Enhancing drug-target affinity prediction through pre-trained language model and gated multi-head attention

Ghina Khoerunnisa [a,1,*], Isman Kurniawan [a,2]

[a] School of Computing, Telkom University, Sukapura, Kec. Dayeuhkolot, Kabupaten Bandung, Jawa Barat, Indonesia
[1] ginnkh@student.telkomuniversity.ac.id; [2] ismankrn@telkomuniversity.ac.id
* corresponding author

ARTICLE INFO

Article history
Received December 20, 2024
Revised February 14, 2025
Accepted March 02, 2025
Available online March 22, 2025

Keywords
Drug target affinity
Pre-trained language model
Gated multi-head attention
Deep learning
Regression

ABSTRACT

Drug development requires accurate drug-target interaction (DTI) information to evaluate a drug's potential. However, existing current methods for estimating DTI are slow and expensive. Deep learning offers an efficient and effective alternative by leveraging sequence data for prediction. Nevertheless, the DTI binary classification approach suffers from a large number of non-interacting pairs, resulting in data imbalance and has a negative impact on performance. To address this issue, DTI is modeled as a regression problem known as drug-target affinity (DTA), which predicts the strength of interactions. While various deep learning methods show competitive results in DTA prediction, they face a challenge in capturing specific drug-target patterns with limited data. To overcome the problem, this study leverages pre-trained language models for enhanced representation. Also, we utilize gated multi-head attention (GMHA), which modifies multi-head attention by including dynamic scaling and a gate process to capture the mutual interactions better. The results show that our proposed method exceeds the benchmark and baseline in all evaluation metrics, with concordance index (CI) of 0.893 and 0.872, and modified r-squared (r_m^2) of 0.673 and 0.723 in Davis and KIBA. Our findings further suggest that pre-trained language models for drug and target receptor representation improve DTA prediction model performance. Also, the GMHA method generally outperforms the simple concatenation method, with more obvious advantages in more complex datasets like KIBA. Our approach provides a competitive enhancement in DTA prediction, suggesting a promising direction for further enhancing drug discovery and development processes.

1. Introduction

Drug target interaction (DTI) refers to an interaction between a drug and a protein, known as a target or receptor, in the human body [1]. Understanding these interactions is essential in the drug development process [2]. According to Kim *et al.* [3], there are more than a million drug compounds that have the potential to become new or repurposed drugs. Meanwhile, the success rate in drug development from phase I clinical trials to therapeutic licensure was relatively low, at only 6.2% of 21,143 compounds [4]. Therefore, accurate prediction of the DTI is an important part of drug development in finding candidate compounds at an early stage [5].

One approach to predict the DTI is conventional approaches conducted in the laboratory. Unfortunately, identifying drug and receptor interactions with such approaches is full of challenges, such

as being time-consuming and cost-expensive [6]. One of the alternative methods for predicting drug and receptor interactions is by extracting sequence information from the drug and receptor and then using the information to recognize the relationship between the sequence and interactions of the drug and receptor. Recent studies highlight deep learning as a potential, cost-effective, and time-efficient method for modeling drug-receptor interactions, demonstrating significant success in addressing the complex, non-linear tasks frequently found in biological and chemical processes [7], [8].

Several studies have been performed to predict sequence-based drug and receptor interactions with deep learning by defining the task as a binary classification problem [9]–[11]. However, this approach has several limitations, such as the lack of experimental valid positive data and the abundance of unvalidated negative data. This approach defines the drug and receptor as having no interaction if the interaction data between them does not exist. However, the lack of interaction data does not guarantee no interaction between drug and receptor. As a result, negative data can be much more numerous, creating a significant class imbalance [12]. Similarly, when using benchmark datasets such as Davis and KIBA [13], [14], researchers defined affinity values above a threshold as interacting pairs [15]. This led to a significant volume of negative data and an imbalanced class distribution [9]. This condition has a negative impact on the performance of the prediction model. Furthermore, the affinity value that represents the strength of the drug-receptor interaction is known as continuous data [16]. Therefore, to address these issues, interactions between drugs and receptors should be defined as a regression problem. Hence, the prediction is conducted on the affinity value of the drug and receptor interaction, which is commonly called Drug Target Affinity (DTA).

Several studies have been performed on sequence-based DTA prediction using deep learning. In 2018, Ozturk *et al.* proposed DeepDTA, which uses CNN to model the SMILES representation of the drug and the amino acid sequence of the receptor. The results were evaluated using the concordance index (CI), which measures how well the model can rank relevant interaction pairs. The DeepDTA model showed good evaluation results with 87% and 86% CI values for the Davis and KIBA datasets, respectively [13]. Then, in 2019, Ozturk *et al.* also proposed WideDTA, a model similar to DeepDTA, except they added features besides SMILES and amino acid sequences, such as ligand max common substructure and protein motifs and domains. WideDTA increased the CI value of DeepDTA by one percent in both datasets [17]. In 2022, Ghimire *et al.* modified the CNN structure by inserting self-attention and produced a CI score of 89% in both the Davis and KIBA datasets [18]. Another study was conducted by D'Souza *et al.* by building a DTA sequence-based prediction model with CNN by adding prior transformers to the drug representation, which resulted in a CI score of 86% [19]. Similar studies have been conducted by modifying the CNN architecture and adding features other than sequences, such as fingerprints, to improve the performance of DTA prediction models. In 2022, Chen *et al.* proposed MultiscaleDTA by utilizing multi-scale CNN and self-attention at each CNN layer. This model achieved a CI value of 89% on the Davis and KIBA datasets [20]. Meanwhile, in 2023, Zhu *et al.* [21] proposed FingerDTA, which enriches the representation of drugs and receptors by incorporating molecular fingerprints into their CNN model. This approach also performed well, with a CI value of 89% on the same dataset.

Apart from the representation of drug and receptor, the inter-interaction or mutual interaction between drug and receptor is also important in predicting DTA. The aforementioned studies combine both representations only with a simple concatenation. This method overlooks the aspects of mutual interaction between drug and receptor representations [22]. In 2020, Abbasi *et al.* proposed DeepCDA, which uses CNN-LSTM as a representation method and a two-sided attention mechanism to achieve mutual interaction. This method resulted in competitive performance, with a CI value of 89% and an r_m^2 value of 64% [23]. In 2023, Zhao *et al.* conducted a similar study by proposing a two-sided attention mechanism to model mutual interactions, namely AttentionDTA, achieving an r^2 value of 74% [14]. In 2021, another study by Zeng *et al.* took a different approach in using attention for mutual interactions, designating drug representations as queries and receptor representations as keys and values, resulting in

a CI of 89% [22]. In 2021, Mahdaddi *et al.* combined SMILES and amino acid sequences before processing them with a CNN-AbiLSTM model, resulting in a CI of 89% and an r_m^2 value of 66% [12].

Although various approaches have been used to predict DTA, two main challenges remain in DTA prediction. First, models such as CNN, CNN-LSTM, and dense layers for representing drugs or receptors are still not effective in capturing specific patterns in drug and receptor sequences, mainly due to the complexity of molecular interactions and the high-dimensional nature of the data. Second, most mutual interaction methods only use simple concatenation without considering the specific mutual relationship between drug and receptor, which is also an important aspect of DTA [24]. As a result, model performance is still not optimal, which leads to room for improvement. Hence, exploring other methods for the representation of the drug and receptor and the mutual interaction model becomes necessary.

Pre-trained language models (PLM), proven powerful in natural language processing (NLP), offer an alternative method for representing drugs and receptors [25]. Pre-trained models such as ChemBERTa-2 and ESM-2 can represent drug molecules on SMILES sequences and proteins on amino acids, respectively [26], [27]. Pre-trained models are useful because they allow the model to use the transferable information encoded in the weights that have been pre-trained with a large amount of data. ChemBERTa-2 was chosen because it was trained with a much larger SMILES dataset compared to other drug molecule PLM models, such as ChemBERTa-1 [28] and MolBERT [29], thus being able to capture molecular representations better. Besides the dataset size, ChemBERTa-2 also shows superior performance compared to ChemBERTa-1 in various SMILES data-driven tasks. In addition, ChemBERTa-2 uses SMILES data from PubChem, which corresponds to the notation used in this study's dataset, ensuring higher compatibility [26]. Meanwhile, ESM-2 was chosen because it has a more complex architecture and a much larger number of parameters than ProtBERT [27], [30], allowing this model to capture the biological features of amino acid sequences more effectively. In addition, ESM-2 was trained using UniRef50, which is more diverse than UniRef100 used by ProtBERT [31], thus improving the generalization ability of the model in understanding receptor characteristics. Pre-trained models can represent sequences effectively, but they only focus on describing interactions within a sequence and ignore interactions between two different sequences, in this case, drug and receptor sequences. Therefore, a multi-head attention mechanism can be used to model the mutual interaction. According to the literature, the attention mechanism can observe the relationship between drugs and receptors simultaneously, thus enabling a more comprehensive understanding of the input data.

This study aims to enhance DTA prediction by addressing two major challenges, which are the limitation of models in capturing the complexity of molecular interactions and the simple concatenation method that ignores the mutual interaction between the drug and the receptor. Therefore, this study implemented two pre-trained language models to obtain drug and receptor representations, ChemBERTa-2 and ESM-2, along with gated multi-head attention, specifically a gated two-sided multi-head cross-attention mechanism (GMHA) to model the mutual interactions between drugs and receptors, allowing the model to simultaneously include within-sequence interactions and interactions between two different sequence types. In this study, GMHA differs from standard multi-head attention mechanisms that rely on fixed scaling by introducing a learnable parameter that makes the scaling process more flexible during training [32]. Also, we added a gate process inspired by the concept of the output gate in the study [33] at the end of the standard multi-head attention to control the proportion of the attention output and input embedding in the final result of GMHA. Finally, this study uses four fully connected layers to predict the DTA.

2. Method

2.1. Dataset

This study used datasets commonly used in DTA prediction studies, namely the Davis and KIBA datasets [14]. Davis and KIBA datasets consist of drug ID, protein or receptor ID, canonical SMILES

sequence, amino acid sequence, and affinity value. The affinity value is a value that represents the strength of the interaction between the drug and the receptor. The measurement variable of the affinity value differs according to the dataset. In the KIBA dataset, affinity measurement is represented by KIBA scores, which statistically combine multiple affinity indicators, including half maximal inhibitory concentration (IC_{50}), inhibition constant (K_i), and dissociation constant (K_d) [34]. On the other hand, the Davis dataset uses K_d values to measure the affinity of the drug and receptor pair [35]. The Davis dataset contains 68 unique SMILES sequences and 365 unique amino acid sequences with 24,956 interactions. Meanwhile, the KIBA dataset has 2068 unique SMILES sequences and 229 unique amino acid sequences with 118,254 interactions. The KIBA dataset contains more data, and the SMILES and amino acid sequences are longer than the Davis dataset.

We performed data preprocessing before developing the DTA prediction model, such as removing missing values and duplicate data. Then, the affinity value (K_d) of the Davis dataset is converted into log space (pK_d), as formulated in Eq. (1), to reduce the large variance inherent in these values. Next, the dataset was divided into train and test sets with a ratio of 80:20. The train set is used to build the DTA prediction model, while the test set is used to evaluate the model.

$$pK_d = -log10\frac{K_d}{10^9} \tag{1}$$

2.2. DTA Prediction Model Framework

The overall framework of our proposed method is illustrated in Fig. 1. The input data consists of SMILES sequences representing drugs and amino acid sequences representing receptors. The first step is the tokenization of both drug and receptor sequences. The tokenization process transforms the raw drug and receptor sequences into a suitable format that our model's subsequent components can process effectively. Then, the results are processed and sent forward to ChemBERTa-2 and ESM-2 encoder models to obtain meaningful embeddings of the drug (E_d) and receptor (E_r). The embeddings are vector representations of each drug or receptor. After obtaining the embeddings for both the drug and the receptor, we derived their respective representations.

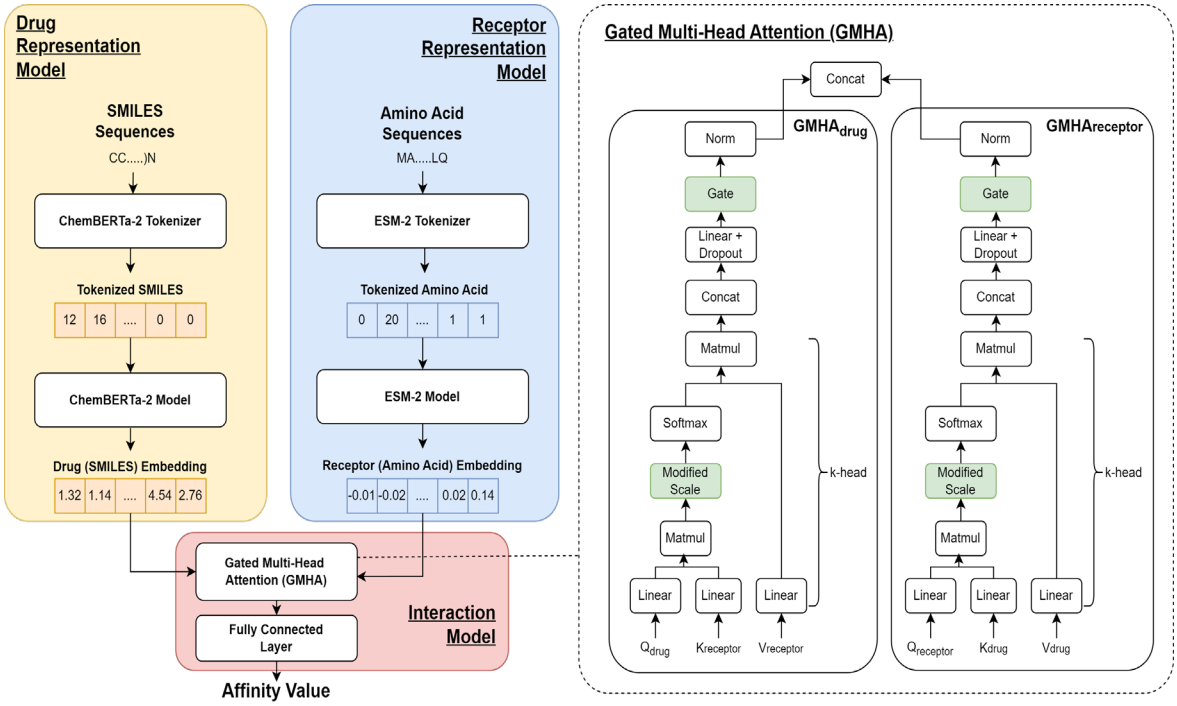

Fig. 1. Illustration of DTA prediction model framework

The next step is defining the interaction model between the drug and the receptor using a gated two-sided multi-head cross-attention mechanism (GMHA). This attention mechanism lets the model

look at drug and receptor embeddings simultaneously, showing how they depend on and affect each other in complex ways. The final process involves feeding the result of the attention mechanism into a fully connected layer to predict the DTA.

The first part of our proposed method framework consists of drug and receptor representation models. The drug and receptor structures were represented using ChemBERTa-2 and ESM-2, respectively. The ChemBERTa-2 is a pre-trained model that has been trained on a large dataset consisting of 77 million SMILES sequences to obtain molecular fingerprints, which is developed based on the RoBERTa model. ChemBERTa-2 has been demonstrated to improve the ability to capture nuanced chemical characteristics, resulting in greater performance across 6 of the 8 SMILES-based tasks. Furthermore, it exhibits high competitiveness in almost all tasks [26]. The ESM-2 model, similar to ChemBERTa-2, is a pre-trained model specifically designed for learning protein sequences. ESM-2 is a protein language model that has been pre-trained on the large Uniref50 and UniRef90 protein sequence datasets, which allows it to understand the complex interactions between proteins [27].

Each representation model consists of a tokenization process and an embedding vector generation process. In tokenization, all characters of each sequence are used by adding padding for sequences to make the length of a shorter sequence become similar to the length of the longest sequence in the data. The tokenizer in ChemBERTa-2 treats sequences as a series of hybrid between character and word-level representations. Therefore, a structure like “Cl” (chlorine) will be tokenized into “Cl”. Meanwhile, ESM-2 treats sequences as a series of characters, so the amino acid sequence MKAV will be broken down into the characters “M”, “K”, “A”, and “V”. To illustrate the process, let us consider the SMILES sequence of "CC...)N" with a length of 55. Using the ChemBERTa-2 tokenizer, the sequence is represented as "12, 16, ..., 0, 0", where the first number represents the start of the sequence; the next numbers represent the SMILES sequence itself, and the number 0 represents padding since the longest SMILES sequence in the dataset is 92 in length. Similarly, the amino acid sequence is processed with the ESM-2 tokenizer. Suppose “MA...LQ” is an amino acid sequence of length 472. Then, after being processed, it is converted to "0, 20, ..., 1, 1". The first number represents the start of the amino acid sequence, and the next number represents the amino acid sequence itself. Then, the number 1 represents padding since the longest amino acid in the dataset is 2549 in length.

Following the tokenization process, the tokenized sequences move through the drug and receptor encoder models, leveraging the capabilities of ChemBERTa-2 and ESM-2, respectively. These models transform the tokenized sequences into embeddings of drugs and receptors. The embeddings capture complex patterns and correlations in the data by encapsulating the important information and properties of the drug and receptor sequences [36]. These embeddings capture the learned representation of each token using the model's pre-trained weights. The result is taken as the average value of all tokens to produce one feature vector that represents the entire compound or molecule. This stage is critical for extracting important information from the input sequences. The vector lengths of drug and receptor embeddings are 384 and 1280, respectively.

Following the extraction of the drug and receptor embedding, the subsequent task is to determine the mutual interaction between the drug and receptor by utilizing a gated multi-head attention (GMHA), which is a modified multi-head attention, specifically two-sided multi-head cross-attention, by adding learnable scaled factor and gate processes. Before that, each embedding result is projected into 256 dimensions. GMHA receives input in the form of a query (Q), key (K), and value (V). Firstly, linear projection on the drug and receptor embedding vectors (E) is used to get the drug and receptor's query, key, and value vectors. Then, use separate weight matrices (W) for each operation with the formula shown in Eq. (2). In this process, we use k equal to 8 and 16 heads on Davis and KIBA, respectively. In every head (H), feed the corresponding Q, K, and V into the attention mechanism with the formula in Eq. (3), where d is the dimension of Q, K, and V to obtain the attention weights and λ is a trainable scale factor used to adjust the level of scaling throughout the training process. Unlike the standard scale factor in multi-head attention, the square root of d, the λ parameter is trained along with other parameters to prevent the attention scores from becoming too small and make the scaling more flexible. If the scores are too small, the softmax distribution becomes nearly uniform, making the model unable

to focus on important features [32]. Next, the results in every head are linearly combined, adding the output of the attention mechanism to the original input, as shown in Eq. (4) [37]. We used the dropout to prevent overfitting. Next, instead of returning the attention result, we used a gate mechanism to control the attention output and original input embedding proportion. This process is based on a gated mechanism in the study [33], which organizes the flow of information from multiple sources. Although the implementations differ in form, the basic idea is similar, which allows the model to adjust the contributions from the attention output and the input embedding. This mechanism is crucial due to the potential loss of important information from the original input embedding during the process. By altering the proportion, we ensure the final result includes essential information from the original input embedding and the attention output. Finally, we performed a normalizing process on the results to produce a final output of the modified multi-head attention layer or $GMHA(Q,K,V)$ [37]. This process results in two outputs of attention: the drug ($GMHA_{drug}$) and the receptor ($GMHA_{receptor}$) attention results. To proceed to the next process, both outputs are concatenated.

$$Q_i = E \cdot W_i^Q; K_i = E \cdot W_i^K; V_i, i = 1, \dots, k \tag{2}$$

$$H_i = attention(Q_i, K_i, V_i) = softmax\left(\frac{Q_i K_i^T}{\lambda\sqrt{d}}\right) V_i, i = 1, \dots, k \tag{3}$$

$$GMHA\ (Q,K,V) norm(gate\big(dropout(concat(H_i, \dots, H_k)W^o)\big)) \tag{4}$$

The final process involves feeding the result of gated multi-head attention into a fully connected layer. We train the model for 300 epochs using the mean squared error (MSE) loss function, as it ensures optimization across the full range of affinity values rather than being biased toward the most frequent ones. Given the skewed distribution of affinity values in Davis (centered around 5.0) and KIBA (around 11–12), a regression-based approach is more suitable than classification to capture the continuous nature of drug and receptor interactions. The Adam optimizer is used for training, and we conducted exploratory experiments to tune hyperparameters for optimal performance. The hyperparameters investigated include learning rate {0.01, 0.001, 0.0001}, batch size {128, 256, 512}, and dropout rate {0.1 to 0.4}. The final settings were a learning rate of 0.0001, batch size of 256 for Davis and 512 for KIBA, dropout rate of 0.3, and weight decay of 0.001, which improved stability and reduced overfitting. The formula for the loss function is shown in Eq. (5), where *n* represents the number of samples, *y'* represents the prediction, and *y* represents the actual affinity score.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i' - y_i)^2 \tag{5}$$

2.3. Experimental Design

This study consisted of four main experiments, i.e., (a) parameter exploration, (b) drug and receptor representation, (c) interaction modeling, and (d) overall comparison. The parameter exploration aims to observe the effects of various parameters in our proposed method, ChemBERTa-2, ESM-2, and gated Multi-head attention DTA (CEMDTA), on the performance of the DTA prediction model. We considered three parameters involved in the model: head of attention, gate scale for the drug ($gate\_d$), and gate scale for the receptor ($gate\_p$). The details of the parameter values used in the experiment are shown in Table 1. We investigated the contribution of those parameters sequentially to determine the optimal combination. Firstly, we conducted experiments for the head parameter while setting the gate scale for both drug and receptor to 0.5. Secondly, the best head value obtained from the previous experiment was used to explore different gate scale values for the drug parameter. Thirdly, using the best head and $gate\_d$ value, we experimented to determine the optimal gate scale for the receptor parameter. Finally, the best model is obtained from the combination of the optimal value of the head, gate scale for the drug ($gate\_d$), and gate scale for the receptor ($gate\_p$), which resulted in the best performance, as evaluated by r_m^2.

Table 1. The details of the parameter exploration experiment

Parameter	Range of Values	Description
Head (H)	[2, 4, 8, 16]	The number of attention heads in a GMHA refers to the parameter that determines the number of parallel attention heads.
Gate scale for the drug (gate_d)	[0.1, 0.3, 0.5, 0.7, 0.9]	Gate scale or contribution control value for the drug side as a query in GMHA
Gate scale for the receptor (gate_p)	[0.1, 0.3, 0.5, 0.7, 0.9]	Gate scale or contribution control value for the receptor side as a query in GMHA

In the drug and receptor representations experiment, we evaluated the method used to represent the drug and receptor sequences. We examined common methods of representing drugs and receptors in DTA prediction topics, such as CNN, LSTM, and CNN-LSTM. Then, we compared those common methods with our proposed method that uses pre-trained language models, specifically ChemBERTa-2 and ESM-2. To ensure a fair comparison, we also included another PLM used by Kang *et al.* [38], namely ChemBERTa-1 (2020) and ProtBERT. Each representation method in this experiment uses the same interaction model as the proposed method GMHA, which allows us to compare the representation methods fairly and consistently. This experiment investigates the impact of different representation methods of drugs and receptors on the performance of DTA prediction. The summary of the comparison of the drug and receptor representation method is presented in Table 2.

Table 2. The summary of the drug and receptor representation method experiment

Model	Drug Repr.	Receptor Repr.	Interaction Model
CNN + GMHA	CNN	CNN	GMHA
LSTM + GMHA	LSTM	LSTM	
CNN-LSTM + GMHA	CNN-LSTM	CNN-LSTM	
ChemBERTaProtBERT + GMHA	ChemBERTa-1 (2020)	ProtBERT	
CEMDTA	**ChemBERTa-2**	**ESM-2**	

We also investigated interaction modeling, focusing on the interactions between drugs and receptors. We evaluated two approaches, namely concatenation, a common method for modeling mutual interaction, and our proposed GMHA mechanism, specifically gated two-sided multi-head cross-attention. We performed this experiment using CNN, LSTM, CNN-LSTM, and the ChemBERTa-2, along with ESM-2 as drug and receptor representations. The interaction modeling aims to determine the effectiveness of GMHA in capturing the mutual interaction relationship between drug and receptor. To further validate GMHA's performance, we conducted an ablation study on the KIBA dataset by comparing GMHA (cross and two-sided) with No Attention, Self-Attention (cross and two-sided), and Multi-Head Attention (cross and two-sided). The summary of the interaction modeling is shown in Table 3.

Table 3. The summary of the interaction modeling experiment

Model	Drug Repr.	Receptor Repr.	Interaction Model
CNN + Concat	CNN	CNN	Concat
CNN + GMHA			GMHA
LSTM + Concat	LSTM	LSTM	Concat
LSTM + GMHA			GMHA
CNN-LSTM + Concat	CNN-LSTM	CNN-LSTM	Concat
CNN-LSTM + GMHA			GMHA
CECDTA	**ChemBERTa-2**	**ESM-2**	Concat
CEMDTA			**GMHA**

Finally, we conduct the method comparison experiment to evaluate CEMDTA against the benchmark models, i.e., AttentionDTA and GraphDTA. AttentionDTA is one of the state-of-the-art models in DTA prediction proposed by Zhao *et al.* in 2023. The model uses CNN to represent drug and receptor sequences, two-sided multi-head attention (MHA), and fully connected layers to model

interactions and predict DTA. Here, the attention outcomes proceed to max pooling separately [14]. Meanwhile, another state-of-the-art DTA, GraphDTA, which was originally proposed by Nguyen *et al.*, uses a graph neural network to process SMILES sequences as a graph representation. In this experiment, we re-run GraphDTA using the graph isomorphism network (GIN) architecture that gives the best performance score in the related paper [39]. For a fair comparison, we adapted and re-executed the publicly available models, adjusting them to be compatible with our dataset while using the original settings described in their respective papers.

Also, we compare the proposed method with several baseline models, i.e., CNN, LSTM, and CNN-LSTM. As for all baseline models, we also considered combining the concatenation and fully connected layers part into the main algorithm. The deep learning method used as the baseline is also frequently used in DTA-related studies. Methods such as CNN and CNN-LSTM are methods that are quite often used to represent drug and receptor sequences. In the literature survey, there are more than 5 studies that use CNN as the main method of drug or receptor representation, with various modifications [13], [14], [17]–[21]. Meanwhile, for CNN-LSTM, there are more than a few studies that at least use the method for drug or receptor representation [12], [13]. Furthermore, we also consider LSTM to be a baseline because it is excellent at capturing long-term dependencies in sequential data. As for interaction modeling, this baseline method uses concatenation and fully connected layers, which is conducted by most current DTA studies. This experiment examines the proposed method's effectiveness compared to these well-known approaches in the DTA prediction model. The summary of the method comparison is shown in Table 4.

Table 4. The summary of the method comparison experiment

Model	Drug Repr.	Receptor Repr.	Interaction Model
AttentionDTA (benchmark)	CNN	CNN	MHA
GraphDTA (benchmark)	GIN	CNN	Concat
CNN + Concat (baseline)	CNN	CNN	Concat
LSTM + Concat (baseline)	LSTM	LSTM	Concat
CNN-LSTM + Concat (baseline)	CNN-LSTM	CNN-LSTM	Concat
CEMDTA	**ChemBERTa-2**	**ESM-2**	**GMHA**

2.4. Evaluation Metrics

In the evaluation process, we measure the model's performance using various metrics. The three performance metrics frequently used in DTA are mean squared error (*MSE*), concordance index (CI) proposed by Gönen *et al.* [40], and regression toward the mean or modified r-squared (r_m^2) proposed by Roy *et al.* [41]. MSE is one of the most common metrics for evaluating regression models. It calculates the average sum of squares of the difference between predicted values y' and actual values y. The smaller the MSE value, the more robust the regression model [16]. MSE is defined as shown in Eq. (5).

Meanwhile, the concordance index (CI) measures the ranking performance of the models that output continuous values [40]. The formula of CI is given in Eq. (6), where b_x is the predicted value for a larger affinity δ_x and b_y is the predicted value for smaller affinity δ_y. The value of Z denotes a normalized constant, and $h(m)$ is a step function [17]. The value for $h(m)$ is defined as shown in Eq. (7). CI has a range of values between 0.5 and 1, with 1 indicating a perfect prediction and 0.5 indicating a random predictor [18].

$$CI = \frac{1}{Z}\sum_{\delta_x>\delta_y} h(b_x - b_y) \tag{6}$$

$$h(m) = \begin{cases} 1 & if \quad m > 0 \\ 0.5 & if \quad m = 0 \\ 0 & if \quad m < 0 \end{cases} \tag{7}$$

The modified r-squared (r_m^2) metric was utilized as another measurement metric to strengthen the evaluation of our model. The r_m^2 is a modified squared correlation coefficient formulated in Eq. (8). The

variables r^2 and r_0^2 indicate the squared correlation coefficient between the predicted and actual values with intercept and without intercept, respectively. The key advantage of r_m^2 is that it provides a more robust measure of predictive ability under different conditions, as it reduces the potential bias introduced by the intercept. Unlike r^2, which can be overestimated due to the intercept influence, r_m^2 gives a more fair and stable evaluation of model performance. A satisfactory model should have a r_m^2 value that exceeds 0.5 [22]. Higher values for both CI and r_m^2 imply better performance.

$$r_m^2 = r^2 \times \left(1 - \sqrt{r^2 - r_0^2}\right) \tag{8}$$

These three metrics provide valuable insight into the model's performance in real-world drug discovery. MSE measures the average squared error, which indicates how close the model predictions are to the actual affinities. CI evaluates the ranking ability of the model, which is useful for prioritizing drug candidates based on their predicted binding strength or affinity. Meanwhile, r_m^2 ensures the model maintains consistent predictive power with and without intercepts. Combining these metrics makes the evaluation more comprehensive, covering both the correctness and ranking ability, which is crucial for real-world drug discovery applications.

3. Results and Discussion

3.1. The Exploration of Parameter Impact

We explored three main parameters in our proposed method, in which the results for both Davis and KIBA datasets are presented in Table 5 and Table 6, respectively.

Table 5. The evaluation results of the parameter exploration experiment on the Davis dataset

Parameter	Value	MSE	CI	r_m^2
Head (H)	2	0.210	0.891	0.663
	4	0.209	0.892	0.670
	8	**0.209**	**0.893**	**0.673**
	16	0.206	0.892	0.670
Gate scale for drug (Gate_d)	0.1	0.212	0.890	0.656
	0.3	0.213	0.891	0.666
	0.5	**0.209**	**0.893**	**0.673**
	0.7	0.215	0.893	0.655
	0.9	0.223	0.887	0.656
Gate scale for receptor (Gate_p)	0.1	0.207	0.892	0.670
	0.3	0.207	0.891	0.666
	0.5	**0.209**	**0.893**	**0.673**
	0.7	0.206	0.893	0.667
	0.9	0.213	0.890	0.656

Subsequently, we explored the gate scale for drug ($gate\_d$) and receptor ($gate\_p$) parameters, which control the proportion of attention result and original input embedding on the drug side as query and receptor side as query. The impact of $gate\_d$ and $gate\_p$ parameters in both datasets show a similar trend, whereas small values lead to poor information utilization, while too large values lead to overfitting. However, in the KIBA dataset, the decrement is sharper, which indicates that the model is more sensitive to changes in gate values. $gate_d$and $gate\_p$ values of 0.5 consistently provide optimal performance on both datasets, retaining enough information without losing the generalizability of the model. According to the results, the best value based on the r_m^2 performance for the head parameter in the Davis dataset is 8, while in KIBA, it is 16. Then, the best gate scale value for the drug and receptor is the same in both datasets, which is 0.5. Combining these parameters resulted in the best model with r_m^2 of 0.673 and 0.723 for Davis and KIBA, respectively.

Table 6. The evaluation results of the parameter exploration experiment on the KIBA dataset

Parameter	Value	MSE	CI	r_m^2
Head (H)	2	0.163	0.877	0.685
	4	0.164	0.875	0.661
	8	0.166	0.876	0.652
	16	**0.166**	**0.872**	**0.723**
Gate scale for drug (Gate_d)	0.1	0.170	0.872	0.689
	0.3	0.162	0.878	0.665
	0.5	**0.166**	**0.872**	**0.723**
	0.7	0.167	0.874	0.638
	0.9	0.169	0.868	0.556
Gate scale for receptor (Gate_p)	0.1	0.164	0.876	0.674
	0.3	0.165	0.878	0.617
	0.5	**0.166**	**0.872**	**0.723**
	0.7	0.163	0.877	0.665
	0.9	0.164	0.874	0.580

After the exploration was completed, the performance of our best model for each dataset was evaluated using the three primary metrics, namely MSE, CI, and r_m^2. The best model shows an MSE value of 0.209 for the Davis dataset and 0.166 for KIBA, reflecting a relatively low prediction error compared to other models. The fairly low error means that our model is better at predicting the actual value of affinity. Nonetheless, MSE alone is insufficient to evaluate the model's excellence, as it lacks a definitive range of values that define a model as excellent. Therefore, we added a concordance index (CI) metric to comprehensively view our model performance in predicting DTA. Our best model achieved a CI value of 0.893 on the Davis dataset, while on the KIBA dataset, it reached 0.872. These values are fairly high and indicate that our model can consistently rank stronger interactions over weaker ones. Moreover, the r_m^2 metric is crucial for assessing how well the model captures the relationship between predicted and actual binding affinity. Our best model achieves an r_m^2 score of 0.673 on Davis and 0.723 on KIBA. These results indicate that the model fairly well identifies biological patterns underlying drug-receptor affinity, leading to predictions that align with actual values. While not in the excellent range, the r_m^2 score suggests that our model remains competitive.

Fig. 2 compares the training loss and validation loss curves of the best model (CEMDTA). The left chart is for the Davis dataset, while the right is for the KIBA dataset. In the loss curves chart on the Davis dataset, it can be seen that the training loss decreases significantly to below 0.1, while the validation loss starts to stabilize around 0.18 to 0.19 after about 200 epochs. However, a gap of about 0.1 between the training loss and validation loss at the near end of the training indicates that the model is slightly overfitting. To address overfitting, we have applied several techniques, including adding dropout layers at multiple stages—after drug and receptor feature embedding, within the GMHA stage and in each hidden layer of the fully connected layer. Additionally, we implemented weight decay with a value of 0.001. Despite these efforts, a small degree of overfitting remains in the Davis dataset. In contrast, in the right-hand chart (KIBA), which has a larger amount of data, the difference between training loss and validation loss is much smaller than the loss curves on the Davis dataset. The validation loss also shows a more steady decrease and is closer to the training loss. This indicates that with a larger dataset, the model can improve its generalization ability with the validation data. These results show that the dataset size has contributed to reducing overfitting, which is relevant as the model is based on deep learning, which has the advantage of large data for improved performance. Based on these observations, further exploration of additional mitigation strategies may be valuable. For example, data augmentation techniques for the SMILES representation, such as SMILES randomization that shuffles the sequence but retains the original molecules, may be considered in future studies to improve generalizability. Validation loss in both datasets shows stability after about 200 epochs.

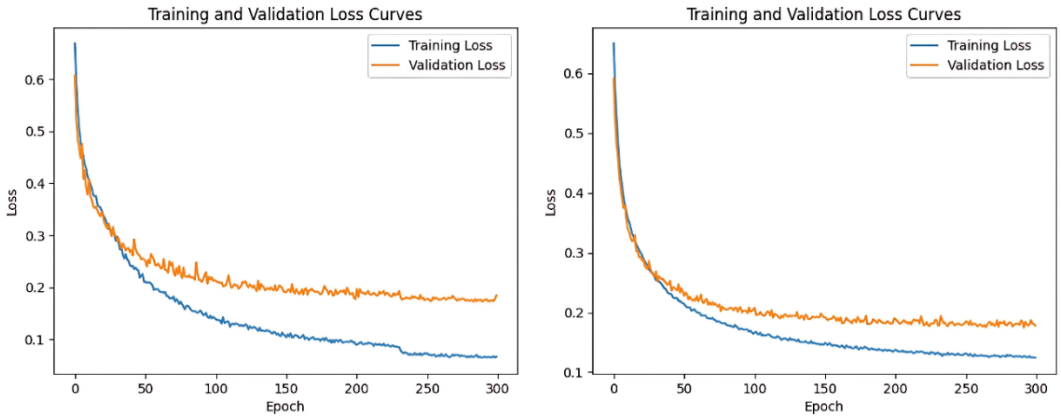

Fig. 2. The training and validation loss curves of the best model on both datasets, Davis (left) and KIBA (right)

3.2. The Drug and Receptor Representation

We compared the drug and receptor representation of our proposed approach, which utilized ChemBERTa-2 and ESM-2, with the common methods used in DTA studies and another PLM (ChemBERTa-1 and ProtBERT). Fig. 3 presents the results of the drug and receptor representation method comparison.

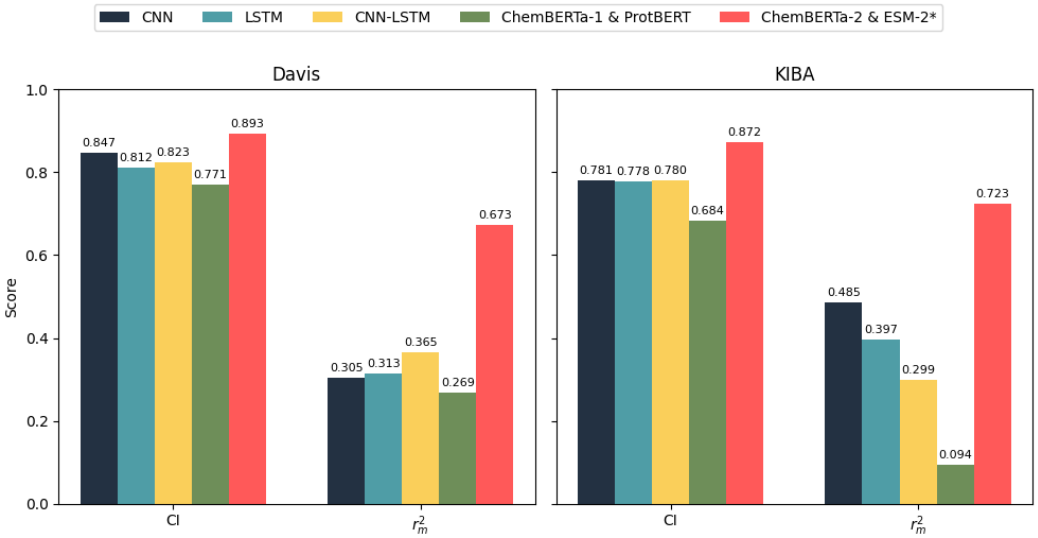

Fig. 3. The evaluation results of the comparison of the proposed drug and receptor representations with other representations in predicting DTA (the proposed representation method is marked with *)

It shows that all models achieved fairly good CI scores, exceeding 0.7 on Davis and 0.6 on KIBA. Nevertheless, our proposed method consistently outperformed other representation methods, with the highest CI score of 0.893 in Davis and 0.872 in KIBA. In the r_m^2 metric, the improvements were even more significant, with our model showing a substantial advantage over the other methods. Particularly, the ChemBERTa-2 and ESM-2 achieved the highest overall scores, while the models with the lowest overall scores were ChemBERTa-1 and ProtBERT on both datasets.

The superior performance of our approach is primarily attributed to the richness of the representation of the ChemBERTa-2 and ESM-2 models. Ahmad *et al.* [26] and Lin *et al.* [27] previously trained the ChemBERTa-2 and ESM-2 models, respectively, using large datasets with the same type (SMILES and amino acid sequences) as ours. Their excellent performance indicates that these pre-trained models leverage relevant and insightful information that enhances DTA prediction performance. The 5.4% and 84.4% improvements in the CI and r_m^2 values, respectively, compared to the CNN method for CI and CNN-LSTM for r_m^2, further demonstrate the effectiveness of our model on Davis. Similarly, on KIBA, our model improved by 11.6% in CI and 49.1% in r_m^2 compared to the second-best CNN model. This

significant enhancement underscores the benefits of leveraging pre-trained models to improve prediction results.

3.3. The Interaction Modeling

We compared the interaction modeling between concatenation and our proposed gated multi-head attention (GMHA) to model the mutual interaction of drug and receptor. Fig. 4 shows the comparison results of the proposed interaction model, a gated two-sided multi-head cross-attention or GMHA with a concatenation-only method.

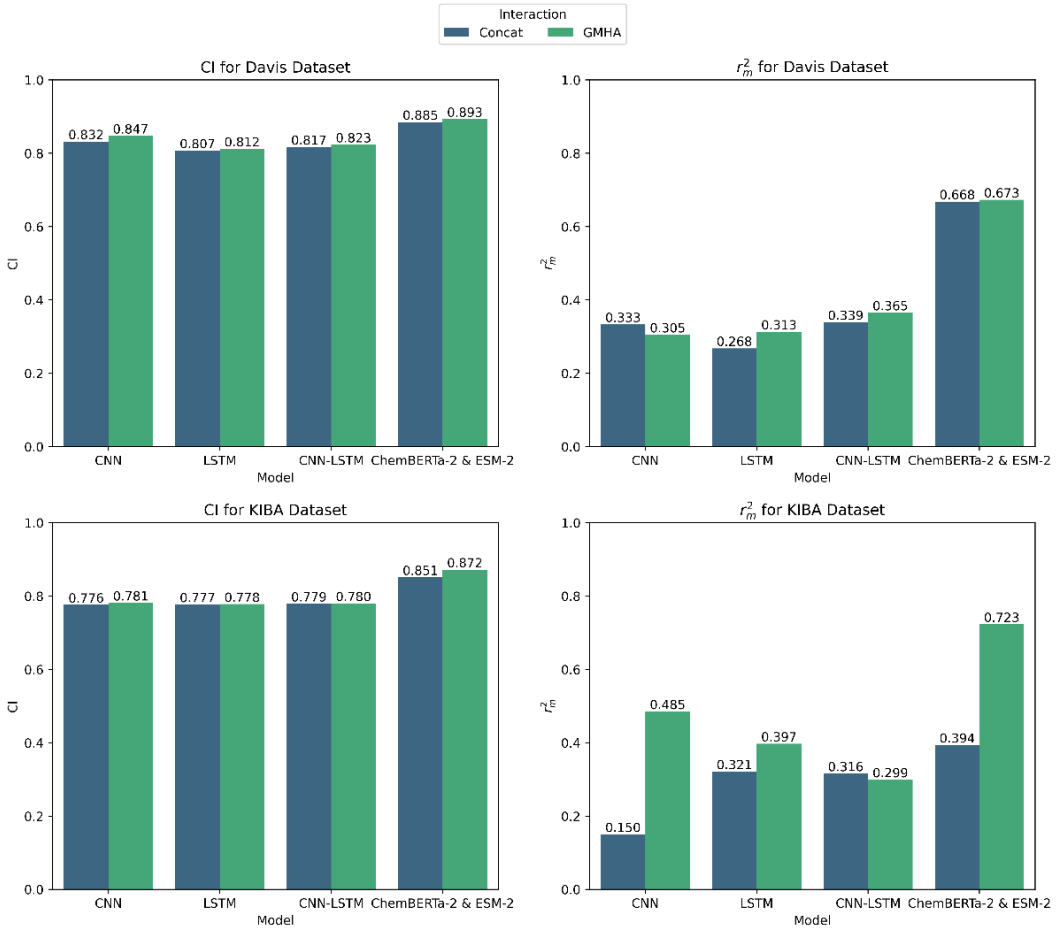

Fig. 4. The evaluation results of the interaction modeling experiment on the Davis and KIBA dataset

Based on the results, we found that GMHA outperformed the concatenation method across most metrics and representation methods for both the Davis and KIBA datasets, with only one model in each dataset showing slightly lower r_m^2 values. For the CI metric, there was a slight improvement in both datasets. Meanwhile, for the r_m^2 metric, on the Davis dataset, GMHA improved the score slightly, while on KIBA, the improvement was more significant. When GMHA was applied to the pre-trained models like ChemBERTa-2 and ESM-2 on Davis, it increased the r_m^2 score from 0.668 to 0.673, indicating a minor enhancement. This suggests that while GMHA can enhance the ability to capture complex mutual interactions, its impact is limited on smaller datasets like Davis. Moreover, the CI score of approximately 0.89 achieved by the concatenation method and GMHA highlights that the pre-trained model already contains most of the necessary information for affinity prediction. However, the benefits of GMHA are more pronounced on larger datasets with longer sequences, such as KIBA, where it increased the CI from 0.851 to 0.872 and r_m^2 from 0.394 to 0.723. These results demonstrate that GMHA is more effective than concatenation in capturing deep interactions between drugs and receptors on datasets with more complex characteristics.

GMHA provides a more significant improvement on the KIBA dataset than Davis because of the limitations of PLMs such as ChemBERTa-2 and ESM-2 in capturing all the important features of the data. Although all sequences were included in the model without truncation, the original design of the PLM was still built with a default input length limit of 512 for ChemBERTa-2 and 1024 for ESM-2. This capacity of ESM-2 is much shorter than the length of the longest protein sequences in both the KIBA (4128) and Davis (2549) datasets. This indicates that although the model can process longer inputs, its pre-trained architecture is likely not fully optimized to capture important features of long sequences, especially in the KIBA dataset. Under these conditions, GMHA provides the advantage of utilizing an attention mechanism that can capture deep relationships between feature embeddings with a gate mechanism to control the proportion of original input embedding and attention results, thus being able to extract important information that pre-trained models potentially miss. In contrast, a simple concatenation method cannot emphasize important features in long and complex sequences. Without an attention mechanism, the concatenation method treats all features equally, thus not considering the relationships between previously extracted features. As a result, important information in complex interactions can be overlooked, especially in datasets like KIBA, which combine data from Davis, Metz, and Anastassiadis [34], leading to diverse interaction patterns that require a more adaptive method. Therefore, GMHA provides a more effective solution for exploring the complex relationships between drugs and receptors, explaining the greater performance improvement on more complex datasets such as KIBA over Davis.

In addition, Fig. 5 displays the attention heatmap to show how the GMHA captures the relationships between features embedding in Davis (top) and KIBA (bottom). The color difference between the two reflects how the model processes the relationship between features. In Davis, attention is spread more evenly, while in KIBA, attention is more focused on specific features, showing a more heterogeneous pattern. The more sparse and specific attention structure of KIBA indicates that GMHA can capture more important features in this dataset.

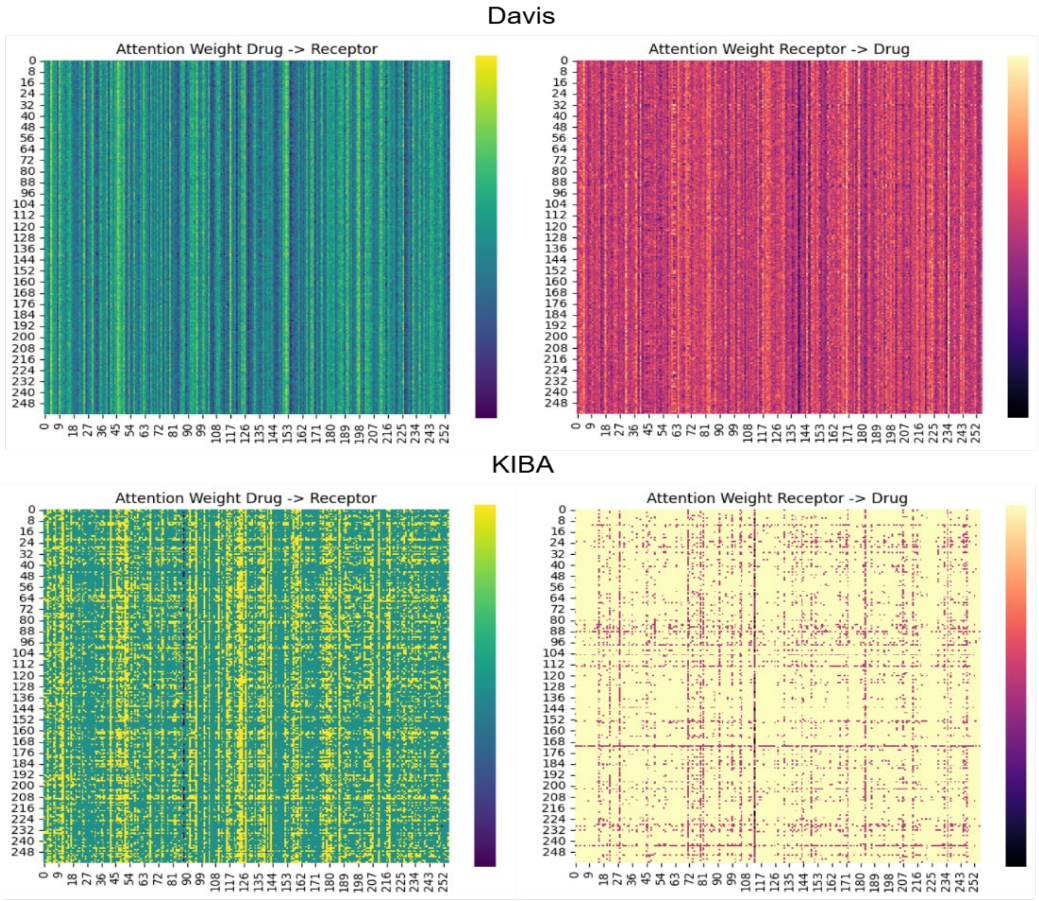

Fig. 5. The attention heatmaps on the Davis and KIBA dataset

Furthermore, to analyze the effectiveness of GMHA, we conducted an ablation study by comparing GMHA with no attention, self-attention, and multi-head attention. As shown in Table 7, attention mechanisms generally improve the model's performance in predicting DTA. Gated Multi-head Attention (GMHA) is the superior approach, with the highest r_m^2. Models without attention only concatenate features, thus failing to capture complex mutual relationships. Self Attention improves this by considering interactions between features, while Multi-head Attention is superior as it can capture patterns from multiple perspectives in parallel. However, the most significant improvement occurs when the GMHA is applied, which adds a gate mechanism to regulate the proportion of information between attention and input embedding. These results prove that GMHA is more effective in understanding complex mutual relationships between features, thus providing better performance than other methods.

Table 7. The evaluation results of the ablation study of GMHA on the KIBA dataset

Model	Description	r_m^2
No Attention	Concat	0.394
Self Attention	Two Sided Self Cross Attention	0.613
Multi-head Attention	Two Sided Multi-head Cross Attention	0.656
Gated Multi-head Attention	**Gated Two Sided Multi-head Cross Attention**	**0.723**

3.4. The Method Comparison

We evaluated our proposed method, CEMDTA, with the benchmark and baseline models in the method comparison. The results of the method comparison evaluation are presented in Table 8.

Table 8. The evaluation results of the method comparison experiment on the Davis and KIBA dataset

Dataset	Model	MSE	CI	r_m^2
Davis	AttentionDTA (benchmark)	0.219	0.878	0.587
	GraphDTA (benchmark)	0.240	0.871	0.559
	CNN + Concat (baseline)	0.364	0.832	0.333
	LSTM + Concat (baseline)	0.393	0.807	0.268
	CNN-LSTM + Concat (baseline)	0.389	0.817	0.339
	CEMDTA (proposed method)	**0.209**	**0.893**	**0.673**
KIBA	AttentionDTA (benchmark)	0.171	0.873	0.595
	GraphDTA (benchmark)	**0.166**	**0.875**	0.606
	CNN + Concat (baseline)	0.369	0.776	0.150
	LSTM + Concat (baseline)	0.381	0.777	0.321
	CNN-LSTM + Concat (baseline)	0.402	0.779	0.316
	CEMDTA (proposed method)	**0.166**	0.872	**0.723**

CEMDTA achieved the best performance on the Davis dataset in all metrics, with the lowest MSE and the highest CI and r_m^2 values. On the KIBA dataset, although CEMDTA shows the best performance in terms of MSE and r_m^2, its CI value is slightly lower than the GraphDTA, differing by only a small margin (0.872 compared to 0.875), which indicates a competitive ranking performance. These results highlight the overall effectiveness of our proposed method in predicting DTA based on sequences, achieving superior performance across evaluation metrics on the Davis dataset, and demonstrating competitive results on the KIBA dataset. This improvement is mainly due to the use of a pre-trained model that is more robust in capturing the features of both drug and receptor sequences, as well as the implementation of gated two-sided multi-head cross-attention (GMHA), which is more effective in modeling their mutual interaction. In contrast, AttentionDTA, another sequence-based model, does not leverage a pre-trained language model or the GMHA mechanism, resulting in sub-optimal feature representation. Nevertheless, AttentionDTA ranks as the second-best model on the Davis dataset, achieving a CI value of 0.878 and an r_m^2 value of 0.587. As for GraphDTA, although it

uses a graph representation to represent the drug, it still relies on the SMILES sequence at an early stage. Graph representation excels in capturing the relationships between atoms. However, our completely sequence-based approach, which works directly at the linear sequence level, can effectively capture the complexity of the relationship between drug compounds and target receptors without the use of graphs. This suggests that our model is superior in DTA prediction effectiveness with a simpler approach than GraphDTA. Nonetheless, on the KIBA dataset, GraphDTA ranked as the second-best model after ours, with a CI value of 0.875 and an r_m^2 value of 0.606.

Furthermore, among the baseline models, CNN with concat performs best on the Davis dataset, achieving MSE, CI, and r_m^2 scores of 0.364, 0.832, and 0.333, respectively. Meanwhile, on the KIBA, the LSTM model with concat emerges as the best baseline, with MSE, CI, and r_m^2 scores of 0.381, 0.777, and 0.321, respectively. However, based on the results, these models show limitations in capturing complex patterns of sequences. Additionally, the concatenation method these models use struggles to effectively model the relationships between drugs and receptors, making it less capable of capturing the critical mutual interactions between them. Therefore, CEMDTA offers a more accurate approach for sequence-based DTA prediction, with better capability in modeling such complex relationships.

The training was performed on a Google Colab Pro with an NVIDIA T4 GPU. Despite using pre-trained models, CEMDTA trains significantly faster than other benchmark models due to the separate feature extraction process for ChemBERTa-2 and ESM-2. CEMDTA takes 2.2 seconds/epoch on the Davis dataset, while AttentionDTA and GraphDTA take 22.5 seconds and 5.3 seconds/epoch, respectively. On KIBA, CEMDTA was trained in 10.3 seconds/epoch, compared to 37.9 seconds and 21 seconds/epoch for the benchmark. Memory usage remains within the 15 GB GPU RAM, making this model computationally efficient.

3.5. Limitations

This study focuses on sequence-based methods for DTA prediction, with AttentionDTA (2023) as the main benchmark. A graph-based model, such as GraphDTA, was included as an additional reference. Furthermore, a model like MolBERT (2021) was not considered due to its focus on DTI classification rather than regression (DTA). The models used for method comparison in this study were selected based on their relevance during the experiments. Models developed or published afterward have not been explored but could be investigated in future studies. Additionally, limited computational resources have constrained further experiments with transformer-based models or other more complex approaches.

This study used two widely used datasets, Davis and KIBA, for training and evaluation. While these datasets are standard benchmarks, they may not fully represent the diversity of real-world drug and receptor affinity. Nevertheless, they provide a strong foundation for evaluating the performance of DTA prediction models. Due to computational constraints, this study was limited to Davis and KIBA. Future studies could explore the impact of incorporating larger and more diverse datasets, such as BindingDB or PDBBind, to further enhance model robustness and generalizability. In addition, interpretability remains a challenge in deep learning-based DTA models. While our approach uses attention mechanisms, it operates on features extracted by PLMs, making it difficult to attribute biological significance to specific features. Improving interpretability in DTA prediction remains an open research direction for future studies.

4. Conclusion

This study aims to develop, evaluate, and analyze a method integrating pre-trained language models, ChemBERTa-2 and ESM-2, for representation of the drug and receptor, respectively, and implemented gated multi-head attention (GMHA) mechanism with dynamic scaling and gate mechanism to regulate attention proportions. Four main experiments, namely parameter exploration, drug and receptor representation, interaction modeling, and method comparison, are performed to evaluate our proposed method. Based on the experimental results, our proposed method has competitive performance. It exceeds the benchmark and baseline models in terms of all evaluation metrics, with CI scores of 0.893 and 0.872 and r_m^2 scores of 0.673 and 0.723 on Davis and KIBA, respectively. This optimal performance

is achieved with a head of 8 on the Davis dataset and 16 on the KIBA dataset and the gate scale for drug and receptor values of 0.5, which balances the capability of extracting information and model generalization. Then, we found that using pre-trained language models for drug and receptor representation improves the DTA prediction model's effectiveness. Furthermore, the implementation of GMHA, two-sided multi-head cross-attention with dynamic scaled and gate process, can improve the overall performance compared to the simple concatenation method. For future research, it is recommended to develop DTA prediction models by utilizing more diverse modalities, such as drug and receptor chemical properties, molecular fingerprints, and graph-based approaches to capture more robust information regarding drug and receptor binding affinities. In addition, while the results are competitive, we should further improve the r_m^2 metric to enhance the model's generalizability.

Declarations

Author contribution. The first author concentrated on conducting the study and writing the manuscript. Meanwhile, the second author is responsible for supervising the course of the study and correcting the manuscript.

Funding statement. The study has received no funding from public, commercial, or non-profit funding agencies.

Conflict of interest. The authors declare no conflict of interest.

Additional information. No additional information is available for this paper.

References

[1] Y.-F. Zhang *et al.*, "SPVec: A Word2vec-Inspired Feature Representation Method for Drug-Target Interaction Prediction," *Front. Chem.*, vol. 7, p. 504142, Jan. 2020, doi: 10.3389/fchem.2019.00895.

[2] H. Khojasteh, J. Pirgazi, and A. Ghanbari Sorkhi, "Improving prediction of drug-target interactions based on fusing multiple features with data balancing and feature selection techniques," *PLoS One*, vol. 18, no. 8, p. e0288173, Aug. 2023, doi: 10.1371/journal.pone.0288173.

[3] S. Kim *et al.*, "PubChem 2019 update: improved access to chemical data," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1102–D1109, Jan. 2019, doi: 10.1093/nar/gky1033.

[4] C. H. Wong, K. W. Siah, and A. W. Lo, "Estimation of clinical trial success rates and related parameters," *Biostatistics*, vol. 20, no. 2, pp. 273–286, Apr. 2019, doi: 10.1093/biostatistics/kxx069.

[5] M. Kalemati, M. Zamani Emani, and S. Koohi, "BiComp-DTA: Drug-target binding affinity prediction through complementary biological-related and compression-based featurization approach," *PLOS Comput. Biol.*, vol. 19, no. 3, p. e1011036, Mar. 2023, doi: 10.1371/journal.pcbi.1011036.

[6] S. Lin, C. Shi, and J. Chen, "GeneralizedDTA: combining pre-training and multi-task learning to predict drug-target binding affinity for unknown drug discovery," *BMC Bioinformatics*, vol. 23, no. 1, p. 367, Sep. 2022, doi: 10.1186/s12859-022-04905-6.

[7] H. Abbasi Mesrabadi, K. Faez, and J. Pirgazi, "Drug–target interaction prediction based on protein features, using wrapper feature selection," *Sci. Rep.*, vol. 13, no. 1, p. 3594, Mar. 2023, doi: 10.1038/s41598-023-30026-y.

[8] L. Douali, "Machine learning for the prediction of phenols cytotoxicity," *Int. J. Adv. Intell. Informatics*, vol. 8, no. 1, p. 58, Mar. 2022, doi: 10.26555/ijain.v8i1.748.

[9] Y. Qian, X. Li, J. Wu, and Q. Zhang, "MCL-DTI: using drug multimodal information and bi-directional cross-attention learning method for predicting drug–target interaction," *BMC Bioinformatics*, vol. 24, no. 1, p. 323, Aug. 2023, doi: 10.1186/s12859-023-05447-1.

[10] Z.-H. Ren *et al.*, "DeepMPF: deep learning framework for predicting drug–target interactions based on multi-modal representation with meta-path semantic analysis," *J. Transl. Med.*, vol. 21, no. 1, p. 48, Jan. 2023, doi: 10.1186/s12967-023-03876-3.

[11] A. Saad, F. A. Maghraby, and Y. M. Omar, "Predicting Drug Target Interaction by Integrating Drug Fingerprint and Drug Side Effect Using Machine Learning," in *Advances in Intelligent Systems and Computing*, vol. 921, Springer, Cham, 2020, pp. 281–290, doi: 10.1007/978-3-030-14118-9_28.

[12] A. Mahdaddi, S. Meshoul, and M. Belguidoum, "EA-based hyperparameter optimization of hybrid deep learning models for effective drug-target interactions prediction," *Expert Syst. Appl.*, vol. 185, p. 115525, Dec. 2021, doi: 10.1016/j.eswa.2021.115525.

[13] H. Öztürk, A. Özgür, and E. Ozkirimli, "DeepDTA: deep drug–target binding affinity prediction," *Bioinformatics*, vol. 34, no. 17, pp. i821–i829, Sep. 2018, doi: 10.1093/bioinformatics/bty593.

[14] Q. Zhao, G. Duan, M. Yang, Z. Cheng, Y. Li, and J. Wang, "AttentionDTA: Drug–Target Binding Affinity Prediction by Sequence-Based Deep Learning With Attention Mechanism," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 20, no. 2, pp. 852–863, Mar. 2023, doi: 10.1109/TCBB.2022.3170365.

[15] T. He, M. Heidemeyer, F. Ban, A. Cherkasov, and M. Ester, "SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines," *J. Cheminform.*, vol. 9, no. 1, p. 24, Dec. 2017, doi: 10.1186/s13321-017-0209-z.

[16] M. A. Thafar, M. Alshahrani, S. Albaradei, T. Gojobori, M. Essack, and X. Gao, "Affinity2Vec: drug-target binding affinity prediction through representation learning, graph mining, and machine learning," *Sci. Rep.*, vol. 12, no. 1, p. 4751, Mar. 2022, doi: 10.1038/s41598-022-08787-9.

[17] H. Öztürk, E. Ozkirimli, and A. Özgür, "WideDTA: prediction of drug-target binding affinity," *arxiv Artif. Intell.*, pp. 1–11, 2019, [Online]. Available at: http://arxiv.org/abs/1902.04166.

[18] A. Ghimire, H. Tayara, Z. Xuan, and K. T. Chong, "CSatDTA: Prediction of Drug–Target Binding Affinity Using Convolution Model with Self-Attention," *Int. J. Mol. Sci.*, vol. 23, no. 15, p. 8453, Jul. 2022, doi: 10.3390/ijms23158453.

[19] S. D'Souza, K. V. Prema, S. Balaji, and R. Shah, "Deep Learning-Based Modeling of Drug–Target Interaction Prediction Incorporating Binding Site Information of Proteins," *Interdiscip. Sci. Comput. Life Sci.*, vol. 15, no. 2, pp. 306–315, Jun. 2023, doi: 10.1007/s12539-023-00557-z.

[20] H. Chen, D. Li, J. Liao, L. Wei, and L. Wei, "MultiscaleDTA: A multiscale-based method with a self-attention mechanism for drug-target binding affinity prediction," *Methods*, vol. 207, pp. 103–109, Nov. 2022, doi: 10.1016/j.ymeth.2022.09.006.

[21] X. Zhu, J. Liu, J. Zhang, Z. Yang, F. Yang, and X. Zhang, "FingerDTA: A Fingerprint-Embedding Framework for Drug-Target Binding Affinity Prediction," *Big Data Min. Anal.*, vol. 6, no. 1, pp. 1–10, Mar. 2023, doi: 10.26599/BDMA.2022.9020005.

[22] Y. Zeng, X. Chen, Y. Luo, X. Li, and D. Peng, "Deep drug-target binding affinity prediction with multiple attention blocks," *Brief. Bioinform.*, vol. 22, no. 5, pp. 1–10, Sep. 2021, doi: 10.1093/bib/bbab117.

[23] K. Abbasi, P. Razzaghi, A. Poso, M. Amanlou, J. B. Ghasemi, and A. Masoudi-Nejad, "DeepCDA: deep cross-domain compound–protein affinity prediction through LSTM and convolutional neural networks," *Bioinformatics*, vol. 36, no. 17, pp. 4633–4642, Nov. 2020, doi: 10.1093/bioinformatics/btaa544.

[24] T. M. Nguyen, T. Nguyen, and T. Tran, "Mitigating cold-start problems in drug-target affinity prediction with interaction knowledge transferring," *Brief. Bioinform.*, vol. 23, no. 4, pp. 1–13, Jul. 2022, doi: 10.1093/bib/bbac269.

[25] B. Min *et al.*, "Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey," *ACM Comput. Surv.*, vol. 56, no. 2, pp. 1–40, Feb. 2024, doi: 10.1145/3605943.

[26] W. Ahmad, E. Simon, S. Chithrananda, G. Grand, and B. Ramsundar, "ChemBERTa-2: Towards Chemical Foundation Models," *arxiv Artif. Intell.*, pp. 1–8, Sep. 2022. [Online]. Available at: https://arxiv.org/abs/2209.01712v1.

[27] Z. Lin *et al.*, "Evolutionary-scale prediction of atomic level protein structure with a language model," *bioRxiv*. Cold Spring Harbor Laboratory, p. 2022.07.20.500902, Jul. 21, 2022, doi: 10.1101/2022.07.20.500902.

[28] S. Chithrananda, G. Grand, and B. R. Deepchem, "ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction," *arxiv Artif. Intell.*, pp. 1–8, Oct. 2020. [Online]. Available at: https://arxiv.org/abs/2010.09885.

[29] B. Fabian *et al.*, "Molecular representation learning with language models and domain-relevant auxiliary tasks," *arXiv*, pp. 1–12, Nov. 2020. [Online]. Available at: https://arxiv.org/abs/2011.13230v1.

[30] A. Elnaggar *et al.*, "ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Learning," *bioRxiv*. Cold Spring Harbor Laboratory, p. 2020.07.12.199554, Jul. 12, 2020, doi: 10.1101/2020.07.12.199554.

[31] B. E. Suzek, H. Huang, P. McGarvey, R. Mazumder, and C. H. Wu, "UniRef: comprehensive and non-redundant UniProt reference clusters," *Bioinformatics*, vol. 23, no. 10, pp. 1282–1288, May 2007, doi: 10.1093/bioinformatics/btm098.

[32] A. Ranjan, M. S. Fahad, and A. Deepak, "Scaled-attention: A novel fast attention mechanism for efficient modeling of protein sequences," *Inf. Sci. (Ny).*, vol. 609, pp. 1098–1112, Sep. 2022, doi: 10.1016/j.ins.2022.07.127.

[33] K. Kurnianingsih *et al.*, "Big data analytics for relative humidity time series forecasting based on the LSTM network and ELM," *Int. J. Adv. Intell. Informatics*, vol. 9, no. 3, p. 537, Nov. 2023, doi: 10.26555/ijain.v9i3.905.

[34] J. Tang *et al.*, "Making Sense of Large-Scale Kinase Inhibitor Bioactivity Data Sets: A Comparative and Integrative Analysis," *J. Chem. Inf. Model.*, vol. 54, no. 3, pp. 735–743, Mar. 2014, doi: 10.1021/ci400709d.

[35] M. I. Davis *et al.*, "Comprehensive analysis of kinase inhibitor selectivity," *Nat. Biotechnol.*, vol. 29, no. 11, pp. 1046–1051, Nov. 2011, doi: 10.1038/nbt.1990.

[36] M. Lee, "Recent Advances in Deep Learning for Protein-Protein Interaction Analysis: A Comprehensive Review," *Molecules*, vol. 28, no. 13, p. 5169, Jul. 2023, doi: 10.3390/molecules28135169.

[37] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017, [Online]. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[38] H. Kang, S. Goo, H. Lee, J.-W. Chae, H.-Y. Yun, and S. Jung, "Fine-tuning of BERT Model to Accurately Predict Drug–Target Interactions," *Pharmaceutics*, vol. 14, no. 8, p. 1710, Aug. 2022, doi: 10.3390/pharmaceutics14081710.

[39] T. Nguyen, H. Le, T. P. Quinn, T. Nguyen, T. D. Le, and S. Venkatesh, "GraphDTA: predicting drug–target binding affinity with graph neural networks," *Bioinformatics*, vol. 37, no. 8, pp. 1140–1147, May 2021, doi: 10.1093/bioinformatics/btaa921.

[40] M. Gönen and G. Heller, "Concordance probability and discriminatory power in proportional hazards regression," *Biometrika*, vol. 92, no. 4, pp. 965–970, Dec. 2005, doi: 10.1093/biomet/92.4.965.

[41] P. P. Roy and K. Roy, "On Some Aspects of Variable Selection for Partial Least Squares Regression Models," *QSAR Comb. Sci.*, vol. 27, no. 3, pp. 302–313, Mar. 2008, doi: 10.1002/qsar.200710043.