www.bindingdb.org/bind/index.jsp`.

2. The BioSNAP (Zitnik et al., 2018) dataset is created from the DrugBank database (Wishart et al., 2008). It is a balanced dataset with validated positive interactions and an equal number of negative samples randomly obtained from unseen pairs. The BioSNAP source is found at `https://github.com/kexinhuang12345/MolTrans`.

3. The Human (Liu et al., 2015; Chen et al., 2020) dataset includes highly credible negative samples. The balanced version of the Human dataset contains the same number of positive and negative samples. The Human source is found at `https://github.com/lifanchen-simm/transformerCPI`.

A.3 How to Obtain the Structure-aware (SA) Sequence of a Protein and the SELFIES of a Drug?

To obtain the SA sequence of a protein, the first step is to obtain Uniprot IDs from the UniProt website using information such as the amino acid sequences or protein names, and then save these IDs in a comma-delimited text file. Subsequently, we use the UniProt IDs to fetch the relevant 3D structure file (.cif) from AlphafoldDB (Varadi et al., 2022) using Foldseek. The SA vocabulary of the protein can then be generated from this 3D structure file.

For drugs, the SELFIES could be derived from SMILES strings. This conversion requires specific Python packages, and upon installation, the SELFIES strings can be generated through appropriate scripts. Please refer to our submission file for detailed procedures, including the necessary code.

Notably, our submission of supplementary material contains step-by-step descriptions and code for generating the SA sequences and SELFIES.

A.4 Baselines

We compare the performance of FusionDTI with the following eight models on the DTI task.

1. Support Vector Machine (Cortes and Vapnik, 1995) on the concatenated fingerprint ECFP4 (Rogers and Hahn, 2010) (extended connectivity fingerprint, up to four bonds) and PSC (Cao et al., 2013) (pseudo-amino acid composition) features.

2. Random Forest (Ho, 1995) on the concatenated fingerprint ECFP4 and PSC features.

3. DeepConv-DTI (Lee et al., 2019) uses a fully connected neural network to encode the ECFP4 drug fingerprint and a CNN along with a global max-pooling layer to extract features from the protein sequences. Then the drug and protein features are concatenated and fed into a fully connected neural network for the final prediction.

[2] `https://pytorch.org/`

Module	Hyperparameter	Value
Mini-batch	Batch size	64 (options: 64, 128)
Drug Encoder	PLM	HUBioDataLab/SELFormer
Protein Encoder	PLM	westlake-repl/SaProt_650M_AF2
BAN	Heads of bilinear attention	3
	Bilinear embedding size	512 (options: 32, 64, 128, 256, 512, 768)
	Sum pooling window size	2
CAN	Attention heads	8
	Hidden dimension	512 (options: 32, 64, 128, 256, 512, 768)
	Integration strategies	Mean pooling (options: Mean pooling, CLS)
	Group size	1 (options: from 1 to 512)
MLP	Hidden layer sizes	(1024, 512, 256)
	Activation	Relu (options: Tanh, Relu)
	Solver	AdamW (options: AdamW, Adam, RMSprop, Adadelta, LBFGS)
	Learning rate scheduler	CosineAnnealingLR (options: CosineAnnealingLR, StepLR, ExponentialLR)
	Initial learning rate	1e-4 (options: from 1e-3 to 1e-6)
	Maximum epoch	200

Table 6: Configuration Parameters

Notations	Description
$\mathbf{D}$	Drug feature
$\mathbf{P}$	Target feature
$\mathbf{q} \in \mathbb{R}^K$	weight vector for bilinear transformation
$Att \in \mathbb{R}^{\rho \times \phi}$	Bilinear attention maps in BAN
$\mathbf{U} \in \mathbb{R}^{N \times K}$	Transformation matrix for drug features
$\mathbf{V} \in \mathbb{R}^{M \times K}$	Transformation matrix for target features
$\mathbf{g}$	The number of tokens per group
$\mathbf{D}^* \in \mathbb{R}^{m \times h}$	Fused drug representations in token-level interaction
$\mathbf{P}^* \in \mathbb{R}^{n \times h}$	Fused target representations in token-level interaction
$\mathbf{Q}_d, \mathbf{K}_d, \mathbf{V}_d \in \mathbb{R}^{m \times h}$	Queries, keys, and values for the drug in token-level interaction
$\mathbf{Q}_p, \mathbf{K}_p, \mathbf{V}_p \in \mathbb{R}^{n \times h}$	Queries, keys, and values for target in token-level interaction
$\mathbf{W}_q^d, \mathbf{W}_k^d, \mathbf{W}_v^d \in \mathbb{R}^{H \times h}$	Projection matrices for drug queries, keys, and values
$\mathbf{W}_q^p, \mathbf{W}_k^p, \mathbf{W}_v^p \in \mathbb{R}^{h \times h}$	Projection matrices for target queries, keys, and values
$\mathbf{F}$	drug-target joint representation
$p \in [0, 1]$	output interaction probability
H	Number of attention heads in token-level interaction
m, n	Sequence lengths for drug and protein respectively
h	Hidden dimension in token-level interaction

Table 7: Notations and Descriptions

4. GraphDTA (Nguyen et al., 2021) uses GNN for the encoding of drug molecular graphs, and a CNN is used for the encoding of the protein sequences. The derived vectors of the drug and protein representations are directly concatenated for interaction prediction.

5. MolTrans (Huang et al., 2021) uses a transformer architecture to encode the drugs and proteins. Then a CNN-based fusion module is adapted to capture DTI interactions.

6. DrugBAN (Bai et al., 2023) use a Graph Convolution Network and 1D CNN to encode

the drug and protein sequences. Then a bilinear attention network (Kim et al., 2018) is adopted to learn pairwise interactions between the drug and protein. The resulting joint representation is decoded by a fully connected neural network.

7. BioT5 (Pei et al., 2023) is a cross-modeling model in biology with chemical knowledge and natural language associations.

8. SiamDTI (Zhang et al., 2024) is a double-channel network structure to acquire local and global protein information for cross-field supervised learning.

A.5 Ablation Study

In Table 8, we compare the performance of two aggregation strategies within the CAN module. The pooling strategy outperforms the CLS-based aggregation, achieving an AUC and AUPRC of 0.989 and 0.990, respectively. This comparison highlights the superior effectiveness of the pooling in aggregating contextual information. Thus, the integration of a CAN module, particularly employing a pooling aggregation strategy, is shown to be essential for making confident and accurate predictions.

A.6 Evaluation of PLMs Encoding

The protein encoder and drug encoder are fundamental for the token-level fusion of representations, as these encoders are responsible for generating fine-grained representations to better explore interaction information. Our proposed model employs two PLMs encoding two biomedical entities: the drug and protein, respectively. In terms of the protein encoders, Figure 7 compares the the performance of the two protein encoders (SaProt (Su et al., 2023) and ESM-2 (Lin et al., 2023)) in combination with three different drug encoders: ChemBERTa-2 (Ahmad et al., 2022), SELFormer (Yüksel et al., 2023) and MoLFormer (Ross et al., 2022). From the figure, we find that SaProt consistently outperforms ESM-2 when combined with all three drug encoders. As can be seen in Figure 8, SELFormer achieves the best performance in encoding the drug sequences among the three advanced drug encoders. Notably, the top-performing combination is SaProt and SELFormer, hence our proposed FusionDTI uses them as drug and protein encoders.

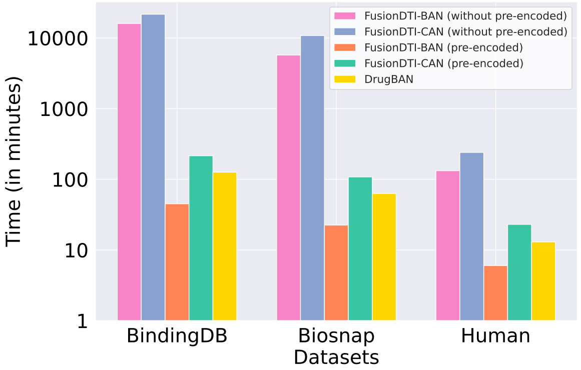

Figure 6: Time comparison on the BindingDB, Human and BioSNAP datasets.

Aggregation	AUC	AUPRC	Accuracy
CLS	0.982	0.983	0.956
Pooling	0.989	0.990	0.961

Table 8: Comparison of aggregation strategies for Fusion-CAN on the BindingDB dataset.

A.7 Efficiency Analysis

Efficiency in computational models is crucial, particularly when handling large-scale and extensive datasets in drug discovery. Our proposed model stores drug representations and target representations in memory for later online training. As evidenced by Figure 6, FusionDTI-CAN and FusionDTI-BAN with pre-encoded representations process the BindingDB dataset much faster than the non-pre-coded models, approximately 45 minutes and 220 minutes, respectively. This stark difference highlights the advantage of pre-encoded, which eliminates the need for real-time data processing and accelerates the overall throughput. While FusionDTI-BAN and DrugBAN have the same fusion module, the pre-encoded FusionDTI-BAN runs faster and predicts more accurately, as shown in Table 1. In addition, FusionDTI-BAN runs faster than FusionDTI-CAN, indicating that the BAN fusion module is more efficient. Ultimately, FusionDTI-BAN with pre-encoded data stands out as a highly efficient approach, offering substantial benefits in scenarios where exists large-scale data.

A.8 Time Complexity Analysis

The feature dimensions of the representations generated by different PLM encoders are fixed, but the size of the feature dimensions may not be the same. Therefore, in order to fuse protein and drug representations, we use two linear layers to keep

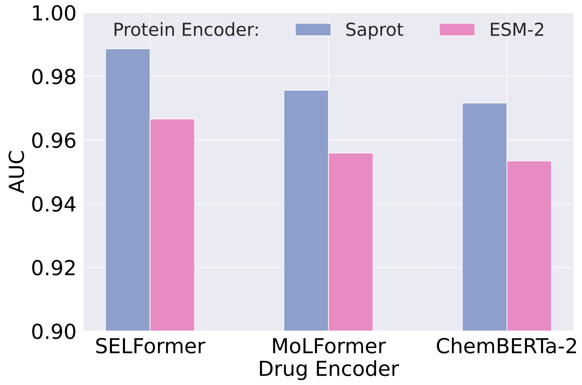

Figure 7: Performance comparison of protein encoders on the BindingDB dataset.

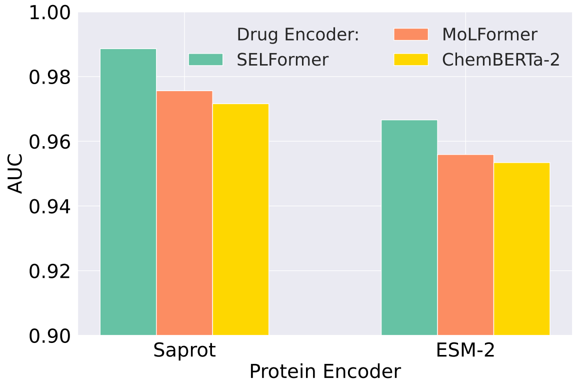

Figure 8: Performance comparison of drug encoders on the BindingDB dataset.

Fusion module	Complexity (O)	Parameters
BAN	$O(\rho \cdot \phi \cdot K)$	790k
CAN	$O(m \cdot n \cdot h)$	1572k

Table 9: Time complexity and parameters comparison of BAN and CAN.

the representations’ feature dimension equal to the token length (512).

The time complexity of BAN depends on the computation of bilinear interaction maps. The bilinear attention involves a Hadamard product and further matrix operations as given in Equation (2). The computation of $U^T P$ and $V^T D$ requires $O(N \cdot \rho \cdot K)$ and $O(M \cdot \phi \cdot K)$ operations, respectively. Here, K denotes the dimensionality of the transformation, which is the rank of the feature space to which the protein and drug features are projected. When the token length is equal to the feature dimension and the dimensions of transformation are two times either, the overall time complexity is $O(\rho \cdot \phi \cdot K)$.

For the token-level interaction in the DTI task, the time complexity is also markedly influenced by the attention mechanisms. It also satisfies the condition that the token length is equal to the feature dimension of the drug and protein. With multi-head attention heads ($H = 8$), the complexity for computing the queries, keys, and values in the Equation (6) and (7), as well as the softmax attention weights, is given by $O(H \cdot n \cdot m \cdot h)$, where $mandn$ represents the token lengths for the drug and protein, respectively, and h is the hidden dimension. Since each head contributes its own set of computations and the attention mechanism operates over all tokens, the $m \cdot n$ term (stemming from the softmax operation across the token length) becomes significant. This leads to a total time complexity of $O(m \cdot n \cdot h)$ per batch for the attention mechanism.

From the above analysis of the time complexity of the two fusion strategies, the time complexity of CAN is lower than BAN in the case of the same input protein and drug features. BAN is markedly affected by the transformation dimension K. When the K is larger than the token and feature dimension, the time complexity of BAN is higher than CAN. However, we observe that the number of parameters in BAN is smaller than that of CAN via the Pytroch package, as shown in Table 9.

A.9 Case Study

The top three predictions (PDB ID: 6QL2 (Kazokaitė et al., 2019), 5W8L (Rai et al., 2017) and 4N6H (Fenalti et al., 2014)) of the co-crystalized ligands are derived from Protein Data Bank (PDB) (Berman et al., 2007). Following the setup of the DrugBAN case study, we only choose X-ray structures with a resolution greater than 2.5 Å corresponding to human proteins. In addition, the co-crystalized ligands are required to have pIC$_{50}$ $\leq$ 100 nM and are not part of the training dataset.

To further DTI in non-small cell lung cancer (NSCLC), we identify ten additional drug-protein pairs from PDB. The selected targets—Epidermal Growth Factor Receptor (EGFR), Anaplastic Lymphoma Kinase (ALK), and ROS1—are well-established oncogenic drivers in NSCLC (Waliany et al., 2025). The corresponding inhibitors, including Erlotinib, Gefitinib, Osimertinib, Crizotinib, and Lorlatinib, exhibit high binding affinities (Herrera-Juárez et al., 2023). Table 10 presents the predicted binding residues for these interac-

Drug-Target (Ligand - PDB ID)	Predicted Binding Residues
VGH - 2YFX	**Glu113**, **Val46**, **Gly117**, **Met115**, **Asp186**, Arg125, Lys225, Gln50, Ala190, Pro319
C6F - 6JQR	**Tyr126**, **Asp209**, **Ala72**, **Glu208**, **Glu197**, Leu219, Pro163, Gln97, Val225, His151
5P8 - 4CLI	**GLu113**, **Leu172**, **Gly118**, **Ala64**, **Asp186**, Ala150, Ile99, Pro290, Ala312, Glu316
0WM - 4G5J	**His296**, **Pro102**, **Pro156**, **Met295**, **Asn116**, Ser92, Thr217, Lys237, His143, Trp188
YY3 - 6LUD	**Phe102**, **Leu151**, **Met1100**, **Lys52**, **Glu111**, Ile22, Pro60, Ala129, Val141, Gly42
AQ4 - 1M17	**Leu155**, **Leu99**, **Met104**, **Phe106**, **Thr165**, Asp111, Lys171, Trp209, Ala61, Asp280
YMX - 5FTO	**Asn162**, **Gly110**, **Phe35**, **Glu118**, **Val38**, His155, ALa197, Met46, Leu112, Asp280
1C9 - 4I23	**Ala50**, **Leu95**, **Met100**, **Pro101**, **Glu69**, Thr247, Tyr120, His177, Pro221, Val49
VGH - 2XP2	**Leu172**, **Gly185**, **Ala116**, **Lys66**, **Asp119**, Pro58, Met82, Pro131, Ala167, Val27
EMH - 3AOX	**Glu143**, **Leu55**, **Gly56**, **Val113**, **Met132**, Glu91, Leu157, Val44, Ala59, Ile166

Table 10: Predicted binding sites for DTI in NSCLC. **Bold** residues are supported by the PDB database, while others remain unverified.

tions, with bolded residues supported by experimental PDB data, while others remain unverified.

A.10 Performance Comparison

Tables 11 and 12 provide a detailed performance evaluation of FusionDTI and baseline models across both in-domain and cross-domain settings. To ensure a comprehensive assessment, we report multiple evaluation metrics, including AUROC and AUPRC as primary indicators, alongside F1-score, Sensitivity, Specificity, and Matthews Correlation Coefficient (MCC). These additional metrics offer deeper insights into model performance across different classification aspects.

Model	AUC	AUPR	Accuracy	F1	Sensitivity	Specificity	MCC
BindingDB							
SVM	0.939±0.001	0.928±0.002	0.825±0.004	0.821±0.004	0.810±0.010	0.840±0.007	0.700±0.012
RF	0.942±0.011	0.921±0.016	0.880±0.012	0.875±0.012	0.870±0.015	0.890±0.010	0.815±0.009
DeepConv-DTI	0.945±0.002	0.925±0.005	0.882±0.007	0.878±0.008	0.870±0.011	0.885±0.010	0.818±0.013
GraphDTA	0.951±0.002	0.934±0.002	0.888±0.005	0.884±0.005	0.880±0.006	0.890±0.004	0.825±0.008
MolTrans	0.952±0.002	0.936±0.001	0.887±0.006	0.882±0.006	0.875±0.009	0.890±0.007	0.820±0.010
DrugBAN	0.960±0.001	0.948±0.002	0.906±0.004	0.901±0.004	0.900±0.008	0.908±0.004	0.872±0.005
SiamDTI	0.961±0.002	0.945±0.002	0.890±0.006	0.886±0.006	0.880±0.007	0.895±0.005	0.830±0.006
BioT5	0.963±0.001	0.952±0.001	0.907±0.003	0.905±0.003	0.900±0.004	0.910±0.003	0.850±0.005
FusionDTI-BAN	0.975±0.002	0.976±0.002	0.933±0.003	0.934±0.002	0.932±0.004	0.935±0.003	0.900±0.003
FusionDTI-CAN	**0.989±0.002**	**0.990±0.002**	**0.961±0.002**	**0.963±0.012**	**0.954±0.003**	**0.955±0.012**	**0.925±0.023**
BioSNAP							
SVM	0.862±0.007	0.864±0.004	0.777±0.011	0.773±0.011	0.760±0.015	0.780±0.008	0.690±0.013
RF	0.860±0.005	0.886±0.005	0.804±0.005	0.800±0.005	0.795±0.008	0.810±0.007	0.715±0.006
DeepConv-DTI	0.886±0.006	0.890±0.006	0.805±0.009	0.801±0.009	0.800±0.013	0.810±0.010	0.718±0.012
GraphDTA	0.887±0.008	0.890±0.007	0.800±0.007	0.796±0.007	0.790±0.010	0.810±0.009	0.712±0.009
MolTrans	0.895±0.004	0.897±0.005	0.825±0.010	0.820±0.010	0.815±0.013	0.830±0.012	0.730±0.011
DrugBAN	0.903±0.005	0.902±0.004	0.834±0.008	0.830±0.009	0.820±0.021	0.847±0.010	0.719±0.007
SiamDTI	0.912±0.005	0.910±0.003	0.855±0.004	0.852±0.004	0.850±0.006	0.860±0.004	0.740±0.006
BioT5	0.937±0.001	0.937±0.004	0.874±0.001	0.870±0.001	0.865±0.002	0.880±0.003	0.765±0.004
FusionDTI-BAN	0.923±0.002	0.921±0.002	0.856±0.001	0.857±0.001	0.854±0.002	0.858±0.002	0.724±0.001
FusionDTI-CAN	**0.951±0.002**	**0.951±0.002**	**0.889±0.002**	**0.890±0.002**	**0.888±0.003**	**0.891±0.002**	**0.778±0.002**
Human							
SVM	0.940±0.006	0.920±0.009	0.895±0.010	0.892±0.011	0.880±0.015	0.910±0.009	0.800±0.012
RF	0.952±0.011	0.953±0.010	0.920±0.012	0.915±0.013	0.910±0.017	0.930±0.014	0.820±0.009
DeepConv-DTI	0.980±0.002	0.981±0.002	0.927±0.007	0.923±0.006	0.920±0.009	0.930±0.008	0.860±0.010
GraphDTA	0.981±0.001	0.982±0.002	0.930±0.008	0.925±0.008	0.920±0.011	0.935±0.009	0.870±0.009
MolTrans	0.980±0.002	0.978±0.003	0.925±0.011	0.920±0.012	0.915±0.016	0.930±0.013	0.855±0.010
DrugBAN	0.982±0.002	0.980±0.003	0.930±0.004	0.903±0.003	0.900±0.005	0.908±0.004	0.810±0.004
SiamDTI	0.970±0.002	0.969±0.003	0.920±0.006	0.915±0.006	0.910±0.008	0.925±0.007	0.840±0.009
BioT5	0.989±0.001	0.985±0.002	0.939±0.008	0.937±0.004	0.929±0.010	0.941±0.004	0.892±0.006
FusionDTI-BAN	0.984±0.002	0.984±0.003	0.938±0.003	0.934±0.002	0.927±0.004	0.931±0.003	0.870±0.003
FusionDTI-CAN	**0.991±0.002**	**0.989±0.002**	**0.947±0.002**	**0.948±0.002**	**0.955±0.033**	**0.950±0.031**	**0.905±0.045**

Table 11: In-domain performance comparison of FusionDTI and the baselines on the BindingDB, Human and BioSNAP datasets (**Best**, Second Best).

Model	AUC	AUPR	Accuracy	F1	Sensitivity	Specificity	MCC
			BindingDB				
SVM	0.490±0.015	0.460±0.001	0.531±0.009	0.521±0.010	0.508±0.015	0.548±0.011	0.150±0.012
RF	0.493±0.021	0.468±0.023	0.535±0.012	0.525±0.013	0.512±0.020	0.550±0.014	0.162±0.015
GraphDTA	0.536±0.015	0.496±0.029	0.472±0.009	0.462±0.008	0.460±0.014	0.478±0.011	0.100±0.012
DeepConv-DTI	0.527±0.038	0.499±0.035	0.490±0.027	0.480±0.026	0.475±0.030	0.495±0.023	0.115±0.020
MolTrans	0.554±0.024	0.511±0.025	0.470±0.004	0.460±0.005	0.455±0.008	0.478±0.007	0.105±0.008
DrugBAN	0.604±0.027	0.570±0.047	0.509±0.021	0.582±0.030	0.565±0.022	0.580±0.025	0.187±0.031
SiamDTI	0.627±0.027	0.571±0.024	0.563±0.033	0.550±0.032	0.540±0.036	0.580±0.028	0.190±0.030
BioT5	0.651±0.002	0.653±0.003	0.621±0.005	0.608±0.004	0.600±0.006	0.635±0.005	0.220±0.007
FusionDTI-BAN	0.659±0.002	0.663±0.002	0.633±0.003	0.587±0.002	0.603±0.003	0.589±0.002	0.276±0.003
FusionDTI-CAN	**0.681±0.005**	**0.680±0.012**	**0.652±0.005**	**0.601±0.005**	**0.628±0.006**	**0.692±0.005**	**0.302±0.005**
			BioSNAP				
SVM	0.602±0.005	0.528±0.005	0.513±0.011	0.502±0.012	0.490±0.014	0.523±0.013	0.150±0.010
RF	0.590±0.015	0.568±0.018	0.499±0.004	0.488±0.005	0.478±0.008	0.513±0.007	0.135±0.008
GraphDTA	0.618±0.005	0.618±0.008	0.535±0.024	0.528±0.023	0.520±0.027	0.550±0.020	0.170±0.025
DeepConv-DTI	0.645±0.022	0.642±0.032	0.558±0.025	0.550±0.024	0.543±0.030	0.573±0.027	0.200±0.028
MolTrans	0.621±0.015	0.608±0.022	0.546±0.032	0.538±0.031	0.530±0.035	0.563±0.033	0.185±0.034
DrugBAN	0.685±0.004	0.713±0.005	0.692±0.006	0.587±0.005	0.522±0.011	0.690±0.012	0.219±0.017
SiamDTI	0.718±0.005	0.725±0.005	0.623±0.007	0.610±0.006	0.600±0.007	0.675±0.006	0.240±0.008
BioT5	0.720±0.008	0.718±0.004	0.715±0.009	0.590±0.010	0.510±0.012	0.710±0.010	0.250±0.011
FusionDTI-BAN	0.723±0.002	0.721±0.002	0.726±0.001	0.597±0.001	0.504±0.012	0.713±0.011	0.254±0.010
FusionDTI-CAN	**0.748±0.021**	**0.766±0.017**	**0.734±0.012**	**0.602±0.012**	**0.531±0.013**	**0.736±0.012**	**0.268±0.011**
			Human				
SVM	0.621±0.036	0.637±0.009	0.533±0.011	0.525±0.012	0.520±0.015	0.546±0.010	0.175±0.011
RF	0.642±0.011	0.663±0.050	0.543±0.014	0.535±0.015	0.530±0.018	0.556±0.013	0.184±0.012
GraphDTA	0.822±0.009	0.759±0.006	0.709±0.016	0.705±0.017	0.702±0.020	0.713±0.015	0.198±0.017
DeepConv-DTI	0.761±0.016	0.628±0.022	0.711±0.030	0.704±0.031	0.704±0.035	0.728±0.027	0.203±0.030
MolTrans	0.810±0.021	0.745±0.034	0.713±0.032	0.725±0.033	0.720±0.037	0.740±0.031	0.215±0.032
DrugBAN	0.833±0.020	0.760±0.031	0.709±0.005	0.713±0.030	0.706±0.022	0.720±0.015	0.242±0.010
SiamDTI	**0.863±0.019**	0.807±0.040	0.720±0.010	0.729±0.015	0.712±0.020	0.736±0.013	0.250±0.015
BioT5	0.856±0.003	**0.853±0.003**	0.715±0.002	**0.741±0.010**	**0.738±0.009**	**0.739±0.013**	0.258±0.013
FusionDTI-BAN	0.784±0.002	0.790±0.003	0.733±0.003	0.725±0.002	0.713±0.004	0.698±0.013	0.212±0.011
FusionDTI-CAN	0.801±0.037	0.803±0.032	**0.738±0.002**	0.736±0.010	0.732±0.013	0.737±0.010	**0.261±0.010**

Table 12: Cross-domain performance comparison of FusionDTI and the baselines on the BindingDB, Human and BioSNAP datasets (**Best**, Second Best).