LANTERN: LEVERAGING LARGE LANGUAGE MODELS AND TRANSFORMERS FOR ENHANCED MOLECULAR INTERACTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding molecular interactions such as Drug-Target Interaction (DTI), Protein-Protein Interaction (PPI), and Drug-Drug Interaction (DDI) is critical for advancing drug discovery and systems biology. However, existing methods often struggle with scalability due to the vast chemical and biological space and suffer from limited accuracy when capturing intricate biochemical relationships. To address these challenges, we introduce **LANTERN** (Leveraging Large **LAN**guage Models and **T**ransformers for **E**nhanced molecula**R** interactio**N**s), a novel deep learning framework that integrates Large Language Models (LLMs) with Transformer-based architectures to model molecular interactions more effectively. LANTERN generates high-quality, context-aware embeddings for drug and protein sequences, enabling richer feature representations and improving predictive accuracy. By leveraging a Transformer-based fusion mechanism, our framework enhances scalability by efficiently integrating diverse interaction data while maintaining computational feasibility. Experimental results demonstrate that LANTERN achieves state-of-the-art performance on multiple DTI and DDI benchmarks, significantly outperforming traditional deep learning approaches. Additionally, LANTERN exhibits competitive performance on challenging PPI tasks, underscoring its versatility across diverse molecular interaction domains. The proposed framework offers a robust and adaptable solution for modeling molecular interactions, efficiently handling a diverse range of molecular entities without the need for 3D structural data and making it a promising framework for foundation models in molecular interaction. Our findings highlight the transformative potential of combining LLM-based embeddings with Transformer architectures, setting a new standard for molecular interaction prediction. The source code and relevant documentation are available at: https://github.com/anonymousreseach99/LANTERN.git.

1 INTRODUCTION

Deciphering molecular interactions—including Drug-Target Interactions (DTI), Protein-Protein Interactions (PPI), and Drug-Drug Interactions (DDI)—is crucial for advancing drug discovery Sachdev & Gupta (2019); Liao et al. (2025), therapeutic innovation Grizzle et al. (2019); Luo et al. (2024), systems biology Hu et al. (2021); Meng et al. (2021), protein design Tran & Hy (2024); Nguyen et al. (2024), and protein-binding ligand generation Khang Ngo & Son Hy (2024) through Generative AI. These interactions are pivotal for uncovering potential drug candidates, elucidating disease pathways, and crafting effective treatments. Yet, the intricate and diverse nature of molecular biology presents substantial challenges in accurately predicting these interactions, necessitating the development of advanced computational strategies. Recent advancements in deep learning have revolutionized computational biology by offering powerful methods for modeling complex biological systems. Transformer architectures Vaswani et al. (2017), originally designed for natural language processing (NLP), have demonstrated remarkable success in capturing long-range dependencies and intricate relationships. Simultaneously, Large Language Models (LLMs) have shown their ability to generate meaningful embeddings for biological sequences, such as SMILES for drugs

Elnaggar et al. (2021); Brandes et al. (2022); Hayes et al. (2025) and amino acid sequences for proteins Chithrananda et al. (2020); Ross et al. (2022); Edwards et al. (2022); Nguyen & Hy (2024); Khang Ngo & Son Hy (2024). These embeddings retain rich biochemical and structural information, providing a promising avenue for understanding molecular interactions.

Existing approaches for molecular interaction prediction leverage various machine learning techniques but face notable limitations. ConPLex Singh et al. (2023) utilizes a pretrained protein language model to predict drug-target interactions by co-locating proteins and drug molecules in a shared feature space, achieving solid performance. However, its reliance on choosing appropriate loss functions based on molecule and protein diversity can lead to data leakage and poor generalization. iNGNN-DTI Sun et al. (2024) applies a nested graph neural network (GNN) for DTI prediction, leveraging pre-trained molecular and protein models, and constructing target graphs from AlphaFold2 3D structures. While it enhances interpretability, it still faces challenges in fully capturing molecular relationships. Its use of a cross-attention-free transformer fails to jointly model drug and protein features due to their distinct distributions, and the reliance on pure MLPs for prediction limits performance, especially in complex cases. SkipGNN Huang et al. (2020) introduces a novel GNN architecture that propagates neural messages via both direct and second-order interactions, improving molecular interaction discovery. However, it still suffers from a lack of biological context, as it does not incorporate pretrained biological language models. MUSE Rao et al. (2024) proposes a multi-scale EM-based framework that integrates structural and network-level information for protein-drug interactions. However, it heavily relies on structural data, which is often unavailable. Moreover, protein structures are merely approximations influenced by experimental techniques like X-ray crystallography Smyth & Martin (2000), cryo-electron microscopy (cryo-EM) Tye et al. (2017), or nuclear magnetic resonance (NMR), all of which introduce uncertainties. In drug design, accurate 3D target structures, such as binding sites, are often missing, further limiting structure-based approaches. Many existing methods, including those by Jha et al. (2022), Zhang et al. (2024), Zhu et al. (2024), and Li et al. (2022), either depend on inaccessible structural data or fail to capture complex relationships due to simplified architectures like MLPs.

To address these challenges, we propose a novel framework that combines the strengths of LLM-based sequence embeddings and Transformer architectures to model molecular interactions. Drugs and proteins are represented as embeddings learned by domain-specific LLMs, capturing their intrinsic biochemical properties. These embeddings are then fused through a Transformer model, which effectively captures the interaction patterns between different molecular entities. This unified approach enables the prediction of DTIs, PPIs, and DDIs with high accuracy and generalization. We evaluate our framework on a diverse set of molecular interaction benchmarks, achieving state-of-the-art (SOTA) performance on multiple DTI and DDI datasets and competitive results on PPI benchmarks. These results highlight the versatility and effectiveness of our approach in addressing various molecular interaction prediction tasks. Our contributions can be summarized as follows:

- **Integration of pretrained LLM embeddings with Transformer-based interaction modeling:** We combine the rich sequence-level representations of drugs and proteins from pretrained large language models (LLMs) with a Transformer encoder layer to jointly model interactions. This approach captures complex relationships between molecular entities and significantly enhances prediction tasks.
- **Broad applicability and SOTA performance:** Our method achieves SOTA performance on three standard drug-target interaction (DTI) datasets and competitive results on protein-protein interaction (PPI) and drug-drug interaction (DDI) benchmarks, demonstrating its generalizability and versatility across fundamental biological prediction tasks and making it an ideal candidate for developing foundation models for molecular interaction.
- **Efficiency and independence from 3D structural data:** Unlike approaches reliant on 3D molecular structures, which are label-intensive and architecturally complex, our framework operates efficiently using only sequence data, making it adaptable to arbitrary drugs and proteins. Its efficacy and versatility make it well-suited for building robust, large-scale foundation models for molecular interactions.

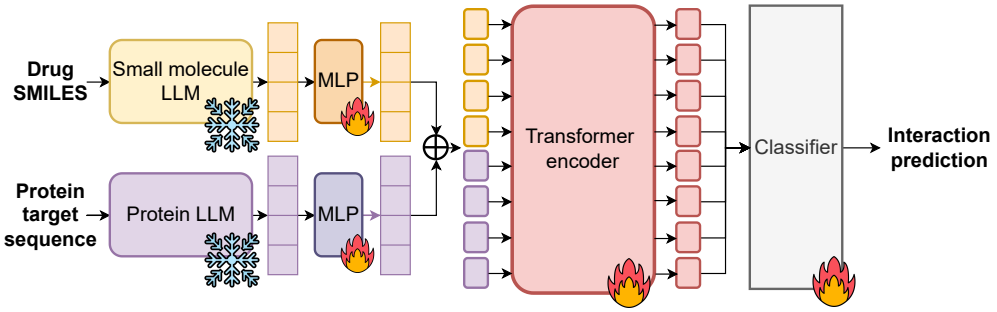

Figure 1: The diagram illustrates the flow of data through different components for the Drug-Target Interaction (DTI) task: Drug SMILES are processed by a Small molecule LLM and then passed through an MLP (Multi-Layer Perceptron). Similarly, Protein target sequences are processed by a Protein LLM and then passed through another MLP. The outputs of these MLPs are concatenated (denoted by $\oplus$) and fed into a Transformer encoder, which then passes the processed data to a Classifier for interaction prediction.

2 METHOD

The Drug-Target Interaction (DTI) prediction task can be framed as a binary classification problem. Given a drug $d \in \mathcal{D}$, represented by its SMILES notation, and a protein $p \in \mathcal{P}$, represented by its amino acid sequence, the goal is to learn a function $f : \mathcal{D} \times \mathcal{P} \to \{0, 1\}$ that predicts whether an interaction exists. Formally, the model aims to approximate $\hat{y} = f(d, p; \theta)$, where $\hat{y} \in \{0, 1\}$ represents the predicted interaction (1 for interaction, 0 for no interaction), and θ refers to the model's learnable parameters.

Drug Representation. The drug d is first encoded using a pretrained Small Molecule LLM, denoted as LLM_{drug}, which transforms the SMILES notation into an embedding: $h_d = \text{LLM}_{\text{drug}}(d)$. This embedding h_d captures the drug's structural and chemical features. The embedding is then further processed by a Multi-Layer Perceptron (MLP), parameterized by ϕ_d, to refine the drug representation: $z_d = \text{MLP}_{\phi_d}(h_d)$. This step enhances the drug feature vector z_d, making it suitable for interaction prediction.

Protein Representation. Similarly, the protein sequence p is encoded by a pretrained Protein LLM, denoted as $\text{LLM}_{\text{protein}}$, which generates a sequence-level embedding: $h_p = \text{LLM}_{\text{protein}}(p)$. The embedding h_p represents the protein's sequence and structure. This sequence-level embedding is further refined by another MLP, parameterized by ϕ_p: $z_p = \text{MLP}_{\phi_p}(h_p)$. The result is a feature vector z_p that encapsulates the protein's characteristics relevant to binding with drugs.

Unified Representation. The drug and protein embeddings are then concatenated to form a unified feature representation: $z_{\text{fusion}} = z_d \oplus z_p$, where $\oplus$ denotes the concatenation operation, combining the drug and protein features into a single vector. The fused representation z_{fusion} is then processed by a Transformer encoder $\mathcal{T}$, which captures complex relationships between the drug and protein features. The attention mechanism in the Transformer allows the model to learn dependencies between the two types of entities (drug and protein), improving the accuracy of predictions: $z_{\text{trans}} = \mathcal{T}(z_{\text{fusion}})$. Finally, a classifier $\mathcal{C}$, typically implemented as an MLP followed by a sigmoid activation function, is applied to the Transformer output to predict the probability of interaction: $\hat{y} = \sigma(\mathcal{C}(z_{\text{trans}}))$, where $\sigma(\cdot)$ is the sigmoid function, ensuring that the output $\hat{y}$ is in the range $(0, 1)$, representing the probability of interaction.

Optimization. The model is trained using binary cross-entropy loss, which measures the difference between the predicted interaction probabilities and the true labels y_i for the i-th drug-target pair. The loss function is:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right],$$

where N is the total number of training samples, y_i is the ground truth label for the i-th pair (1 for interaction, 0 for no interaction), and $\hat{y}_i$ is the predicted probability for the interaction. This loss function encourages the model to make accurate predictions by penalizing errors in both positive and negative classifications.

2.1 Generalization to Other Interaction Tasks

The proposed architecture is not limited to drug-target interaction prediction. It can be generalized to other interaction prediction tasks, such as protein-protein and drug-drug interactions. For protein-protein interactions, both the Small Molecule LLM and the Protein LLM in Figure 1 are replaced with two instances of the Protein LLM, enabling the extraction of biologically meaningful features from both input protein sequences. Similarly, for drug-drug interactions, the architecture employs two Small Molecule LLMs to process the SMILES representations of the interacting drugs.

3 Experiments

3.1 Datasets

The study leverages a variety of benchmark datasets to assess the performance of the proposed methods in different interaction prediction tasks. For Drug-Target Interaction (DTI) prediction, the DAVIS Davis et al. (2011), KIBA He et al. (2017), and BioSNAP Zitnik et al. (2018) datasets provide comprehensive drug-protein interaction data, summarized in Table 1. For Protein-Protein Interaction (PPI) tasks, the Yeast PPI Ito et al. (2001) dataset is employed, containing 2,497 proteins and 11,188 interactions. In the case of Drug-Drug Interaction (DDI) prediction, the DeepDDI Rao et al. (2024) dataset provides data on drug-drug interactions and side effects, sourced from DrugBank Wishart et al. (2018), making it an essential resource for studying potential adverse drug reactions.

Dataset	#Drugs	#Proteins	#Interactions	#Positives	#Negatives
DAVIS	68	442	30,056	1506	28,550
KIBA	2068	229	118,254	22,729	95,525
BioSNAP	4510	2180	27,428	13,817	13,611

Table 1: Summary of the DAVIS, KIBA, and BioSNAP datasets used for drug-target interaction (DTI) tasks, including the number of drugs, proteins, interactions, and the distribution of positive and negative interactions.

3.2 Architecture and Implementation Details

The detailed architectural configurations, hyperparameter selections, and implementation specifics are thoroughly documented in Appendix A.1. This section covers critical aspects such as the computational resources utilized, including the type of GPUs employed, the selection of large language models (LLMs), optimization strategies, and the number of layers in the proposed architecture.

3.3 Ablation study

BioSNAP and DAVIS were selected for the ablation study due to their distinct characteristics and computational feasibility. DAVIS, which consists of a small number of drugs with dense interaction data, provides a controlled setting for evaluating model performance on well-characterized targets. In contrast, BioSNAP offers a more balanced distribution of positive and negative interactions, enabling a comprehensive assessment of model generalization.

Effect of LLMs selection: To identify the optimal large language model (LLM) for interaction prediction tasks, we conducted an ablation study evaluating the performance of various LLMs tailored for proteins and drugs. For proteins, we considered models such as ProtT5 Elnaggar et al. (2021), ProtBERT Brandes et al. (2022) and ESM3 Hayes et al. (2025) (Evolutionary Scale Modeling version 3), while for drugs, we evaluated models like ChemBERTa Chithrananda et al. (2020), MoLFormer Ross et al. (2022) and MolT5 Edwards et al. (2022). The evaluation was performed on

two benchmark datasets, DAVIS and BioSNAP, which provide diverse and complementary data for drug-target interaction prediction tasks. The results from these datasets allowed for a robust comparison of the LLMs, highlighting their strengths and weaknesses in capturing biochemical interactions effectively.

		BioSNAP		DAVIS	
Proteins	**Drugs**	AUROC (↑)	AUPRC (↑)	AUROC (↑)	AUPRC (↑)
ProtT5	ChemBERTa	0.9911	0.9907	0.989	0.827
	MoLFormer	**0.9953**	**0.9961**	*0.991*	*0.878*
	MolT5	0.9946	0.9955	0.991	0.872
ProtBERT	ChemBERTa	0.9807	0.9857	0.990	0.849
	MoLFormer	*0.9947*	*0.9953*	0.991	0.857
	MolT5	0.9842	0.9880	0.991	0.867
ESM3	ChemBERTa	0.9927	0.9938	0.990	0.871
	MoLFormer	<u>0.9948</u>	<u>0.9958</u>	**0.995**	**0.905**
	MolT5	0.9887	0.9908	<u>0.993</u>	<u>0.882</u>

Table 2: Performance comparison of large language models (LLMs) for proteins and drugs on the BioSNAP and DAVIS datasets. The metrics reported are the Area Under the ROC Curve (AUROC) and the Area Under the Precision-Recall Curve (AUPRC). The highest performance values are marked in bold, the second-highest are <u>underlined</u>, and the third-highest are *italicized*, showcasing the relative effectiveness of each model combination.

Table 2 illustrates the performance of various large language models (LLMs) for proteins and drugs was evaluated on the BioSNAP and DAVIS datasets using AUROC and AUPRC metrics. The results indicate that the combination of ProtT5 and MoLFormer achieved the highest performance on BioSNAP, with AUROC and AUPRC values of 0.9953 and 0.9961, demonstrating its effectiveness in predicting interactions in this dataset. On the DAVIS dataset, ESM3 paired with MoLFormer emerged as the top-performing model, attaining the highest AUROC (0.995) and AUPRC (0.905). MoLFormer consistently delivered strong results across both datasets, showcasing its robustness as a drug representation model. Among the protein models, ProtT5 and ESM3 were particularly effective, with ESM3 excelling in DAVIS and ProtT5 in BioSNAP. Additionally, ProtBERT paired with MoLFormer also showed competitive performance, making it a viable alternative for specific scenarios. Based on these findings, ProtT5 with MoLFormer is recommended for BioSNAP, while ESM3 with MoLFormer is best suited for DAVIS. For applications requiring a single versatile combination, MoLFormer paired with either ProtT5 or ESM3 provides a robust solution.

Impact of Transformer-based Encoding: We evaluated the performance of Transformer architectures against traditional Multi-Layer Perceptrons (MLPs) for encoding tasks. Our findings indicate that transformer-based models with the combination of ProtT5 for protein encoding and MoLFormer for drug encoding, significantly outperformed MLP-based approaches on the BioSNAP dataset. Table 3 presents the performance metrics of different model configurations, highlighting the effect of removing or replacing Transformer components. The results demonstrate that the full Transformer-based model achieves the highest AUROC (0.9953) and AUPRC (0.9961), along with superior sensitivity (0.9511) and specificity (0.9562).

Model Configuration	**AUROC**	**AUPRC**	**Sensitivity**	**Specificity**
Remove Feed Forward and Add & Norm	0.9942	0.9915	0.9496	0.9275
Remove Multi-head Attention	0.8571	0.8606	0.8072	0.8613
Remove Whole Transformer	0.7601	0.7541	0.5669	0.7852
Replace Whole Transformer by MLP	0.8485	0.8559	0.8805	0.8541
Use Whole Transformer	**0.9953**	**0.9961**	**0.9511**	**0.9562**

Table 3: Performance metrics (AUROC and AUPRC) for different model configurations, evaluating the impact of removing or replacing Transformer components.

Ablation studies reveal a substantial decline in performance when key Transformer components are removed. Eliminating the Feed Forward and the Add & Normalization steps results in a slight

reduction in AUROC (0.9942) and AUPRC (0.9915), suggesting that these components contribute to fine-tuning the model's performance. In contrast, removing the Multi-head Attention mechanism causes a significant drop in AUROC (0.8571) and AUPRC (0.8606), emphasizing the crucial role of attention mechanism in feature extraction. The most severe degradation is observed when the entire Transformer structure is removed, leading to an AUROC of 0.7601 and an AUPRC of 0.7541, with markedly reduced sensitivity (0.5669).

Furthermore, replacing the entire Transformer with a MLP results in an AUROC of 0.8485 and an AUPRC of 0.8559, which is still lower than any Transformer-based configuration, reinforcing the superiority of Transformer architectures over traditional MLPs for encoding tasks. These findings underscore the critical role of Transformer components, particularly Multi-head Attention, in achieving optimal performance in drug-protein interaction prediction.

Appendix A.2 provides an in-depth analysis of the advantages of Transformer-based fusion models compared to traditional MLPs. The results presented in Table 3 further corroborate these findings, emphasizing the critical role of various Transformer components in encoding performance. Specifically, the removal of key elements, such as Multi-head Attention, led to a substantial decline in predictive accuracy, thereby reinforcing the theoretical insights discussed in Appendix A.2.

Figure 2 presents t-SNE visualizations of data representations, illustrating the impact of MLP and Transformer models on feature distribution and clustering. Observing the transformations, it is evident that the MLP modifies the data distribution to a certain extent, enhancing separation but still exhibiting some overlap. In contrast, the Transformer model demonstrates a more pronounced clustering effect, indicating its superior capability in capturing complex relationships and structural patterns within the data. These visualizations underscore the effectiveness of Transformer-based models in producing well-defined feature representations compared to traditional MLP approaches.

By integrating both theoretical analysis and empirical validation, our study underscores the enhanced capability of Transformer-based models in capturing complex feature interactions, presenting a compelling case for their adoption in drug-target interaction tasks.

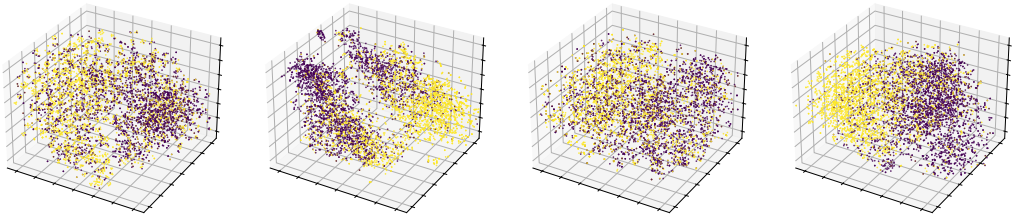

Figure 2: t-SNE visualizations of data representations: (a) before applying the MLP, (b) after applying the MLP, (c) before applying the Transformer, and (d) after applying the Transformer. Data points are color-coded according to their labels, where ■ represents label 0, ■ represents label 1.

3.4 Benchmarks

3.4.1 Drug-target interaction benchmark

Tables 4, 5, and 6 present a comprehensive performance comparison of various drug-target interaction (DTI) prediction methods across the BioSNAP, DAVIS, and KIBA datasets. The results highlight the superiority of our Transformer-based approach over existing models, including DeepDTA Öztürk et al. (2018), Moltrans Huang et al. (2021), ML-DTI Yang et al. (2021), DGraphGTA Jiang et al. (2020), and iNGNN-DTI Sun et al. (2024).

Across all datasets, our method consistently achieves the highest AUROC and AUPRC, with substantial improvements over state-of-the-art approaches. Notably, on the BioSNAP dataset, our model attains an AUROC of 0.995 and an AUPRC of 0.996, surpassing iNGNN-DTI, the second-best model, by a significant margin. Similar trends are observed for the DAVIS dataset, where our method outperforms competing models with a notable increase in both AUROC (0.995) and AUPRC (0.905). The most pronounced performance gain is seen on the KIBA dataset, where our approach achieves an AUROC of 0.976 and an AUPRC of 0.977, demonstrating its robustness across different benchmarks.

These results underscore the effectiveness of our model in accurately capturing drug-protein interactions, significantly outperforming MLP-based and graph-based methods. The observed improvements in sensitivity and specificity further validate the model's reliability in both detecting positive interactions and minimizing false positives, making it a promising tool for DTI prediction.

Method	**AUROC**	**AUPRC**	**Sensitivity**	**Specificity**
DeepDTA	0.897 ± 0.0027	0.900 ± 0.0046	0.859 ± 0.0089	0.786 ± 0.0197
Moltrans	0.887 ± 0.0034	0.881 ± 0.0085	0.824 ± 0.0106	0.809 ± 0.0104
ML-DTI	0.911 ± 0.0053	0.911 ± 0.0112	0.851 ± 0.0054	0.828 ± 0.0215
DGraphGTA (Alphafold2)	0.913 ± 0.0022	0.917 ± 0.0024	0.858 ± 0.0175	0.831 ± 0.0151
iNGNN-DTI	0.934 ± 0.0021	0.939 ± 0.0022	0.872 ± 0.0189	0.854 ± 0.0200
Our method	$\mathbf{0.995 \pm 0.0045}$	$\mathbf{0.996 \pm 0.0036}$	$\mathbf{0.951 \pm 0.0409}$	$\mathbf{0.956 \pm 0.0329}$

Table 4: Performance comparison of various methods on the DTI task using the BioSNAP datasets. The table reports the AUROC and AUPRC with their respective standard deviations.

Method	**AUROC**	**AUPRC**	**Sensitivity**	**Specificity**
DeepDTA	0.892 ± 0.0066	0.378 ± 0.0231	0.854 ± 0.0066	0.792 ± 0.0291
Moltrans	0.898 ± 0.0050	0.371 ± 0.0067	0.865 ± 0.0050	0.783 ± 0.0387
ML-DTI	0.910 ± 0.0034	0.381 ± 0.0247	0.895 ± 0.0034	0.795 ± 0.0183
DGraphGTA (Alphafold2)	0.885 ± 0.0099	0.316 ± 0.0447	0.894 ± 0.0034	0.724 ± 0.0467
iNGNN-DTI	0.931 ± 0.0027	0.473 ± 0.0167	0.922 ± 0.0155	0.802 ± 0.0240
Our method	$\mathbf{0.995 \pm 0.0037}$	$\mathbf{0.905 \pm 0.0238}$	$\mathbf{0.976 \pm 0.0159}$	$\mathbf{0.964 \pm 0.0207}$

Table 5: Performance comparison of various methods on the DTI task using the DAVIS datasets. The table reports the AUROC and AUPRC with their respective standard deviations.

Method	**AUROC**	**AUPRC**	**Sensitivity**	**Specificity**
DeepDTA	0.912 ± 0.0037	0.743 ± 0.0127	0.881 ± 0.0056	0.780 ± 0.0127
Moltrans	0.899 ± 0.0022	0.691 ± 0.0142	0.872 ± 0.0116	0.760 ± 0.0160
ML-DTI	0.909 ± 0.0020	0.727 ± 0.0108	0.878 ± 0.0111	0.779 ± 0.0113
DGraphGTA (Alphafold2)	0.911 ± 0.0004	0.739 ± 0.0043	0.881 ± 0.0183	0.784 ± 0.0277
iNGNN-DTI	0.915 ± 0.0016	0.753 ± 0.0071	0.888 ± 0.0183	0.779 ± 0.0146
Our method	$\mathbf{0.976 \pm 0.0154}$	$\mathbf{0.977 \pm 0.0088}$	$\mathbf{0.959 \pm 0.0268}$	$\mathbf{0.965 \pm 0.0074}$

Table 6: Performance comparison of various methods on the DTI task using the KIBA datasets. The table reports the AUROC and AUPRC with their respective standard deviations.

3.4.2 DRUG-DRUG INTERACTION BENCHMARK

Table 7 presents a comparative analysis of drug-drug interaction (DDI) prediction methods on the DeepDDI dataset, evaluated using AUROC and AUPRC metrics. Our proposed method achieves an AUROC of 0.998 and an AUPRC of 0.995, outperforming all other approaches. Notably, while MUSE attains the same AUROC, our method achieves the highest AUPRC, demonstrating superior precision in ranking positive interactions. These results highlight the effectiveness of our approach in enhancing DDI prediction accuracy.

Method	**AUROC** (↑)	**AUPRC** (↑)
SSI-DDI Rao et al. (2024)	0.868	0.871
MIRACLE Wang et al. (2021)	0.944	0.895
CGIB Lee et al. (2023)	0.950	0.961
MUSE Rao et al. (2024)	**0.998**	0.993
Our method	**0.998**	**0.995**

Table 7: Performance comparison of drug-drug interaction prediction methods using the DeepDDI datasets, evaluated by AUROC and AUPRC metrics. Bold values indicate the best performance in each category.

3.4.3 PROTEIN-PROTEIN INTERACTION BENCHMARK

Table 8 presents a comparative evaluation of various methods on the Yeast PPI dataset across multiple performance metrics, including accuracy (Acc), precision (Pre), sensitivity (Sen), specificity (Spe), F1-score (F1), Matthews correlation coefficient (MCC), and area under the curve (AUC). The TAGPPI method demonstrates superior overall performance, achieving the highest scores in most metrics. In contrast, our proposed methods, particularly *ProtBERT* and *ProtT5*, exhibit exceptional sensitivity, reaching a near-perfect 99.82%, along with competitive AUC values. However, this comes at the cost of lower precision and accuracy, making our approach particularly well-suited for applications that prioritize high recall, such as identifying novel protein-protein interactions.

Method	Acc	Pre	Sen	Spe	F1	MCC	AUC
MCD-SVM You et al. (2014)	91.36	91.94	90.67	NA	91.30	84.21	97.07
RF-LPQ Wong et al. (2015)	93.92	96.45	91.10	NA	93.70	88.56	NA
kNN-CTD Yang et al. (2010)	86.15	90.24	81.03	NA	85.39	NA	NA
EELM-PCA You et al. (2013)	86.99	87.59	86.15	NA	86.86	77.36	NA
DeepPPI Du et al. (2017)	94.43	96.65	92.06	NA	94.30	88.97	97.45
SAE Sun et al. (2017)	67.17	66.90	68.06	66.30	67.44	34.39	NA
DPPI Hashemifar et al. (2018)	94.55	96.68	92.24	NA	94.41	NA	NA
DNN-PPI Li et al. (2018)	76.61	75.10	79.63	73.59	77.29	53.32	74.35
PIPR Chen et al. (2019)	97.09	97.00	97.17	97.00	97.09	94.17	NA
TAGPPI Song et al. (2022)	**97.81**	**98.10**	98.26	**98.10**	**97.80**	**95.63**	97.74
Our method (*ProtT5*)	77.35	68.87	**99.82**	54.99	81.50	61.23	96.23
Our method (*ProtBERT*)	79.27	70.74	**99.82**	59.95	82.80	64.21	**97.93**
Our method (*ESM3*)	84.47	93.60	71.85	95.08	81.29	68.82	94.61

Table 8: Performance comparison of different methods using the Yeast PPI datasets. NA means the corresponding metric is not available from the original paper. Bold font indicates the best result in the column.

4 CONCLUSION

In this work, we introduced **LANTERN**, a novel framework that leverages Large Language Models (LLMs) and Transformer architectures to enhance molecular interaction prediction. By integrating pretrained embeddings with a Transformer-based fusion mechanism, LANTERN effectively captures complex biochemical relationships across diverse interaction tasks, including Drug-Target Interactions (DTI), Protein-Protein Interactions (PPI), and Drug-Drug Interactions (DDI).

Our extensive evaluations demonstrate that LANTERN achieves state-of-the-art (SOTA) performance on multiple DTI and DDI benchmarks and exhibits competitive results in PPI tasks, underscoring its robustness and versatility. The ablation studies further highlight the importance of Transformer-based encoding over traditional MLP architectures, validating the superior representation learning capabilities of attention mechanisms.

Beyond achieving strong predictive performance, LANTERN offers a scalable and generalizable solution that does not require 3D structural data, making it highly applicable for drug discovery, therapeutic development, and network biology. Integrating self-supervised pretraining strategies may enhance adaptability to new molecular interaction domains. Another key direction is leveraging multiple LLM models within the same data type, such as drugs, to harness the complementary knowledge from various models, addressing individual model limitations and optimizing predictions for a more comprehensive understanding of the problem.

MEANINGFULNESS STATEMENT

A meaningful representation of life captures biological entities through rich embeddings that enable the discovery of novel interactions and insights via computational methods, validated in real-world applications. Our work lays the groundwork for developing a large-scale foundation model for molecular interaction prediction, advancing our ability to represent and understand molecular entities at an unprecedented scale.

REFERENCES

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

Muhao Chen, Chelsea J-T Ju, Guangyu Zhou, Xuelu Chen, Tianran Zhang, Kai-Wei Chang, Carlo Zaniolo, and Wei Wang. Multifaceted protein–protein interaction prediction based on siamese residual rcnn. *Bioinformatics*, 35(14):i305–i314, 2019.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *CoRR*, abs/2010.09885, 2020. URL https://arxiv.org/abs/2010.09885.

Mindy I Davis, Jeremy P Hunt, Sanna Herrgard, Pietro Ciceri, Lisa M Wodicka, Gabriel Pallares, Michael Hocker, Daniel K Treiber, and Patrick P Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology*, 29(11):1046–1051, 2011.

Xiuquan Du, Shiwei Sun, Changlin Hu, Yu Yao, Yuanting Yan, and Yanping Zhang. Deepppi: boosting prediction of protein–protein interactions with deep neural networks. *Journal of chemical information and modeling*, 57(6):1499–1510, 2017.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 375–413, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.26. URL https://aclanthology.org/2022.emnlp-main.26/.

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. 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*, pp. 1–1, 2021. doi: 10.1109/TPAMI.2021.3095381.

Amy J Grizzle, John Horn, Carol Collins, Jodi Schneider, Daniel C Malone, Britney Stottlemyer, and Richard David Boyce. Identifying common methods used by drug interaction experts for finding evidence about potential drug-drug interactions: web-based survey. *Journal of medical Internet research*, 21(1):e11182, 2019.

Somaye Hashemifar, Behnam Neyshabur, Aly A Khan, and Jinbo Xu. Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics*, 34(17):i802–i810, 2018.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, pp. eads0018, 2025.

Tong He, Marten Heidemeyer, Fuqiang Ban, Artem Cherkasov, and Martin Ester. Simboost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. *Journal of cheminformatics*, 9:1–14, 2017.

Lun Hu, Xiaojuan Wang, Yu-An Huang, Pengwei Hu, and Zhu-Hong You. A survey on computational models for predicting protein–protein interactions. *Briefings in bioinformatics*, 22(5): bbab036, 2021.

Kexin Huang, Cao Xiao, Lucas M Glass, Marinka Zitnik, and Jimeng Sun. Skipgnn: predicting molecular interactions with skip-graph networks. *Scientific reports*, 10(1):21092, 2020.

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

Takashi Ito, Tomoko Chiba, Ritsuko Ozawa, Mikio Yoshida, Masahira Hattori, and Yoshiyuki Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences*, 98(8):4569–4574, 2001.

Kanchan Jha, Sriparna Saha, and Hiteshi Singh. Prediction of protein–protein interaction using graph neural networks. *Scientific Reports*, 12(1):8360, 2022.

Mingjian Jiang, Zhen Li, Shugang Zhang, Shuang Wang, Xiaofeng Wang, Qing Yuan, and Zhiqiang Wei. Drug–target affinity prediction using graph neural network and contact maps. *RSC advances*, 10(35):20701–20712, 2020.

Nhat Khang Ngo and Truong Son Hy. Multimodal protein representation learning and target-aware variational auto-encoders for protein-binding ligand generation. *Machine Learning: Science and Technology*, 5(2):025021, apr 2024. doi: 10.1088/2632-2153/ad3ee4. URL https://dx.doi.org/10.1088/2632-2153/ad3ee4.

Namkyeong Lee, Dongmin Hyun, Gyoung S Na, Sungwon Kim, Junseok Lee, and Chanyoung Park. Conditional graph information bottleneck for molecular relational learning. In *International Conference on Machine Learning*, pp. 18852–18871. PMLR, 2023.

Hang Li, Xiu-Jun Gong, Hua Yu, and Chang Zhou. Deep neural network based predictions of protein interactions using primary sequences. *Molecules*, 23(8):1923, 2018.

Xue Li, Peifu Han, Gan Wang, Wenqi Chen, Shuang Wang, and Tao Song. Sdnn-ppi: self-attention with deep neural network effect on protein-protein interaction prediction. *BMC genomics*, 23(1): 474, 2022.

Qian Liao, Yu Zhang, Ying Chu, Yi Ding, Zhen Liu, Xianyi Zhao, Yizheng Wang, Jie Wan, Yijie Ding, Prayag Tiwari, et al. Application of artificial intelligence in drug-target interactions prediction: A review. *npj Biomedical Innovations*, 2(1):1, 2025.

Huimin Luo, Weijie Yin, Jianlin Wang, Ge Zhang, Wenjuan Liang, Junwei Luo, and Chaokun Yan. Drug-drug interactions prediction based on deep learning and knowledge graph: A review. *Iscience*, 2024.

Xiangmao Meng, Wenkai Li, Xiaoqing Peng, Yaohang Li, and Min Li. Protein interaction networks: centrality, modularity, dynamics, and applications. *Frontiers of Computer Science*, 15:1–17, 2021.

Viet Thanh Duy Nguyen and Truong Son Hy. Multimodal pretraining for unsupervised protein representation learning. *Biology Methods and Protocols*, 9(1):bpae043, 06 2024. ISSN 2396-8923. doi: 10.1093/biomethods/bpae043. URL https://doi.org/10.1093/biomethods/bpae043.

Viet Thanh Duy Nguyen, Nhan D. Nguyen, and Truong Son Hy. Proteinrediff: Complex-based ligand-binding proteins redesign by equivariant diffusion-based generative models. *Structural Dynamics*, 11(6):064102, 11 2024. ISSN 2329-7778. doi: 10.1063/4.0000271. URL https://doi.org/10.1063/4.0000271.

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

Jiahua Rao, Jiancong Xie, Qianmu Yuan, Deqin Liu, Zhen Wang, Yutong Lu, Shuangjia Zheng, and Yuedong Yang. A variational expectation-maximization framework for balanced multi-scale learning of protein and drug interactions. *Nature Communications*, 15(1):4476, 2024.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022. doi: 10.1038/s42256-022-00580-7.

Kanica Sachdev and Manoj Kumar Gupta. A comprehensive review of feature based methods for drug target interaction prediction. *Journal of biomedical informatics*, 93:103159, 2019.

Rohit Singh, Samuel Sledzieski, Bryan Bryson, Lenore Cowen, and Bonnie Berger. Contrastive learning in protein language space predicts interactions between drugs and protein targets. *Proceedings of the National Academy of Sciences*, 120(24):e2220778120, 2023.

MS Smyth and JHJ Martin. x ray crystallography. *Molecular Pathology*, 53(1):8, 2000.

Bosheng Song, Xiaoyan Luo, Xiaoli Luo, Yuansheng Liu, Zhangming Niu, and Xiangxiang Zeng. Learning spatial structures of proteins improves protein–protein interaction prediction. *Briefings in bioinformatics*, 23(2):bbab558, 2022.

Tanlin Sun, Bo Zhou, Luhua Lai, and Jianfeng Pei. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC bioinformatics*, 18:1–8, 2017.

Yan Sun, Yan Yi Li, Carson K Leung, and Pingzhao Hu. ingnn-dti: prediction of drug–target interaction with interpretable nested graph neural network and pretrained molecule models. *Bioinformatics*, 40(3):btae135, 2024.

Thanh V. T. Tran and Truong Son Hy. Protein design by directed evolution guided by large language models. *IEEE Transactions on Evolutionary Computation*, pp. 1–1, 2024. doi: 10.1109/TEVC.2024.3439690.

Bik-Kwoon Yeung Tye, Yuanliang Zhai, et al. Developing cryo-electron microscopy for the high-resolution structure determination of biomolecules in solution: 2017 nobel prize in chemistry. 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Yingheng Wang, Yaosen Min, Xin Chen, and Ji Wu. Multi-view graph contrastive representation learning for drug-drug interaction prediction. In *Proceedings of the web conference 2021*, pp. 2921–2933, 2021.

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082, 2018.

Leon Wong, Zhu-Hong You, Shuai Li, Yu-An Huang, and Gang Liu. Detection of protein-protein interactions from amino acid sequences using a rotation forest model with a novel pr-lpq descriptor. In *Advanced Intelligent Computing Theories and Applications: 11th International Conference, ICIC 2015, Fuzhou, China, August 20-23, 2015. Proceedings, Part III 11*, pp. 713–720. Springer, 2015.

Lei Yang, Jun-Feng Xia, and Jie Gui. Prediction of protein-protein interactions from protein sequence using local descriptors. *Protein and peptide letters*, 17(9):1085–1090, 2010.

Ziduo Yang, Weihe Zhong, Lu Zhao, and Calvin Yu-Chian Chen. Ml-dti: mutual learning mechanism for interpretable drug–target interaction prediction. *The Journal of Physical Chemistry Letters*, 12(17):4247–4261, 2021.

Zhu-Hong You, Ying-Ke Lei, Lin Zhu, Junfeng Xia, and Bing Wang. Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. In *BMC bioinformatics*, volume 14, pp. 1–11. Springer, 2013.

Zhu-Hong You, Lin Zhu, Chun-Hou Zheng, Hong-Jie Yu, Su-Ping Deng, and Zhen Ji. Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set. In *BMC bioinformatics*, volume 15, pp. 1–9. Springer, 2014.

Chengcheng Zhang, Tianyi Zang, and Tianyi Zhao. Kge-unit: toward the unification of molecular interactions prediction based on knowledge graph and multi-task learning on drug discovery. *Briefings in Bioinformatics*, 25(2):bbae043, 2024.

Jing Zhu, Chao Che, Hao Jiang, Jian Xu, Jiajun Yin, and Zhaoqian Zhong. Ssf-ddi: a deep learning method utilizing drug sequence and substructure features for drug–drug interaction prediction. *BMC bioinformatics*, 25(1):39, 2024.

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

A Appendix

A.1 Implementation Details and Model Configurations

In this section, we provide a comprehensive overview of the implementation details and model configurations of the proposed LANTERN framework.

A.1.1 LLM choosing

To featurize the inputs, we utilize Molformer, ChemBERTa, and MolT5 for small molecules (embedding sizes per SMILES string: 768, 384, and 1024, respectively) and ProtBERT, ProtT5, and ESM-3 for proteins (embedding sizes per amino acid: 1024, 1024, and 1536, respectively). Our framework is designed to be flexible, allowing for various biological pretrained LLMs, and we propose the choice of LLMs for each task based on empirical observations.

For most language models, we use the output from their final embedding layer. However, for ProtT5 and MolT5, we specifically use the final embedding layer of their encoder components. All models generate per-amino-acid features for proteins or per-SMILES string features for small molecules. These features are averaged along the sequence length to produce fixed-length vectors for downstream tasks.

- **DTI Benchmark:** Based on performance across different datasets, we use ProtT5 for proteins in BioSNAP, ProtBERT for proteins in KIBA, and ESM3 for proteins in DAVIS. For small molecules, MoLFormer is selected as the best-performing model across all datasets.
- **DDI Benchmark:** MoLFormer is selected due to its robust and consistent performance across experiments.
- **PPI Benchmark:** ProtT5, ProtBERT, and ESM3 are all utilized in this task.

A.1.2 Evaluation Metrics

Model performance was evaluated using standard metrics, including the accuracy (Acc), precision (Pre), sensitivity (Sen), specificity (Spe), F1-score (F1), Matthews correlation coefficient (MCC), area under the curve (AUC) and area under the precision-recall curve (AUC-PR). The formulas for these metrics are as follows:

- **Accuracy (Acc)** measures the proportion of correctly classified samples:
$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN}$$
- **Precision (Pre)** represents the proportion of true positives among all predicted positives:
$$\text{Pre} = \frac{TP}{TP + FP}$$
- **Sensitivity (Sen)** or **Recall** measures the ability to correctly identify positive samples:
$$\text{Sen} = \frac{TP}{TP + FN}$$
- **Specificity (Spe)** measures the ability to correctly identify negative samples:
$$\text{Spe} = \frac{TN}{TN + FP}$$
- **F1-Score (F1)** is the harmonic mean of precision and sensitivity:
$$\text{F1} = 2 \cdot \frac{\text{Pre} \cdot \text{Sen}}{\text{Pre} + \text{Sen}}$$
- **Matthews Correlation Coefficient (MCC)** is a balanced measure that considers all four quadrants of the confusion matrix:
$$\text{MCC} = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

- **Area Under the ROC Curve (AUC)** evaluates the trade-off between sensitivity and specificity across thresholds. While no specific formula is shown here, it is calculated using the area under the Receiver Operating Characteristic (ROC) curve.
- **Area Under the Precision-Recall Curve (AUC-PR)**: Similar to AUC, but focuses on precision-recall trade-offs, especially useful for imbalanced datasets.

A.1.3 Hardware and Software Environment

Experiments were conducted on a system equipped with NVIDIA A100 GPUs and 160GB RAM. The implementation was carried out using PyTorch, with additional libraries such as Hugging Face Transformers for embedding extraction and RDKit for molecular processing.

A.1.4 Hyperparameter Tuning

For all three tasks, a linear layer is employed to project the pretrained embeddings from the large language models (LLMs), including the protein language model and the small molecule language model, into a shared latent space of size 384. The two projected representations are then concatenated and passed through an Transformer encoder layer with 8 attention heads. The classifier, a linear layer of shape (768, 1), transforms the encoder output into a logit. The model is trained and optimized for 100 epochs, with a learning rate initialized at 1e-4 and subsequently reduced following a linear annealing schedule after 30, 60, and 80 epochs, using a decay coefficient of 0.8. The batch size is set to 64 for all datasets except the DeepDDI dataset, where it is increased to 512. Dropout is set to 0.1 for all datasets, except for the Yeast dataset, where it is increased to 0.2.

A.2 Advantages of Transformer-Based Fusion over MLP

The feature fusion step plays a critical role in learning meaningful interactions between drug and protein representations. While traditional approaches such as Multi-Layer Perceptrons (MLPs) offer simple non-linear transformations, they fall short in capturing complex dependencies between input modalities. In contrast, Transformer-based architectures, driven by the attention mechanism, provide several advantages, including the ability to model long-range dependencies, dynamic weighting of input features, and improved representation learning. This section presents a mathematical formulation to justify the superiority of Transformers over MLPs in the proposed architecture.

A.2.1 Limitations of MLP-Based Fusion

An MLP processes concatenated drug and protein feature representations via successive linear transformations followed by non-linear activations. Given the concatenated feature vector $z_{\text{fusion}} \in \mathbb{R}^d$, an MLP of L layers with weight matrices $W^{(l)} l = 1^L$ and biases $b^{(l)} l = 1^L$ produces an output representation:

$$h_{\text{MLP}}^{(l)} = \sigma\left(W^{(l)} h_{\text{MLP}}^{(l-1)} + b^{(l)}\right),$$

where $\sigma(\cdot)$ denotes a non-linear activation function, such as ReLU. Despite its expressiveness, an MLP applies fixed learned weights to all input features, lacking the flexibility to capture contextual relationships between different segments of the input. Consequently, it struggles with:

- **Static Weighting:** The same weights are applied to all input pairs, disregarding potential interactions between drug and protein features.
- **Lack of Interpretability:** MLPs lack mechanisms to quantify feature importance, making it challenging to understand which features contribute most to the interaction prediction.
- **Poor Scalability:** As input dimensions grow, MLPs require exponentially more parameters to capture feature dependencies effectively.

A.2.2 Advantages of Attention Mechanism for Feature Fusion

The self-attention mechanism in Transformers provides a more expressive and efficient method for feature fusion compared to traditional Multi-Layer Perceptrons (MLPs). Unlike MLPs, which apply fixed weights to all input features, self-attention dynamically computes context-dependent feature

interactions, leading to improved representation learning. In this section, we provide a mathematical justification of why attention is a more suitable choice for feature fusion in the Drug-Target Interaction (DTI) task.

Attention Formulation Given the concatenated feature representation $z_{\text{fusion}} \in \mathbb{R}^d$, the self-attention mechanism operates by computing attention scores across all feature dimensions. The attention scores are computed as:

$$\alpha_{ij} = \frac{\exp\left(\frac{q_i k_j^\top}{\sqrt{d}}\right)}{\sum_{j=1}^{d} \exp\left(\frac{q_i k_j^\top}{\sqrt{d}}\right)},$$

where:

- $q_i = W_Q z_i$ and $k_j = W_K z_j$ are the query and key projections of the feature representations,
- $W_Q, W_K \in \mathbb{R}^{d \times d}$ are learnable weight matrices,
- α_{ij} represents the attention weight assigned to feature j when computing the representation of feature i,
- $\sqrt{d}$ serves as a scaling factor to prevent gradient vanishing.

The output of the attention mechanism is computed as a weighted sum of value vectors:

$$z_{\text{attn},i} = \sum_{j=1}^{d} \alpha_{ij} v_j, \quad \text{where } v_j = W_V z_j,$$

where $W_V \in \mathbb{R}^{d \times d}$ is the learnable value projection.

Expressive Power of Attention Compared to MLP MLPs apply a fixed transformation to the input of the form:

$$z_{\text{MLP}} = \sigma(W z_{\text{fusion}} + b).$$

This transformation assumes a linear mapping followed by a non-linear activation $\sigma(\cdot)$, which fails to capture interactions across features in a dynamic fashion. In contrast, attention mechanisms adaptively model feature interactions through:

$$z_{\text{attn}} = \sum_{j=1}^{d} \alpha_{ij} W_V z_j.$$

Complexity Analysis Let d be the input feature dimension, and assume a hidden dimension of h. The computational complexity of an MLP is $\mathcal{O}(d \cdot h)$, whereas the complexity of the self-attention mechanism is $\mathcal{O}(d^2)$. Although the attention mechanism has a quadratic complexity with respect to input size, the added computational cost is justified by the superior feature interaction modeling capabilities and improved generalization.

Given an arbitrary function $f : \mathbb{R}^d \to \mathbb{R}^d$, it has been demonstrated that a multihead-attention mechanism, when equipped with a sufficient number of heads and layers, can approximate f with lower sample complexity than a Multilayer Perceptron (MLP) of comparable depth. Specifically, for any $\epsilon > 0$, the approximation error of attention-based models satisfies the following bound:

$$\sup_{x \in \mathcal{X}} |f(x) - f_{\text{attention}}(x)| \leq \epsilon.$$

Moreover, the scaling of the approximation error for attention-based models is more favorable than that of MLPs. In particular, the error for MLPs decreases at a rate of $\mathcal{O}\left(\frac{1}{\sqrt{d}}\right)$, whereas the error in attention-based models diminishes at a faster rate of $\mathcal{O}\left(\frac{1}{d}\right)$:

$$\text{Error}_{\text{MLP}} \sim \mathcal{O}\left(\frac{1}{\sqrt{d}}\right), \quad \text{Error}_{\text{Attention}} \sim \mathcal{O}\left(\frac{1}{d}\right)$$

These results indicate that Transformers, which rely on self-attention mechanisms, require fewer parameters to achieve a similar level of approximation accuracy compared to MLPs. This property highlights the parameter efficiency of attention-based models, making them particularly well-suited for high-dimensional input spaces.

Empirical Validation To empirically validate the advantages of attention-based feature fusion, we conducted experiments comparing the Transformer and MLP architectures in terms of model accuracy and loss convergence. The results in Table 3 demonstrate that the attention-based model achieves significantly better performance, supporting the theoretical findings.

Conclusion The self-attention mechanism provides a mathematically superior alternative to MLP-based fusion by offering:

- **Dynamic feature weighting** that adjusts to the importance of different drug and protein features,
- **Efficient long-range dependency modeling** without the need for exponentially large parameters,
- **Stronger generalization abilities** due to better function approximation properties.

Thus, replacing MLPs with attention-based mechanisms in the proposed DTI model enhances feature interaction learning, ultimately leading to improved prediction performance.

Given the above theoretical and empirical justifications, the Transformer-based fusion approach provides a more effective solution for capturing drug-target interactions than MLPs. The proposed model capitalizes on attention mechanisms to dynamically learn intricate feature relationships, leading to improved interaction prediction performance.