Network Modeling Analysis in Health Informatics and Bioinformatics (2025) 14:24
https://doi.org/10.1007/s13721-025-00516-2

ORIGINAL ARTICLE

Optimizing drug-target binding affinity prediction for kinase proteins: a novel all-MLP-transformer approach

Nassima Aleb[1]

Received: 10 June 2024 / Revised: 5 March 2025 / Accepted: 11 March 2025

Abstract

Drug discovery is crucial for identifying therapeutic agents that effectively interact with disease-linked proteins. This interaction, quantified by drug-target binding affinity (DTBA), is a key indicator of a drug's potential efficacy. Traditional methods, like high-throughput screening, are resource-intensive and costly. While machine learning approaches have emerged for DTBA prediction, significant challenges remain. To address these challenges, we introduce TransMLP-DTBA, an innovative All-MLP Mixing based Transformer architecture for DTBA prediction. TransMLP-DTBA overcomes limitations of existing methods by avoiding computationally expensive operations like high-dimensional multiplications and Multihead-Attention. Instead, it employs MLP Mixing, reducing computational demands and streamlining the process, making it suitable for large-scale screenings and drug repurposing. TRansMLP-DTBA leverages 1D sequences of drugs and targets, represented by SMILES strings and protein sequences, respectively, bypassing the need for complex 3D structural data. Extensive testing on public datasets demonstrates TransMLP-DTBA superiority over state-of-the-art methods like DeepDTA and GraphDTA, achieving significantly lower mean square error, higher concordance index, and higher R-squared, highlighting its robustness and predictive accuracy. This research unlocks potential for further advancements in drug discovery, emphasizing the potential of MLP-Mixing based transformer models in biomedical applications.

Keywords Drug-target binding affinity (DTBA) · Deep learning in drug discovery · MLP-mixer transformer · Predictive modeling in biomedicine · High-throughput screening (HTS)

1 Introduction and related work

In the pursuit of novel therapeutic agents, drug discovery fundamentally revolves around identifying molecules capable of effectively binding to specific disease-associated proteins. The strength of this binding, quantified by drug-target binding affinity (DTBA), serves as a crucial indicator of a drug's potential efficacy. Predicting binding affinity between drugs and their targets is essential in accelerating the drug discovery process. Historically, drug discovery has relied on high-throughput screening (HTS) (Shortridge 2008), which, despite its effectiveness, is resource-intensive, time-consuming, and costly (Tavella 2021). To address these challenges, computational methods leveraging existing biochemical data have been developed to predict DTBA, reducing dependence on experimental assays (Parenti 2012; Angelopoulos 2011).

Early computational methods utilized various innovative approaches, including similarity-based approaches (Yamanishi 2008), kernel-driven regression methods (Peng Peng et al. 2015), and semi-supervised frameworks (Chawla 2005). Matrix factorization techniques (Ezzat et al. 2016) and Kronecker Regularized Least Squares (KronRLS) (Pahikkala et al. 2015) emerged as powerful tools for interaction prediction. To improve predictive accuracy, gradient boosting machines like SimBoost (He et al. 2017) were employed. In parallel, docking simulations such as Glide, MOE-Dock, and AutoDock Vina have been widely used, although these methods are computationally expensive and limited to proteins with known 3D structures (Friesner et al. 2004; Eberhardt et al. 2021).

With the advent of machine learning, several ML-based approaches emerged, using algorithms like Random Forest (RF) and Support Vector Machines (SVM) (Ballester et al. 2010; Meli et al. 2021; Shar et al. 2016). Deep learning

✉ Nassima Aleb
alebn@rcjy.edu.sa

[1] Jubail Industrial College, Jubail, Saudi Arabia

Published online: 16 April 2025

techniques, such as DeepDTA (Öztürk et al. 2018), utilized convolutional neural networks (CNNs) with SMILES strings and protein sequences. DeepConvDTI (Lee et al. 2019) replaced CNNs with Multi-Layer Perceptron (MLPs). Further advancements incorporated attention mechanisms, like in DeepCDA (Rifaioglu 2020) and MATT-DTI (Kwon et al. 2020). Transformer-based models, such as DeepPurpose and TransformerCPI (Huang et al. 2021; Chen et al. 2020), demonstrated improved accuracy.

Despite these advancements, current methods face significant limitations in performance, robustness, and scalability (Vamathevan 2019). Machine learning-based methods, such as Random Forest (RF) and Support Vector Machines (SVM), depend heavily on extensive feature engineering and known ligand data (Wang and Zhang 2017). Proposed deep learning (DL) models have improved prediction performance, yet they require large datasets and intensive computing power.

Transformer models, one of the most advanced deep learning architectures are highly effective in capturing long-range dependencies within sequential data and have become a leading approach in various bioinformatics tasks. Their success stems from their ability to process information efficiently, leading to improved accuracy compared to earlier methods. However, the computational demands of standard transformer architectures, particularly those relying on self-attention mechanisms, present scalability challenges. These challenges are multifaceted: the quadratic complexity of self-attention with respect to sequence length directly impacts computational cost and memory requirements, especially when dealing with large datasets of long protein sequences or complex molecules. Furthermore, the need for substantial computational resources (powerful GPUs and large memory capacity) can limit accessibility for researchers with limited infrastructure.

This work builds upon the strengths of the transformer architecture while directly addressing its scalability limitations. We capitalize on the renowned transformer architecture's ability to capture long-range dependencies but mitigate its inherent quadratic complexity and memory intensiveness by significantly reducing computational costs and memory footprint without sacrificing predictive accuracy. This enhanced scalability makes our model more accessible to a wider range of researchers and datasets, enabling broader application in drug discovery and high-throughput screening.

To overcome these limitations, we propose a novel end-to-end, all-MLP-based transformer architecture inspired by the MLP-Mixer model (Cazenavette et al. 2021). This approach simplifies the process by utilizing 1D sequences for both drugs and targets, encoding protein targets and compounds using protein sequences and SMILES strings, respectively. Through MLPs and permutations, MLP-Mixer operates by independently combining information across both token and feature dimensions. The model significantly reduces computational costs and parameters, allowing for faster training and inference, as well as minimizing the risk of overfitting. This scalability makes the model viable even on varying and large datasets. The introduction of MLPs and permutation-based operations within the transformer architecture is a notable contribution, as it effectively captures the nonlinear interactions between drug and target sequences without the memory-intensive and quadratic scaling associated with most attention-based mechanisms. This improvement aims to create a model compatible with high-throughput screening. The MLP-Mixing framework seamlessly integrates these interactions, utilizing methods to capture nonlinear dependencies within sequential data. Furthermore, by processing both drug and protein sequences within the same chemical space, interactions are better understood and modeled with improved precision. The MLP-Mixer architecture incorporates mixer blocks, each consisting of two Multi-Layer Perceptron (MLPs) with skip connections, facilitating token and feature mixing. This setup allows for efficient processing by leveraging long-range interactions, presenting a memory-efficient alternative to traditional self-attention mechanisms. Our architecture excludes the decoder block due to our regression-focused task. Employing MLP Mixing in DTBA prediction, rather than Multi-Head Attention, addresses challenges related to computational complexity and model size. Extensive experimental evaluations using Davis and KIBA datasets demonstrate our approach's efficiency and effectiveness, making it ideal for large-scale screenings and drug repositioning. The following sections detail our proposed methodology, experiments, and results.

2 Materials and methods

2.1 Datasets

Studies on drug-target binding affinity (DTBA) often utilize datasets comprising SMILES sequences for drugs, protein sequences for targets, and a continuous value indicating the binding strength between them. This value is typically represented as the dissociation constant (Kd), inhibition constant (Ki), or half-maximal inhibitory concentration (IC50), where lower values indicate stronger binding (Öztürk et al. 2018). However, in some datasets, including those used in this study, the binding scores are transformed using Formulas (1) and (2) below, showing that higher scores indicate better binding affinities for drug-target pairs.

$$pK_d = -log_{10}\left(K_d/10^9\right) \tag{1}$$

$$pK_i = -log_{10}(K_i/10^9) \quad (2)$$

Our research used two benchmark datasets widely recognized in DTBA prediction studies: Davis dataset (Davis et al. 2011) and KIBA dataset (Tang et al. 2014).

- Davis dataset, developed by Davis et al. (2011), contains 30,056 Kd binding affinity scores for interactions between 68 compounds and 442 kinases, with the scores ranging from 5 to 10.8. This dataset focuses on kinase selectivity assays, emphasizing the Kd values for each drug-target pair.
- KIBA dataset, introduced by Tang et al. (2014), is based on the Kinase Inhibitor Bio-Activity (KIBA) scoring system which integrates data from three biochemical assays (IC50, Ki, and Kd) to evaluate kinase inhibitors activity. The original dataset includes 246,088 KIBA scores for 52,498 compounds and 467 targets. A revised version of this dataset by He et al (2017), used in our study, contains 160,296 bio-activity scores for 2,111 compounds and 229 targets, with scores ranging from 0 to 17.2, where higher scores denote stronger binding affinities.

It is worth noting that our datasets include SMILES sequences enriched with critical stereochemical details, such as chirality indicators ('@' and '@@') and double bond geometry notations ('/' and '\'). These symbols encode essential aspects of three-dimensional molecular structures, capturing the spatial orientation of chiral centers and the planar arrangement of substituents around double bonds. While this does not constitute a full 3D representation, it effectively incorporates key stereochemical features that significantly influence molecular interactions. This representation allows for a computationally efficient approach, avoiding the complexity and resource intensity of full 3D modeling, while retaining critical structural information necessary for robust predictive performance. By leveraging these detailed annotations, our model learns and utilizes stereochemical features to enhance the accuracy of predictions, achieving an optimal trade-off between computational efficiency and structural precision.

Table 1 provides a comprehensive summary of the datasets used in this study, presenting key statistics that offer insights into their scale, binding score distribution, and other relevant characteristics. Additionally, the table includes a brief description of each dataset along with references to their respective sources.

2.2 Input representation and tokenization

We used the Simplified Molecular-Input Line-Entry System (SMILES) for drug representations, a notation system effective in describing chemical structures using printable characters (Weininger 1988). SMILES is a compact and standardized format for representing molecular structures as text strings, functioning as a true language, with a small vocabulary size and few grammar rules. The tokenization process was performed by constructing a dictionary comprising 43 unique SMILES characters, with each character assigned a distinct integer value. This method facilitates an efficient encoding process that preserves the structural integrity of the chemical compounds represented. Similarly, for protein sequences, which consist of 20 different amino acids, each amino acid is encoded into an integer. This consistent encoding framework supports a streamlined and efficient approach to input representation, enhancing the overall computational efficiency of our model while maintaining the necessary precision for effective prediction.

To effectively manage the variability in sequence length within our dataset, we instituted a maximum length criterion of 1200 for proteins and 100 for drugs. This threshold was strategically chosen to encompass approximately 95% of the sequences in our dataset, as determined by their length distribution shown in Fig. 1. Sequences shorter than these limits are post-padded with zeros, while those exceeding the maximum length are truncated, ensuring consistent input sizes for model processing. While the truncation results in the loss of information for around 5% of the sequences, this trade-off was deemed acceptable as it enables the model to achieve scalable and efficient processing. Variable-length sequences, while preserving all information, would significantly increase computational demands, reduce scalability, and complicate batch processing, particularly in high-throughput workflows. Moreover, the truncated sequences likely represent outliers and do not constitute the majority of the dataset, minimizing their overall impact on the model's

Table 1 Datasets summary

Dataset	Drugs	Kinase proteins	Bindings	Source	Description
Davis	68	442	30,056	Davis et al. (2011)	Dataset of drug-target affinities for kinase inhibitors, providing pKd scores for kinase inhibitor binding affinities
Kiba	2111	229	118,254	He et al. (2017)	Comprehensive kinase inhibitor bioactivity dataset

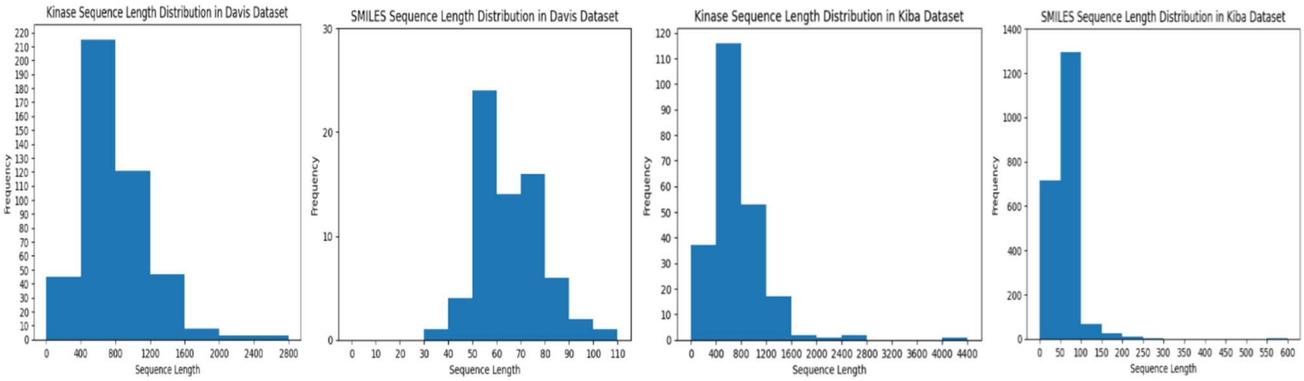

Fig. 1 Distribution of the lengths of Kinase and SMILES strings for Davis and KIBA datasets

predictive performance. Thus, the chosen approach represents a balance between computational efficiency and data integrity, prioritizing scalability and broad applicability in practical settings.

2.3 Proposed model

The TransMLP-DTBA framework is designed for the regression task of predicting binding affinity scores between drug-target pairs represented by sequences. We employed a transformer architecture. Our problem involves two modalities: drug information and target protein information. We investigated two distinct strategies.

(a) Separate Learning for Each Modality: Distinct transformer architectures are utilized for drugs and targets. The learned representations from each modality are concatenated to produce the final prediction, allowing specialized handling of the unique features inherent to each modality (Liu et al. 2023).
(b) Joint Learning of Modalities: This integrates drug and target sequences at the outset, embedding them within a unified representational space. A single transformer architecture then learns the composite representation of the drug-target pair for subsequent predictive analysis (Zhu et al. 2023a, b).

Our preliminary testing showed that the joint learning approach outperforms the separate learning strategy across all metrics, making it the focus of our study. Initially, sequences are processed through an Embedding block, converting sequence characters into 128-dimensional dense vectors while preserving positional information. These vectors are then forwarded to the Transformer-Encoder block, which applies multiple rounds of MLP operations and layer normalization on both the input and its permutation, aggregating the results. Our architecture, based on the MLP-Mixer idea, presents a novel approach to sequence processing by replacing the computationally intensive self-attention mechanism of transformers with a more efficient multi-layer perceptron (MLP) system. This leverages the inherent simplicity and scalability of MLPs, addressing the limitations of self-attention, particularly in large-scale data processing. Our novelty lies in the application of MLP mixing within the Transformer architecture. Unlike traditional MLPs, the MLP-Mixer employs separate token-mixing and channel-mixing MLPs. This technique enables the model to capture both local (within-token) and global (across-token) relationships, making it suitable for diverse tasks such as drug-target binding affinity prediction. The token-mixing MLPs capture interactions between different tokens in a sequence, while the channel-mixing MLPs handle interactions within each token's features. One of the primary advantages of the MLP-Mixer is its simplicity and efficiency. The self-attention mechanism, commonly used in standard transformer models, has a computational complexity of $O(n^2)$ relative to the sequence length. This quadratic complexity arises because the self-attention mechanism requires computing pairwise interactions between all tokens in a sequence, resulting in significant computational and memory overhead for lengthy protein sequences (Rabe and Staats 2021). In contrast, MLP mixing provides a computationally efficient alternative for processing token interactions. By using MLPs to mix information both across tokens and across features, our approach significantly reduces the computational cost while maintaining robust performance. The permutation layers are fundamental to the MLP-Mixer's success. They enable the switching of dimensions between the token-mixing and channel-mixing MLPs, facilitating the efficient and effective processing of information along both axes. This allows the model to capture nuanced relationships, enhancing its ability to learn complex patterns and dependencies. Our empirical investigations demonstrated that MLP mixing achieves an ideal equilibrium between computational efficiency and predictive performance, particularly when applied to extended protein sequences (Jiang and Xu 2023).

To further enhance performance and mitigate overfitting, our architecture incorporates residual and skip connections, a standard practice in transformer models.

In traditional transformers, the final step often involves applying global average pooling or using special tokens to generate the encoder output sequence. However, we introduced a novel method for generating the transformer output sequence by summarizing the outputs of the transformer block into a fixed-size vector. We calculate the Softmax over the first dimension (by rows), obtaining a one-dimensional vector where each element represents the Softmax of the corresponding row. More details are provided in Sect. 2.3.3.

Summary of our model's key features

Transformer Architecture Our model capitalizes on the renowned Transformer architecture, acclaimed for its proficiency in addressing complex challenges efficiently. Adaptations to the standard configuration include MLP Mixing in place of Multihead Attention, Softmax Sequence Pooling, and the omission of special tokens. These modifications enhance the model's functionality and scalability.

MLP Mixing over Multihead Attention We favored MLP mixing due to the computational challenges associated with processing extensive sequences, particularly protein data. Multihead Attention's computational complexity, denoted as O(n^2) relative to sequence length, makes it suboptimal for lengthy protein sequences (Rabe and Staats 2021). MLP mixing provides a computationally efficient alternative for processing token interactions, maintaining robust performance while addressing scalability concerns. Our empirical investigations validated that MLP mixing strikes an optimal balance between computational efficiency and predictive performance, particularly for extended protein sequences (Jiang and Xu 2023).

Exclusion of Special Tokens We deliberately excluded special tokens like [end], [start], or [class] from the input sequence representation, streamlining the model architecture and reducing computational overhead and potential error sources without diminishing predictive effectiveness. This decision facilitated more efficient use of computational resources and smoother integration of drug and target sequence data (Mehta et al. 2019).

Softmax Sequence Pooling To refine our model's accuracy in predicting drug-target binding affinity, we opted for Softmax sequence pooling over conventional global average pooling. Softmax pooling enhances precision by weighting the significance of each token's representation based on learned weights, prioritizing informative tokens and minimizing the impact of less relevant ones (Marin et al. 2021). Our experimental results consistently demonstrated that Softmax sequence pooling outperforms global average pooling in predictive performance.

Subsequent sections provide a detailed exploration of each component within our model. Figure 2 offers a visual depiction of the overall model's architecture, Algorithm 1 presents the structure of one transformer encoder block.

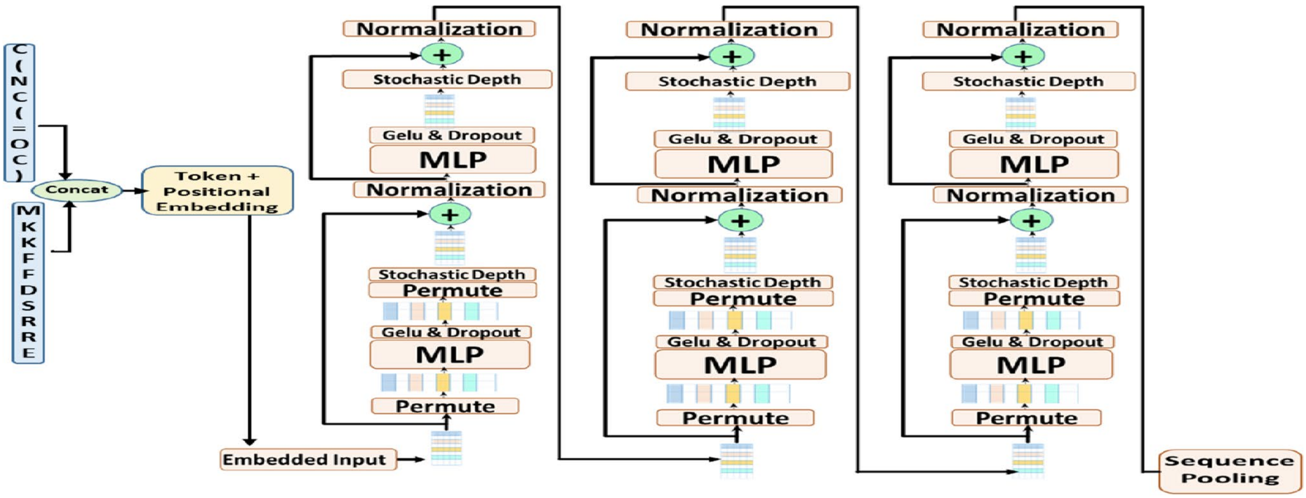

Fig. 2 Model architecture

2.3.1 Embedding block

After the preprocessing and the tokenization of the protein and SMILES sequences, the concatenated sequence is subsequently fed into an embedding block, which is responsible for transforming each encoded token position into a 128-dimensional vector. This transformation is facilitated by a learnable dictionary matrix, which dynamically adjusts during training to generate high-dimensional embeddings that capture the key features of each token.

Due to the Transformer-Encoder's inherent inability to recognize the order of tokens in a sequence, it is imperative to integrate positional information to maintain the contextual relevance of the sequence data (Zhou et al. 2019). To address this, our model includes a linear positional embedding, which incorporates spatial context to the sequence data, Eqs. 3 and 4. This embedding approach ensures that each SMILES (S) and protein (P) sequence not only retains the structural and chemical information encoded by integer tokens, but also includes positional data critical for the Transformer-Encoder. By implementing this encoding and embedding technique, our model is equipped to more accurately interpret and learn from the sequential data, thereby enhancing its ability to predict drug-target interactions with greater precision.

$$E(S) = E_{tok}(S) + E_{pos}(S) \tag{3}$$

$$E(P) = E_{tok}(P) + E_{pos}(P) \tag{4}$$

where $\mathrm{E}_{\mathrm{tok}}(\mathrm{S})$ and $\mathrm{E}_{\mathrm{pos}}(\mathrm{P})$ are the token and positional embeddings.

2.3.2 Transformer encoder

In conventional Transformer encoders, each layer consists of Multi-Head Self-Attention (MHSA) and a Multi-Layer Perceptron (MLP) head, with Layer Normalization (LN) applied before each sub-layer to stabilize learning. Residual connections promote information and gradient flow. MHSA, while effective in capturing complex token relationships, introduces significant computational challenges, scaling quadratically with sequence length ($\mathrm{O}(\mathrm{n}^2)$), making it impractical for long sequences, such as protein data (Wu et al. 2022). MHSA also struggles with long-range dependencies and parallelization, limiting efficiency on modern hardware. To address these limitations, we adapted the transformer by replacing traditional attention with MLP Mixing, simplifying computational complexity while maintaining performance (Wang et al. 2022). This method, inspired by its success in vision models, reduces computational overhead and enhances efficiency. Our model uses MLP Mixing to balance performance and computational demands, particularly for lengthy sequences in drug-target interactions. It comprises two MLP Mixing Sub-Blocks: Token-Mixing for original input and Channel-Mixing for permuted sequences, leveraging transformer strengths in a manageable framework.

MLP Mixing Block The output of the embedding block has the shape (seq_length, hidden_dimension), with the hidden dimension set to 128. The MLP Mixing block operates through two main stages: Token-Mixing MLPs and Channel-Mixing MLPs. Token-Mixing MLPs mix information across tokens (rows) independently of feature dimensions (columns), treating each feature dimension separately and enabling combinations between different tokens in the sequence. This captures dependencies between tokens within the sequence. On the other hand, Channel-Mixing MLPs mix information across feature dimensions within each token, processing each token independently and enabling combinations between the features within that token, thus capturing relationships between features within a single token. Thus, while mixing information across feature dimensions (columns), the MLPs do not consider combinations between the tokens, and vice-versa. This method alternates between these operations, ensuring comprehensive combination by first mixing information within each feature dimension independently and then within each token independently.

The workflow begins with the input sequence shaped (seq_length, hidden_dimension). To enable the MLP Mixing process, we performed permutations to rearrange the sequence. Initially, we applied the first permutation, which reoriented the sequence to focus on tokens, allowing the Token-Mixing MLPs to mix information between tokens across the sequence length. After processing through the Token-Mixing MLPs, we applied a second permutation to focus on features, allowing the Channel-Mixing MLPs to mix information between the feature dimensions within each token. By alternating between these two stages, the MLP Mixing block ensures effective communication and mixing of information both across tokens (sequence length) and within their respective features (hidden dimension). The benefits of MLP Mixing include increased efficiency, as it reduces computational complexity compared to traditional attention mechanisms while capturing both inter-token and intra-token relationships. Additionally, it offers superior scalability, handling long sequences more efficiently, making it suitable for large-scale bioinformatics datasets.

In the MLP-Mixer architecture, the Token-Mixing and Channel-Mixing MLPs are the primary components for mixing information. These two MLP types alternate to ensure comprehensive processing. The architecture is built around these mixing sub-blocks, along with layer normalization, skip connections, and the final output layer for predictions. No additional MLP head is needed, as the mixing blocks effectively handle the data processing and transformation.

Both Token-Mixing MLPs and Channel-Mixing MLPs use the Gaussian Error Linear Unit (GELU), presented in Eq. (5), as the activation function (Zhong et al. 2022; Hendrycks 2016). GELU is favored in many deep learning applications for its non-linear properties and smoothness, often providing superior performance compared to traditional activation functions such as the Rectified Linear Unit (ReLU). This preference for GELU is driven by its ability to enable more effective capture and representation of complex patterns in the data, which is crucial for the advanced computational tasks our model undertakes (Lee 2023).

$$GELU(x) = 0.5 * x * (1 + tanh\left(\sqrt{\frac{2}{x}} * \left(x + 0.044715 * x^{3}\right)\right) \tag{5}$$

GELU enhances model performance by introducing non-linearity to the input x through a mechanism that involves both the hyperbolic tangent function (tanh) and a Gaussian distribution-like transformation. Specifically, the input is scaled and shifted in a manner that approximates the cumulative distribution function of a Gaussian distribution. This transformation involves a scaling factor of $\sqrt{\frac{2}{\pi}}$ and a shifting factor of 0.044715. These adjustments allow GELU to emulate the smoothness of Gaussian functions, providing a robust method for handling the variability and complexity inherent in sequence data. The application of GELU in our MLP block significantly enhances the ability of neurons to process inputs non-linearly. This capability is essential for learning intricate patterns and relationships within the dataset, particularly in tasks involving natural language processing (NLP) and those based on transformer models. The effective utilization of GELU in these contexts has contributed to its widespread adoption across various deep learning applications, especially in areas where understanding and processing complex and nuanced data structures are critical. This makes GELU a vital component of our model's architecture, ensuring that it can achieve high levels of accuracy and performance in predicting drug-target interactions.

Skip Connection Sub-Block Stochastic Depth (Huang et al. 2016) is an innovative regularization technique that we have incorporated to enhance the robustness and efficiency of our deep neural network's training process. This method builds upon the foundational concept of skip or residual connections, which traditionally involve using all layers of a network. Stochastic Depth, however, introduces a probabilistic element to this process by randomly omitting certain layers during the training phase. The probability of a layer being skipped is initially high and gradually decreases as the training progresses. This technique of randomly skipping layers during training effectively reduces the network's depth for the duration of the training, which simplifies the model's complexity and helps to prevent overfitting. Overfitting is a common challenge in deep learning, where a model performs well on training data but poorly on unseen data. By reducing the effective number of parameters that need to be optimized during training, Stochastic Depth helps to enhance the model's generalization capabilities to new, unseen data. Moreover, like dropout, Stochastic Depth serves as a form of regularization by introducing noise into the training process. This noise forces the network to not rely on any single path or pattern, thus encouraging the learning of more robust and diverse features. This aspect is particularly beneficial in the context of very deep networks, which are prone to overfitting and often struggle with prolonged training times and convergence issues. The application of Stochastic Depth in our network underscores our commitment to deploying cutting-edge techniques to improve training outcomes. This approach has shown significant advantages in complex tasks within the realm of computer vision, such as image classification and object detection, where deep networks are especially beneficial. By implementing Stochastic Depth, we aim to ensure that our model not only achieves high accuracy but also maintains its performance robustly across different datasets and conditions.

Algorithm 1 Transformer-encoder

```
OneBlock (input: x):
    # First Permutation
    x2 = Permute (x)
    # First MLP Block with Dropout and activated by Gelu
    x3= Gelu (MLP_Drop(x2))
    # Second Permutation
    x4 = Permute (x3)
    # Residual connection 1
    x4=StochasticDepth(x4)
    x5=x4+x
    x5=LayerNormalization (x5)
    # Second MLP Block with Dropout and activated by Gelu
    x6=Gelu (MLP_Drop(x5))
    # Residual connection 2
    x6=StochasticDepth(x6)
    x7=x6+x5
    x8=LayerNormalization (x7)
```

2.3.3 Sequence pooling

In our Transformer architecture, contextual enrichment of each token's representation is a pivotal aspect achieved by capturing the intricate interactions among tokens within the sequence. To effectively synthesize these rich, contextually-enhanced representations into a singular, fixed-size vector, we employ an advanced method of sequence pooling, which aggregates the extensive sequence representations into a more manageable form, facilitating subsequent processing and analysis. While traditional BERT-inspired models (Devlin et al. 2018) often rely on a learnable special token, such as [class], [start] and [end]; or global average pooling, we have developed a refined method that leverages projection and Softmax, given in Eq. (6), to enhance the way sequence information is condensed. Our pooling technique, presented in Algorithm 2, involves projecting the sequence representations into a transformed space where the significance of each position is accentuated by weights that are calculated across all embedding dimensions. This weighted projection allows for each input token's contribution to the final representation to be adjusted based on its contextual relevance and importance. These calculated weights are applied to each token, producing a final, weighted representation that encapsulates the most pertinent features of the input sequence.

Algorithm 2 Sequence Pooling

```
SP (input: rep, where rep ∈ $R^{bs×l×hd}$) :
#bs: Batch size
#l: sequence length
#hd: Hidden dimension
    # Apply a linear layer D with one output:
    Y=  D (1) (rep)   # Y ∈ $R^{hd×1}$
    # Apply a Softmax activation to the output:
    Scores= Softmax (Y)
    # The scores are used as weights for each token to obtain
      the weighted representation:
    Weighted_rep= Scores* rep          # z ∈ $R^{bs×hd}$
```

By applying Softmax to the first dimension, we transform the representation into a probability distribution. The Softmax function, is defined as:

$$Softmax(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}} \tag{6}$$

where z_i represents the vector elements and j ranges over the first dimension, has several beneficial properties. It converts arbitrary real-valued scores into normalized probabilities, ensuring that each element's contribution is proportional to its relative magnitude. This is crucial for providing a stable basis for the final prediction layer. Softmax also, tends to amplify the differences between high and low values, making the resulting feature map more discriminative. This is particularly useful in identifying significant patterns in the input sequences. The resulting probabilities provide an intuitive measure of importance for each feature, aiding interpretability and explainability of the model's predictions.

By leveraging Softmax, we ensure that each row of the obtained representation contributes meaningfully to the final prediction, enhancing the robustness and accuracy of the binding affinity score. The resulting pooled sequence is then passed to the prediction layer, which consists of a single neuron that outputs the binding affinity score.

Thus, the SP function's capability to selectively emphasize critical features of the sequence ensures that the model's predictive accuracy is maintained, while also minimizing computational overhead and complexity.

3 Experiments and results

This section outlines the experimental setup and presents key results. The experiments were conducted using TensorFlow/Keras for the core model, along with various Python libraries for data manipulation.

3.1 Evaluation metrics

To evaluate our model and compare its performance against other state-of-the-art methods, we employed the following standard evaluation metrics:

3.1.1 Mean squared error: MSE

Mean square error (MSE), introduced in Eq. (7), is commonly used in regression tasks to measure how close the line obtained by connecting the predicted values to the actual data points. The formula below defines the MSE, where P denotes the vector of predicted values, Y denotes the vector of the ground-truth binding affinity scores, and n is the number of samples. The smaller the MSE, the better the performance of the regressor, MSE = 0 corresponds to the perfect model prediction.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(P_i - Y_i)^2 \tag{7}$$

Here P_i represents the predicted value for the ith sample, Y_i represents the actual ground-truth value for the i^{th} sample, and n is the number of samples.

3.1.2 Concordance index: CI

The concordance index (CI) (Harrell et al. 1982) extends the concept used in the area under the Receiver Operating Characteristic (ROC) curve (AUC) from binary classification to scenarios where the ranking of outcomes is essential. This metric is particularly valuable in the field of drug-target binding affinity (DTBA) prediction, where the goal often involves ranking potential interactions by likelihood rather than predicting their exact affinity values. The CI, introduced in Eq. (8), assesses a model's discrimination power by evaluating its ability to correctly order pairs of drug-target interactions based on their binding affinities. It calculates the probability that for any two randomly selected data points, the one predicted to have a higher binding affinity actually does have a higher actual affinity than another data point with a lower predicted affinity. This involves comparing the predicted rankings (f_i for higher predicted affinity and f_j for lower predicted affinity) with the actual order of the binding affinity scores (y_i for higher actual affinity and y_j for lower actual affinity).

$$CI = \frac{1}{Z}\sum_{\gamma_i > \gamma_j} h(f_\mathrm{i} - f_\mathrm{j}) \tag{8}$$

Table 2 Hyperparameters

Parameter	Options	Value
Transformer-encoders	[2, 3, 4]	3
Protein embedding dim	[128, 256]	128
SMILES embedding dim	[128, 256]	128
MLP bloc layers	[1, 2, 3]	3
MLP activation function	[ReLu, GELU]	GELU
Dropout rate	[0.1, 0.01, 0.03]	0.1
Loss function	[MSE,MAE]	MSE
Optimizer	[Adam, AdamW]	AdamW
Optimizer learning rate	[1e−3, 1e−4, 1e−5, dynamic]	Dynamic
Batch size	[128, 256, 512, 1024]	256
Epochs	[100, 200, 300]	200

A key component of the CI calculation is the normalization constant, Z, which adjusts for the total number of pairs that can be ordered differently—those pairs where the actual outcomes differ. Additionally, the calculation of CI incorporates a step function, $h(u)$, which is 1 if $u > 0$ and 0 if $u \leq 0$. This function helps in counting the number of concordant pairs, where the predicted and actual rankings align. CI values range from 0.5 to 1.0: A CI of 0.5 indicates that the model's predictions are no better than random. This level of performance suggests that the model fails to learn effectively from the data, effectively guessing without discernible insights. A CI of 1.0 corresponds to the perfect order of predictions regardless of the quality of the prediction accuracy, as the concordance index is interested exclusively in the order of the predictions, it does not give any insight on the quality of the predictions themselves. Thus, for instance, it will be equal to the maximal value 1 if the order is correct even if the predictions are inaccurate.

When the objective is the prediction of the labels relative order rather than their values, the CI measure is a suitable performance metric. For example, in DTBA investigations, it is desirable to have an estimation of the interaction likelihood of a given drug (or target) with all the targets and to rank accordingly the targets (or drugs). The CI is the proportion of the concordant pairs over the total number of pairs. Its formula is given in Eq. (8).

3.1.3 R-squared: R^2

The R-squared statistic, given in Eq. (9), quantifies how well a regression model explains the variance in the dependent variable based on the independent variables. It is computed as the squared correlation between the observed values and the values predicted by the model. A higher R-squared value indicates a stronger fit of the model to the data, implying that a greater proportion of the variability in the dependent variable is accounted for by the predictors.

In the context of Quantitative Structure–Activity Relationship (QSAR) models, R-squared serves as a critical external validation parameter (Chirico and Gramatica 2012). It assesses the model's ability to accurately predict the activity of chemical compounds based on their structural attributes. An R-squared value exceeding 0.5 for a QSAR model tested on a separate dataset generally suggests that the model is robust and reliable, demonstrating that its predictions are not random but are statistically significant.

The calculation of R-squared involves comparing the actual values (Yi), the mean of the actual values ($\overline{Y}$), and the predicted values (Pi). It takes into account the total sum of the squared differences between actual and predicted values, and the total variance in the observed data.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (P_i - Y_i)^2}{\sum_{i=1}^{n} (Y_i - \overline{Y}_i)^2} \tag{9}$$

3.2 Hyperparameter settings

The performance of deep learning models is critically dependent on the correct configuration of hyperparameters (Agaebrahimian et al. 2019). These include, but are not limited to, the learning rate, number of layers, types of activation functions, and the choice of optimizer. For our deep learning model, we adopted a dynamic approach to adjust the learning rate: starting from an initial value, the learning rate is progressively reduced during the training process whenever the model's loss shows signs of plateauing. To achieve the best possible model performance, we manually fine-tuned the hyperparameters. This fine-tuning process was guided by the model's performance metrics on the validation sets. By focusing on reducing the Mean Square Error (MSE) across these validation datasets, we identified the optimal set of hyperparameters that minimized the MSE, indicating a better predictive performance and model accuracy. The final selection of hyperparameters was thus based on their ability to achieve the lowest average MSE, ensuring that the model is not only tailored to perform well on the data it has seen but also generalizes effectively to new, independent test data. The specific details of these hyperparameters, such as the exact starting learning rate, the conditions under which it is reduced, the types of layers and activation functions used, and the choice of optimizer, are outlined in Table 2.

3.3 Results and comparison

In the current study, we proposed a novel architecture that augments the transformer framework by integrating Multilayer Perceptrons (MLP) and permutations, which facilitates the efficient learning of molecular and protein representations for the prediction of binding affinities. The performance of our model was evaluated using experiments conducted on the datasets referenced earlier, and the results were delineated in conjunction with comparisons to contemporary state-of-the-art methodologies.

Figure 3 shows a graph plotting the predicted versus actual binding affinity values for both the Davis and KIBA datasets. The dashed line represents the ideal prediction scenario (y = x) where predictions perfectly match the actual values. The solid line, derived from linear regression, shows how closely the predictions align with the actual values. Analysis of the KIBA dataset reveals that the regression fitting curve closely matches the ideal line, as it does for the Davis dataset. This indicates that our predictive model is

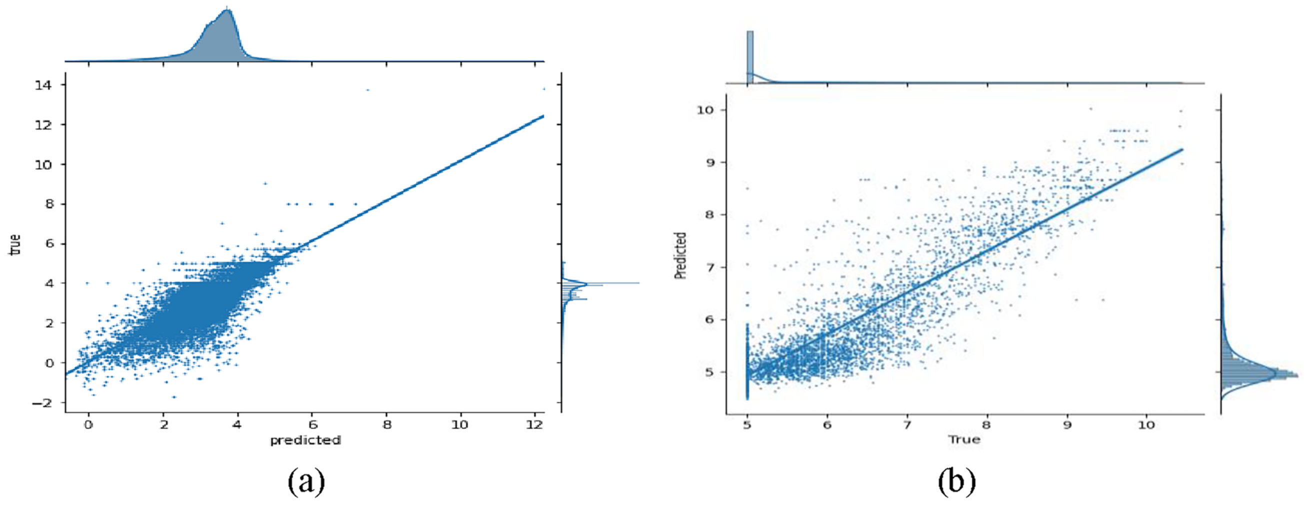

Fig. 3 **a** Actual/predicted KIBA dataset; **b** actual/predicted Davis dataset

accurate, with predictions closely mirroring the actual data for both datasets.

To validate our novel architecture and assess its prediction efficiency for drug-target binding affinities, we conducted a comparison against several state-of-the-art (SOTA) methods that are recognized leaders in this domain. Each of these methods employs unique strategies for handling drug and target representations, and our goal was to establish where our model stands relative to these advanced approaches. Here is a summary of the methods we considered in our comparative analysis:

- DeepDTA (Öztürk et al. 2018): This method utilizes SMILES strings for drugs and amino acid sequences for targets, which are processed through convolutional neural networks (CNN) to learn their respective representations. The learned representations are then concatenated and fed into a fully connected neural network to predict the binding affinities. This approach forms a foundational comparison for our model due to its direct focus on learning from sequence data.
- WideDTA (Öztürk et al. 2019): Building on the DeepDTA framework, WideDTA expands the model's capabilities by incorporating additional chemical and biological text information. This enhancement aims to improve the representation learning of drugs and targets, providing a more comprehensive data integration, which helps in capturing a wider array of interactions and dependencies in the binding process.
- DeepCDA (Abbasi et al. 2020): This method advances further by integrating CNNs with Long Short-Term Memory (LSTM) networks within its representation learning module. By combining these two powerful neural network architectures, DeepCDA is designed to capture both the spatial features of molecular structures and the sequential characteristics of protein chains, offering a more sophisticated approach to understanding drug-target interactions.
- GraphDTA (Nguyen 2021): Distinguishing itself in the landscape, GraphDTA utilizes molecular graphs along with graph neural networks (GNN) for drug representation, and combines these with protein sequences processed through 1D CNNs for target representation. This hybrid approach leverages the strengths of graph-based learning for drugs and conventional sequence processing for targets, culminating in predictions made through a multi-layer fully connected neural network.

Table 3 The performance of the TransMLP-DTBA model with state-of-the-art models

Dataset	Methods	MSE	CI	R^2
Davis	DeepDTA	0.261	0.878	0.630
	WideDTA	0.262	0.886	0.633
	DeepCDA	0.248	0.891	0.649
	GraphDTA	0.241	0.887	0.679
	TransMLP-DTBA	0.239	0.889	0.675
KIBA	DeepDTA	0.194	0.863	0.673
	WideDTA	0.179	0.875	0.675
	DeepCDA	0.176	0.889	0.682
	GraphDTA	0.151	0.883	0.687
	TransMLP-DTBA	0.150	0.887	0.690

The benchmarking of our TransMLP-DTBA model against established state-of-the-art methods provided us with a comprehensive perspective on its performance relative to current leading models in the field of drug-target interaction prediction. In this assessment, we used a variety of metrics to evaluate model performance comprehensively. Table 3 provides a detailed comparison of the TransMLP-DTBA model's performance alongside five other state-of-the-art models. The results from this comparative analysis underscore the advanced predictive capabilities of our model, as it consistently outperformed the competing models across all datasets and evaluation metrics.

As shown in Table 3, on the Davis dataset, our TransMLP-DTBA model achieved a Mean Squared Error (MSE) of 0.239, demonstrating its accuracy in prediction compared to DeepDTA, WideDTA, DeepCDA, and GraphDTA. The Concordance Index (CI) of 0.889 and R-squared (R^2) of 0.675 further indicate the model's strong performance in capturing data variability. Similarly, on the KIBA dataset, TransMLP-DTBA achieved an MSE of 0.150, highlighting its predictive accuracy. The CI of 0.887 and R^2 of 0.690 suggest that our model maintains reliable performance across different datasets. These results underscore the effectiveness of the TransMLP-DTBA model.

4 Discussion

In this study, we developed the novel TransMLP-DTBA model for drug-target binding affinity (DTBA) prediction, integrating several advanced techniques to enhance performance. This model leverages an MLP Mixing architecture within the transformer framework, circumventing the computational demands of multi-head self-attention mechanisms. By utilizing sequence data alone, including SMILES strings for drugs and amino acid sequences for proteins, the TransMLP-DTBA model demonstrates notable scalability for high-throughput drug screening, making it particularly advantageous for drug repurposing applications.

We assessed the predictive performance of TransMLP-DTBA using key metrics: Mean Squared Error (MSE), Concordance Index (CI), and R-squared (R^2). While recent models in drug-target prediction have adopted diverse computational architectures such as attention mechanisms, Graph Neural Networks (GNNs), and hybrid models to achieve higher accuracy, TransMLP-DTBA held its ground with competitive performance, particularly considering its simplicity and generalizability. These results demonstrate that the TransMLP-DTBA model not only competes with but also surpasses existing state-of-the-art solutions. Additionally, we compared our model to recent approaches using similar sequence representations. For instance, SAM-DTA, which achieved an MSE of 0.4261 and an R^2 of 0.7984 on sequence data (Hu et al. 2022). In comparison, TransMLP-DTBA outperforms SAM-DTA on MSE for both Davis (0.239) and KIBA (0.150), although SAM-DTA achieves a higher R^2 value. This suggests that while SAM-DTA captures more variance, TransMLP-DTBA provides lower prediction error, indicating a trade-off between error minimization and model complexity.

Ensemble models, such as DeepFusionDTA, combine structural and sequential features, achieving a CI increase of 1–1.5% on the KIBA and Davis datasets relative to DeepDTA (Pu et al. 2021). DataDTA similarly leverages a multi-feature aggregation approach to achieve higher CI and correlation scores (Zhu et al. 2023a, b). While both models achieve strong correlation and CI metrics, TransMLP-DTBA matches these models in prediction accuracy without the need for complex feature fusion, indicating a balance of performance with lower data demands.

4.1 Model limitations

While TransMLP-DTBA demonstrates robust predictive accuracy and marks a significant advancement in DTBA prediction, our design choices involve common trade-offs worthy of discussion. Our model relies exclusively on sequence data, using SMILES strings for drugs and amino acid sequences for proteins, to achieve remarkable scalability and computational efficiency. This approach enables high-throughput prediction while capturing key sequence-derived features. For example, our drug dataset incorporates chirality indicators (such as "@" and "/" symbols) that provide valuable insights into stereochemistry and partially reveal aspects of a molecule's three-dimensional structure. In contrast, protein data is limited to primary sequences and does not convey secondary, tertiary, or quaternary structural details. Although this choice supports efficiency and feasibility for processing large-scale data, it may challenge the capture of spatial interactions in complex cases, such as with multi-domain proteins or compounds with multifaceted binding sites. In addition, the model's performance is highly sensitive to the quality of input data; errors in SMILES representations or sequence alignments can propagate inaccuracies. Finally, our reliance on the KIBA and Davis datasets, widely recognized benchmarks that predominantly focus on kinases, introduces a bias that may restrict the generalizability of our findings to other protein families. These considerations underscore established trade-offs in the field while also revealing both opportunities for future refinement and the inherent strengths of our approach.

5 Future directions

To enhance predictive accuracy, integrating multi-scale techniques that operate at different resolution levels of the sequence data may help capture both high-level patterns and fine-grained details, thereby improving overall performance. For instance, representing sequences using n-grams of varying lengths is one promising approach. In addition, incorporating biological pathway information can provide valuable context for drug-target interactions by illuminating the broader biological processes involved, potentially refining model predictions further. To improve interpretability, explainable AI techniques can offer insights into the model's internal decision-making processes. These enhancements will collectively position TransMLP-DTBA as a versatile tool for drug discovery, supporting both high-throughput screening and detailed mechanistic investigations across diverse biomedical applications. Furthermore, in forthcoming research, we are planning to address dataset bias by expanding training data to include a broader representation of protein families; this effort will involve identifying and integrating publicly available datasets that cover non-kinase proteins, followed by a rigorous evaluation of model performance and generalizability on the expanded dataset.

6 Conclusion

The study introduces an innovative approach to drug-target binding affinity (DTBA) prediction through the transformer-based TransMLP-DTBA model. By integrating an MLP Mixing architecture, this model circumvents the computational demands of multi-head self-attention mechanisms, enabling robust and efficient prediction with reduced resource requirements. Utilizing sequence data alone, including SMILES strings for drugs and amino acid sequences for proteins, the TransMLP-DTBA model demonstrates notable scalability for high-throughput drug screening, making it particularly advantageous for drug repurposing applications. Rigorous benchmarking against state-of-the-art methods highlights the model's superior performance. The TransMLP-DTBA model achieved remarkable results across the Davis and KIBA datasets, with notable performance across all metrics. The model's strong generalization capabilities are essential for practical applications in drug discovery and repurposing, offering a promising avenue for future research and development in pharmacological studies. Future work will focus on refining the model's architecture and exploring additional datasets. This thoughtful approach seeks to explore the model's potential in advancing drug discovery and development processes, aiming for a scalable and cost-effective solution for early-stage drug discovery.

Data availability The data supporting the findings of this study are available from the corresponding author upon request.

Declarations

Conflict of interest All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Abbasi K, Razzaghi P, Poso A, Amanlou M, Ghasemi JB, Masoudi-Nejad A (2020) DeepCDA: deep cross-domain compound–protein affinity prediction through LSTM and convolutional neural networks. Bioinformatics 36(17):4633–4642. https://doi.org/10.1093/bioinformatics/btaa544

Abdel-Basset M, Hawash H, Elhoseny M, Chakrabortty R, Ryan M (2020) DeepH-DTA: deep learning for predicting drug-target interactions: a case study of COVID-19 drug repurposing. IEEE Access 8:170433–170451. https://doi.org/10.1109/ACCESS.2020.3024238

Aghaebrahimian A, Cieliebak M (2019) Hyperparameter tuning for deep learning in natural language processing. In: 4th swiss text analytics conference (swisstext 2019), winterthur, June 18–19 2019. SwissText. https://doi.org/10.21256/zhaw-18993

Ballester PJ, Mitchell JB (2010) A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. Bioinformatics 26(9):1169–1175. https://doi.org/10.1093/bioinformatics/btq112

Blay V, Tolani B, Ho S, Arkin M (2020) High-Throughput Screening: today's biochemical and cell-based approaches. Drug Discov Today. https://doi.org/10.1016/j.drudis.2020.07.024

Carnero A (2006) High throughput screening in drug discovery. Clin Transl Oncol 8:482–490. https://doi.org/10.1007/S12094-006-0048-2

Cazenavette G, De Guevara ML (2021) MixerGAN: an MLP-based architecture for unpaired image-to-image translation. arXiv preprint arXiv:2105.14110. https://doi.org/10.48550/arXiv.2105.14110

Chawla NV, Karakoulas G (2005) Learning from labeled and unlabeled data: an empirical study across techniques and domains. J Artif Intell Res 23:331–366. https://doi.org/10.1613/jair.1509

Chen XW, Lin X (2014) Big data deep learning: challenges and perspectives. IEEE 2:514–525

Chen L, Tan X, Wang D, Zhong F, Liu X, Yang T, Luo X, Chen K, Jiang H, Zheng M (2020) TransformerCPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. Bioinformatics 36(16):4406–4414. https://doi.org/10.1093/bioinformatics/btaa524

Chirico N, Gramatica P (2012) Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. J Chem Inf Model 52(8):2044–2058. https://doi.org/10.1021/ci300084j

Dash S, Shakyawar SK, Sharma M, Kaushik S (2019) Big data in healthcare: management, analysis and future prospects. J Big Data 6(1):1–25. https://doi.org/10.1186/s40537-019-0217-0

Davis MI, Hunt JP, Herrgard S, Ciceri P, Wodicka LM, Pallares G, Hocker M, Treiber DK, Zarrinkar PP (2011) Comprehensive analysis of kinase inhibitor selectivity. Nat Biotechnol 29(11):1046–1051. https://doi.org/10.1038/nbt.1990

Deng L, Zeng Y, Liu H, Liu Z, Liu X (2022) DeepMHADTA: prediction of drug-target binding affinity using multi-head self-attention and convolutional neural network. Curr Issues Mol Biol 44:2287–2299. https://doi.org/10.3390/cimb44050155

Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

Du Y, Guo X, Shehu A, Zhao L (2020) Interpretable molecule generation via disentanglement learning. In: Proceedings of the 11th ACM international conference on bioinformatics, computational biology and health informatics, pp 1–8. https://doi.org/10.1145/3388440.3414709

Eberhardt J, Santos-Martins D, Tillack AF, Forli S (2021) AutoDock Vina 1.2.0: new docking methods, expanded force field, and python bindings. J Chem Inf Model 61(8):3891–3898. https://doi.org/10.1021/acs.jcim.1c00203

Ezzat A, Zhao P, Wu M, Li XL, Kwoh CK (2016) Drug-target interaction prediction with graph regularized matrix factorization. IEEE/ACM Trans Comput Biol Bioinform 14(3):646–656. https://doi.org/10.1109/TCBB.2016.2530062

Feng Q, Dueva E, Cherkasov A, Ester M (2018) Padme: a deep learning-based framework for drug-target interaction prediction. arXiv preprint arXiv:1807.09741. https://doi.org/10.48550/arXiv.1807.09741

Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT et al (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 47(7):1739–1749. https://doi.org/10.1021/jm0306430

Grčar M, Mladenič D, Fortuna B, Grobelnik M (2006) Data sparsity issues in the collaborative filtering framework. In: Advances in web mining and web usage analysis: 7th international workshop on knowledge discovery on the web, WebKDD 2005, Chicago, IL, USA, August 21, 2005. Revised papers 7. Springer, Berlin, pp 58–76

Green, L., Bell, C., & Janjić, N. (2001). Aptamers as reagents for high-throughput screening.. BioTechniques, 30 5, 1094–6, 1098, 1100 passim . https://doi.org/10.2144/01305dd02.

Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA (1982) Evaluating the yield of medical tests. JAMA 247(18):2543–2546. https://doi.org/10.1001/jama.1982.03320430047030

He T, Heidemeyer M, Ban F, Cherkasov A, Ester M (2017) SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. J Cheminform 9:1–14. https://doi.org/10.1186/s13321-017-0209-z

Hendrycks D, Gimpel K (2016) Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415

Hu Z, Liu W, Zhang C, Huang J, Zhang S, Yu H, Xiong Y, Liu H, Ke S, Hong L (2022) SAM-DTA: a sequence-agnostic model for drug-target binding affinity prediction. Brief Bioinform. https://doi.org/10.1093/bib/bbac533

Hua Y, Song X, Feng Z, Wu XJ, Kittler J, Yu DJ (2022) CPInformer for efficient and robust compound-protein interaction prediction. IEEE/ACM Trans Comput Biol Bioinf 20(1):285–296

Huang K, Fu T, Glass LM, Zitnik M, Xiao C, Sun J (2020) DeepPurpose: a deep learning library for drug–target interaction prediction. Bioinformatics 36(22–23):5545–5547. https://doi.org/10.1093/bioinformatics/btaa1005

Huang K, Xiao C, Glass LM, Sun J (2021) MolTrans: molecular interaction transformer for drug–target interaction prediction. Bioinformatics 37(6):830–836. https://doi.org/10.1093/bioinformatics/btaa880

Huang G, Sun Y, Liu Z, Sedra D, Weinberger KQ (2016) Deep networks with stochastic depth. In: Computer Vision—ECCV 2016: 14th European conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. Springer, pp 646–661

Inglese J, Auld D, Jadhav A, Johnson R, Simeonov A, Yasgar A, Zheng W, Austin C (2006) Quantitative high-throughput screening: a titration-based approach that efficiently identifies biological activities in large chemical libraries. Proc Natl Acad Sci USA 103(31):11473–11478. https://doi.org/10.1073/PNAS.0604348103

Jiang Y, Xu Y (2023) Revenge of MLP in sequential recommendation. arXiv preprint arXiv:2305.14675. https://doi.org/10.48550/arXiv.2305.14675

Kim S, Zhao Z (2013) Unified inference for sparse and dense longitudinal models. Biometrika 100(1):203–212. https://doi.org/10.1093/BIOMET/ASS050

Kwon Y, Shin WH, Ko J, Lee J (2020) AK-score: accurate protein-ligand binding affinity prediction using an ensemble of 3D-convolutional neural networks. Int J Mol Sci 21(22):8424. https://doi.org/10.3390/ijms21228424

L'heureux A, Grolinger K, Elyamany HF, Capretz MA (2017) Machine learning with big data: challenges and approaches. IEEE Access 5:7776–7797

Lee I, Keum J, Nam H (2019) DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. PLoS Comput Biol 15(6):e1007129. https://doi.org/10.1371/journal.pcbi.1007129

Lee M (2023) GELU activation function in deep learning: a comprehensive mathematical analysis and performance. arXiv preprint arXiv:2305.12073. https://doi.org/10.48550/arXiv.2305.12073

Liu J, Lu Y, Guan S, Jiang T, Ding Y, Fu Q, Cui Z, Wu H (2023) Drug-target interaction prediction via combining transformer and graph neural networks. https://doi.org/10.2174/1574893618666230912141426

Ma D, Li S, Chen Z (2023) Drug-target binding affinity prediction method based on a deep graph neural network. Math Biosci Eng MBE 20(1):269–282. https://doi.org/10.3934/mbe.2023012

Mahmud SH, Chen W, Jahan H, Dai B, Din SU, Dzisoo AM (2020) DeepACTION: A deep learning-based method for predicting novel drug-target interactions. Anal Biochem 610:113978. https://doi.org/10.1016/j.ab.2020.113978

Marin D, Chang JH, Ranjan A, Prabhu A, Rastegari M, Tuzel O (2021) Token pooling in vision transformers. arXiv preprint arXiv:2110.03860

Mayr L, Bojanic D (2009) Novel trends in high-throughput screening. Curr Opin Pharmacol 9(5):580–588. https://doi.org/10.1016/j.coph.2009.08.004

McInnes C (2007) Virtual screening strategies in drug discovery. Curr Opin Chem Biol 11(5):494–502. https://doi.org/10.1016/J.CBPA.2007.08.033

Mehta S, Koncel-Kedziorski R, Rastegari M, Hajishirzi H (2019) Define: deep factorized input token embeddings for neural sequence modeling. arXiv preprint arXiv:1911.12385

Meli R, Anighoro A, Bodkin MJ, Morris GM, Biggin PC (2021) Learning protein-ligand binding affinity with atomic environment vectors. J Cheminform 13(1):59. https://doi.org/10.1186/s13321-021-00536-w

Meng M, Wei Z, Li Z, Jiang M, Bian Y (2019) Property prediction of molecules in graph convolutional neural network expansion. In: 2019 IEEE 10th international conference on software engineering and service science (ICSESS), pp 263–266. IEEE. https://doi.org/10.1109/ICSESS47205.2019.9040723

Merrouche W, Harrou F, Taghezouit B, Sun Y (2024) Improved lithium-ion battery health prediction with data-based approach. Adv Electr Eng Electron Energy. https://doi.org/10.1016/j.prime.2024.100457

Mukherjee S, Ghosh M, Basuchowdhuri P (2022) DeepGLSTM: deep graph convolutional network and LSTM based approach for predicting drug-target binding affinity. In: Proceedings of the 2022 SIAM international conference on data mining (SDM). Society for Industrial and Applied Mathematics, pp 729–737. https://doi.org/10.1137/1.9781611977172.82

Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. Front Neurorobot 7:21. https://doi.org/10.3389/fnbot.2013.00021

Nguyen T, Le H, Quinn TP, Nguyen T, Le TD, Venkatesh S (2021) GraphDTA: predicting drug–target binding affinity with graph neural networks. Bioinformatics 37(8):1140–1147. https://doi.org/10.1093/bioinformatics/btaa921

Öztürk H, Özgür A, Ozkirimli E (2018) DeepDTA: deep drug–target binding affinity prediction. Bioinformatics 34(17):821–829. https://doi.org/10.1093/bioinformatics/bty593

Öztürk H, Ozkirimli E, Özgür A (2019) WideDTA: prediction of drug-target binding affinity. arXiv preprint arXiv:1902.04166

Pahikkala T, Airola A, Pietilä S, Shakyawar S, Szwajda A, Tang J, Aittokallio T (2015) Toward more realistic drug–target interaction predictions. Brief Bioinform 16(2):325–337. https://doi.org/10.1093/bib/bbu010

Peng L, Liao B, Zhu W, Li Z, Li K (2015) Predicting drug–target interactions with multi-information fusion. IEEE J Biomed Health Inform 21(2):561–572. https://doi.org/10.1109/JBHI.2015.2513200

Pu Y, Li J, Tang J, Guo F (2021) DeepFusionDTA: drug-target binding affinity prediction with information fusion and hybrid deep-learning ensemble model. IEEE/ACM Trans Comput Biol Bioinf 19:2760–2769. https://doi.org/10.1109/TCBB.2021.3103966

Rabe MN, Staats C (2021) Self-attention does not need O (n^2) memory. arXiv preprint arXiv:2112.05682

Rifaioglu AS, Nalbat E, Atalay V, Martin MJ, Cetin-Atalay R, Doğan T (2020) DEEPScreen: high performance drug-target interaction prediction with convolutional neural networks using 2-D structural compound representations. Chem Sci 11(9):2531–2557. https://doi.org/10.1039/C9SC03414E

Shar PA, Tao W, Gao S, Huang C, Li B, Zhang W, Shahen M, Zheng C, Bai Y, Wang Y (2016) Pred-binding: large-scale protein–ligand binding affinity prediction. J Enzyme Inhib Med Chem 31(6):1443–1450. https://doi.org/10.3109/14756366.2016.1144594

Sugita S, Ohue M (2021) Drug-target affinity prediction using applicability domain based on data density. In: 2021 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB), pp 1–6. https://doi.org/10.26434/chemrxiv.14498688.v1

Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23(10):1282–1288. https://doi.org/10.1093/bioinformatics/btm098

Tang J, Szwajda A, Shakyawar S, Xu T, Hintsanen P, Wennerberg K, Aittokallio T (2014) Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. J Chem Inf Model 54(3):735–743

Üstün B, Melssen WJ, Buydens LM (2006) Facilitating the application of support vector regression by using a universal Pearson VII function-based kernel. Chemom Intell Lab Syst 81(1):29–40. https://doi.org/10.1016/j.chemolab.2005.09.003

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30

Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y (2017) Graph attention networks. arXiv preprint arXiv:1710.10903. https://doi.org/10.17863/CAM.48429

Vert J (2011) 3D ligand-based virtual screening with support vector machines, pp 35–45. https://doi.org/10.4018/978-1-61520-911-8.CH003

Vilar S, Cozza G, Moro S (2008) Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery. Curr Top Med Chem 8(18):1555–1572. https://doi.org/10.2174/156802608786786624

Wang Z, Jiang W, Zhu YM, Yuan L, et al (2022) Dynamixer: a vision MLP architecture with dynamic mixing. In: Proceedings of the 39th international conference on machine learning, PMLR 162, pp 22691–22701. https://proceedings.mlr.press/v162/wang22i.html

Wang C, Zhang Y (2017) Improving scoring-docking-screening powers of protein–ligand scoring functions using random forest. J Comput Chem 38(3):169–177

Wei B, Zhang Y, Gong X (2022) DeepLPI: a novel deep learning-based model for protein–ligand interaction prediction for drug repurposing. Sci Rep 12:23014. https://doi.org/10.1038/s41598-022-23014-1

Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 28(1):31–36. https://doi.org/10.1021/ci00057a005

Wieder O, Kohlbacher S, Kuenemann M, Garon A, Ducrot P, Seidel T, Langer TA (2020) compact review of molecular property prediction with graph neural networks. Drug Discov Today Technol 37:1–2. https://doi.org/10.1016/j.ddtec.2020.11.009

Wu YH, Liu Y, Zhan X, Cheng MM (2022) P2T: pyramid pooling transformer for scene understanding. IEEE Trans Pattern Anal Mach Intell

Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M (2008) Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. Bioinformatics 24(13):i232–i240. https://doi.org/10.1093/bioinformatics/btn162

Yan S, Zheng X, Chen D, Wang Y (2013) Exploiting two-faceted web of trust for enhanced-quality recommendations. Expert Syst Appl 40(17):7080–7095. https://doi.org/10.1016/j.eswa.2013.06.035

Yang S, Zhu F, Ling X, Liu Q, Zhao P (2021) Intelligent health care: applications of deep learning in computational medicine. Front Genet 12:607471. https://doi.org/10.3389/fgene.2021.607471

Yuan W, Chen G, Chen C (2021) FusionDTA: attention-based feature polymerizer and knowledge distillation for drug-target binding affinity prediction. Brief Bioinform. https://doi.org/10.1093/bib/bbab506

Zeng Y, Chen X, Luo Y, Li X, Peng D (2021) Deep drug-target binding affinity prediction with multiple attention blocks. Brief Bioinform 22(5):117. https://doi.org/10.1093/bib/bbab117

Zhao Q, Zhao H, Zheng K, Wang J (2022) HyperAttentionDTI: improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism. Bioinformatics 38(3):655–662. https://doi.org/10.1093/bioinformatics/btab715

Zhong YD, Zhang T, Chakraborty A, Dey B (2022) A neural ode interpretation of transformer layers. arXiv preprint arXiv:2212.06011. https://doi.org/10.48550/arXiv.2212.06011

Zhou P, Fan R, Chen W, Jia J (2019) Improving generalization of transformer for speech recognition with parallel schedule sampling and relative positional embedding. arXiv preprint arXiv:1911.00203

Zhou H, Zhang S, Peng J, Zhang S, Li J, Xiong H, Zhang W (2021) Informer: beyond efficient transformer for long sequence time-series forecasting. In: Proceedings of the AAAI conference on artificial intelligence, vol 35, no 12, pp 11106–11115. https://api.semanticscholar.org/CorpusID:229156802 Accessed 26 June 2023

Zhu Z, Yao Z, Qi G, Mazur N, Yang P, Cong B (2023a) Associative learning mechanism for drug-target interaction prediction. CAAI Trans Intell Technol 8(4):1558–1577

Zhu Y, Zhao L, Wen N, Wang J, Wang C (2023b) DataDTA: a multi-feature and dual-interaction aggregation framework for drug–target binding affinity prediction. Bioinformatics. https://doi.org/10.1093/bioinformatics/btad560

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.