Pre-trained Inspired MocFormer: Efficient and Predictive Models of Drug-target Interactions

Yi-Lun Zhang[1*], Wen-Tao Wang[2*], Jia-Hui Guan[1], Hao-Wen Yang[1]
[1] The Chinese University of Hong Kong, Shenzhen
[2] Chongqing University of Posts and Telecommunications

***Abstract*—Numerous deep learning (DL) methods have been proposed to identify drug-target interactions (DTIs). However, these methods often face challenges due to the diversity and complexity of drugs and proteins and the presence of noise and bias in the data. Limited labeled data and extracting meaningful features from datasets also pose difficulties. These limitations hinder the development of accurate and general deep-learning models for DTI prediction. To address these challenges, a novel framework is introduced for identifying DTIs. The framework incorporates pre-trained molecular representation models and a transformer module inspired by pre-training. By pre-training the model, it can acquire a more comprehensive feature representation, enabling it to handle the diversity and complexity of drugs and proteins effectively. Moreover, the model mitigates noise and bias in the data by learning general feature representations during pre-training, improving prediction accuracy. In addition to pre-training, a transformer mechanism called MocFormer is proposed. MocFormer extracts feature matrices from drug and protein sequences obtains decision vectors, and makes predictions based on these decision vectors. Experiments were conducted using public datasets from DrugBank to evaluate the framework's effectiveness. The results demonstrate that the proposed framework outperforms state-of-the-art methods regarding accuracy, area under the ROC curve (AUC), recall, and the area under the precision-recall curve (AUPRC). The code for the framework can be accessed from the following GitHub repository: GitHub Repository.**

***Index Terms*—drug–target interactions, pre-training, transformer**

I. Introduction

Drug discovery and drug repurposing are highly valued in the current field of biomedicine. Identifying drug–target interactions (DTIs) is critical in drug discovery and repurposing. But generally, the process is always costly and in high risk [1], the mean cost of developing a new drug needs a mean investment of $1335.9 million, which is mainly from repeated laboratory experimental procedures. Then some computer-aided methods were developed, such as virtual screening (VS) [2]. However, the VS method is based on structure with limited speed. Then, the deep learning method was introduced into the field of drug discovery.

Recently, deep learning has achieved superior performance compared with classical methods in many fields, such as computer vision and natural language processing [3]–[8]. With the production of a large amount of biological activity data in recent years, predicting DTIs through deep learning technology has become research. Initially, researchers usually only used manual annotation to label proteins and small molecules with manual descriptors in limited datasets. Tian et al. proposed using a fully connected neural network (FCNN) to represent drugs and proteins based on hand annotation. Drugs and proteins based on hand-crafted descriptors for prediction. Later on, with the further development of deep learning itself, transformer [9] and GNN [10] were proposed, and attempts were made to encode and decode molecules and proteins separately through transformer [11]. Encoding and decoding [11], [12] to learn their high-dimensional structures and input them into neural networks for iteration to simulate their interactions. Meanwhile, graph neural networks are also the usual means to study DTI, where one constructs its 2D structure by treating atoms as nodes and chemical bonds as edges. The attention mechanism has been widely used in both approaches, which is thought to capture the key sites where its small molecules bind to proteins [13].

The main methods of studying DTIs by deep learning are three categories: Sequence-based, structure-based, and net-word-based.

1) Sequence-based methods try to analyze features from the sequence data of drugs in SMILES [14] and protein amino acids sequence. The function and structure information is believed to be included in sequence simplicity.
2) Structure-based methods utilize 3D structure data of proteins and ligand molecules to study the interaction details to predict the binding affinity [15]. Net-work-based methods aim to contain drugs, targets, and other biological entities into a graph-based network and try to extract the biochemical functional information [16].

Despite advancements in the field, predicting drug-target interactions (DTIs) continues to encounter challenges. These include effectively handling the diversity and complexity of drugs and proteins, addressing noise and bias in the data, utilizing limited labeled data efficiently, and extracting meaningful features from the datasets. These obstacles impede the development of accurate and generalized deep-learning models for DTI prediction. Overcoming these challenges is crucial for advancing the field and enhancing the performance of DTI prediction models.

To address the abovementioned challenges, inspired by the success of transfer learning and pre-training in computer vision and natural language processing tasks [17]–[20], this paper introduces a novel approach called pre-trained inspired MocFormer for predicting drug-target interactions (DTIs).

Identify applicable funding agency here. If none, delete this.

The proposed model takes the SMILES string of drugs and the amino acid sequence of proteins as input. Initially, both inputs undergo processing by the Molecule pre-trained module (Uni-Mol) [21] and the protein pre-trained module (ESM-2) [22], respectively. Each amino acid and SMILES character is transformed into its corresponding embedding vector.

Through pre-training, the model can acquire a more comprehensive feature representation, enabling it to handle the diversity and complexity of drugs and proteins effectively. Moreover, the model can mitigate noise and bias in the data by learning general feature representations during the pre-training stage, thereby improving prediction accuracy. Additionally, a transformer mechanism called MocFormer is proposed. MocFormer is utilized to extract feature matrices from drug and protein sequences, obtain decision vectors, and subsequently make predictions based on these decision vectors.

In summary, this paper presents the following contributions: 1) To the best of our knowledge, a pre-trained inspired transformer is proposed for the first time, to achieve transfer learning based drug and protein interactions prediction, termed MocFormer. 2) A counterintuitive phenomenon is discovered, referred to as the 'one-sided trap'. It is observed that solely employing molecular pre-training or protein pre-training models results in inferior performance on the DTI task. Possible explanations are provided for this phenomenon. 3) The MocFormer pipeline outperforms the state-of-the-art (SOTA) methods in the DTI task, demonstrating superior performance.

II. Methods

Figure1 provides an overview of our framework for identifying drug–target interactions (DTIs) from the SMILES string of drugs and amino acid sequence of proteins. The framework consists of three main modules: a molecule pre-trained module, a protein pre-trained module, and a biologically inspired attention module. Each module is described in detail below. Given the drug's SMILES strings and protein's amino acid sequences, CNN block extracts feature matrices from the sequences of drugs and proteins. And finally, the prediction results will be the output.

A. Molecule Pre-trained Module

Uni-Mol is a 3D molecule representation learning framework with three main components. Firstly, it utilizes a transformer-based backbone, which takes atoms and atom pairs as inputs and incorporates the SE(3) method to reduce the 3D conformation of a molecule. Secondly, the model is trained on a large dataset comprising 209 million molecules and 3 million proteins. Lastly, the trained model is fine-tuned using downstream tasks such as predicting the drug-target inter-right and inter-wrong sites and their corresponding 3D structures.

In the MocFormer pipeline, the grid search method was employed to fine-tune the pre-trained model provided by Uni-Mol on the Davis dataset. The pre-trained model from the Davis dataset underwent fine-tuning using the random forest regression method, and the learning rate was selected from the range [1e-5, 1e-4, 4e-4, 1e-3]. Furthermore, different batch sizes, namely [8, 16, 32], were experimented with. To ensure robustness, the five-fold cross-validation technique was utilized. This technique allowed for the selection of three sets of optimal characterization results. These optimal sets of representation vectors were then used as input for MocFormer's model inference, and the final choice was determined based on the best performance.

After the Molecule Pre-trained Module processes the input, the drug's embedding matrix, denoted as f_D, is obtained. The computation can be summarized using Equation (1), where f represents the size of the embeddings for drug strings, and 512 means the embedding dimensions.

$$f_D \in R^{\,512\times f} \tag{1}$$

B. Protein Pre-trained Module

ESM-2 is developed based on the belief that the information regarding structure and function can be found in amino acid sequences, making LLMs (Large Language Models) a handy tool for this task. ESM-2 remains a transformer-based model with a maximum of 15 billion parameters. It utilizes approximately 138 million sequences for training and employs an equivalent transformer to represent the protein's three-dimensional structure. This results in an attention pattern corresponding to the protein's three-dimensional structure.

In the MocFormer model, the chosen variant of ESM-2 is a large language model with 36 layers and 3 billion parameters. It is fine-tuned using the DTI (Drug-Target Interaction) task on the Davis dataset. The selected method for fine-tuning is the K-neighborhood algorithm, which is optimized using the grid search approach. The hyperparameters being searched include the batch size (options: 8, 16, 32), the number of neighbors (options: 5, 10), the weighting strategy (options: uniform, distance), and the algorithm type (options: ball_tree, kd_tree, brute). The leaf size is also considered for the algorithm (options: BallTree, KDTree). Finally, three sets of vector representations are selected and fed into the subsequent model to determine the best set of representation vectors.

After the Protein Pre-trained Module processes the input, the protein's embedding matrix, denoted as f_P, is obtained. The computation can be summarized using Equation (2), where f represents the size of the embeddings for protein strings, and 2560 means the embedding dimensions.

$$f_P \in R^{\,2560\times f} \tag{2}$$

C. Transformer Module

In this pipeline, two transformers are employed to encode and decode the drug and target within the transformer module. The transformer module utilizes a multi-attention mechanism to capture the most significant vector dimensions for the drug-target prediction task. This mechanism assigns higher weights to these dimensions from the 512-dimensional drug vectors and the 2,560-dimensional protein vectors. Moreover, the multi-head attention mechanism within the transformer further enhances this process, ensuring that the more critical vector dimensions are emphasized for the drug-target prediction task.

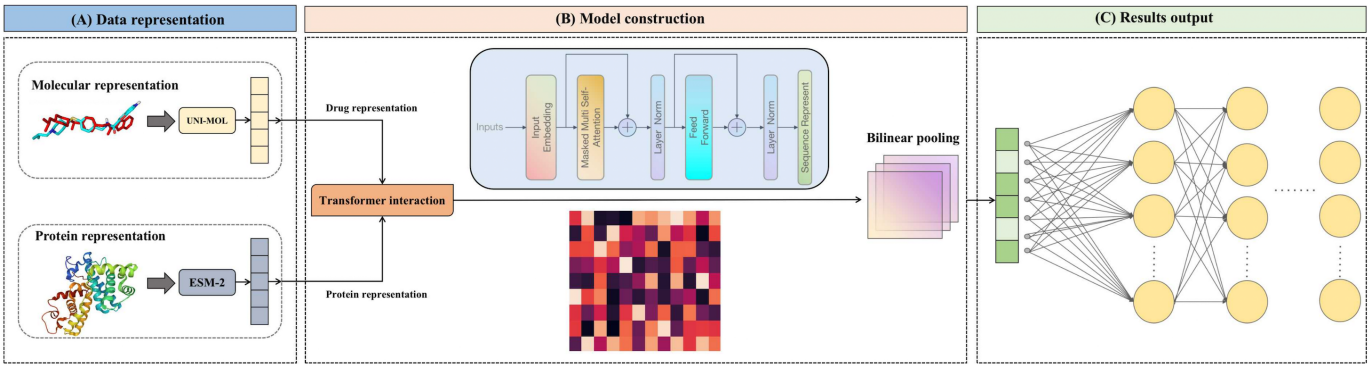

Fig. 1. An overview of Pre-trained Inspired MocFormer.

The allocation of weights facilitates MocFormer in learning the intrinsic patterns associated with drug-target interactions.

The computation can be summarized using Equation 3–6. The query (Q), key (K), and value (V) are defined as follows: Q represents the query, K represents the key, and V represents the value of the protein and drug. The weight matrices are denoted as W^Q, W^K, and W^V, while d_k means the dimensions of the vectors.

$$Q = f_D \times W^Q \tag{3}$$

$$K = f_D \times W^K \tag{4}$$

$$V = f_D \times W^V \tag{5}$$

$$Attention = softmax(\frac{Q \times K^T}{\sqrt{d_k}}) \times V \tag{6}$$

The multi-head attention is then introduced and summarized using Equation 7–8. For each head, there are weight matrices W_i^Q, W_i^K, and W_i^V, with dimensions $W_i^Q \in \mathbb{R}^{d_{512} \times d_{64}}$, $W_i^K \in \mathbb{R}^{d_{512} \times d_{64}}$, and $W_i^V \in \mathbb{R}^{d_{512} \times d_{64}}$. Additionally, a linear transformation matrix $W_i^O \in \mathbb{R}^{d_{512} \times d_{512}}$ is utilized.

$$Head_i = Attention(Q \times W_i^Q, K \times W_i^K, V \times W_i^V) \tag{7}$$

$$MultiHead = Concat(Head_1, ...Head_8) \times W_Q \tag{8}$$

The fully connected feed-forward network comprises two dense layers, each followed by a ReLU activation function, allowing for nonlinear transformations. This can be summarized using Equation 9. The weight matrices W_1 and W_2 have dimensions of $\mathbb{R}^{f \times f}$, and bias terms b_1 and b_2 are also included.

$$FFN_D = max(0, x \times W_1 + b_1)W_2 + b_2 \in R^{\ 256 \times f} \tag{9}$$

The handling of protein embedding vectors is also something to consider.

$$Q = f_P \times W^Q \tag{10}$$

$$K = f_P \times W^K \tag{11}$$

$$V = f_P \times W^V \tag{12}$$

W_i^Q,W_i^K,W_i^V are all weight matrices, $W_i^Q \in R^{\ d_{2560} \times d_{512}}$, $W_i^K \in R^{\ d_{2560} \times d_{512}}$, $W_i^V \in R^{\ d_{2560} \times d_{512}}$, $W_i^O \in R^{\ d_{2560} \times d_{2560}}$is the linear transformation matrix.

$$Head_i = Attention(Q \times W_i^Q, K \times W_i^K, V \times W_i^V) \tag{13}$$

$$MultiHead = Concat(Head_1, ...Head_8) \times W_Q \tag{14}$$

Also, the network comprises two dense layers, each followed by a ReLU activation function, allowing for nonlinear transformations. This can be summarized using Equation 9. The weight matrices W_1 and W_2 have dimensions of $\mathbb{R}^{f \times f}$, and bias terms b_1 and b_2 are also included.

$$FFN_P = max(0, x \times W_1 + b_1)W_2 + b_2 \in 512 \times f \tag{15}$$

D. Bilinear Pooling and Full Connected Layer

The bilinear pooling technique fuses features from the drug and protein decoders. It involves bilinearly multiplying the first two features at the same position to obtain the matrix **B**. Then, sum pooling is applied to all positions in **B** to get the matrix ξ. The matrix ξ is further transformed into a vector, referred to as the bilinear vector **x**. Additionally, moment normalization and L2 normalization operations are performed on **x** to obtain the fused features **Z**. The bilinear pooling method is utilized to merge the output of the drug and protein decoders.

Then, the merged vector representation will be fed into a multi-layer, fully connected layer network. The activation function is relu, a dropout layer is added after each layer to prevent overfitting, and a binary cross entropy is used to output the final prediction results.

$$B(x, f_P, f_D) = f_P \times f_D^T \tag{16}$$

$$\xi = \sum_x^A f_P \times f_D^T \tag{17}$$

$$m = vec(\xi) \tag{18}$$

$$y = sign(x)\sqrt{|x|} \tag{19}$$

$$y = \frac{y}{||y||_2} \in R^{\,2560\times512\times1} \tag{20}$$

III. Experimental Results

This section presents the results obtained by applying the proposed methods to the DrugBank dataset. The experimental dataset and evaluation metrics will be explained in Section A. The implementation details of the experiments will be discussed in Section B. In addition, Section C will present the results of the ablation study, while Section D will provide a comprehensive comparison with the current state of the art.

A. Dataset and Evaluation Metrics

The experimental dataset for this study was derived by extracting drug and target data from the DrugBank database [23], as presented in Table I. The dataset used in this research corresponds to the data released on January 3, 2020 (version 5.1.5). Inorganic compounds and tiny molecule compounds (e.g., Iron [DB01592] and Zinc [DB01593]) were manually discarded, along with drugs having SMILES strings that could not be recognized by the RDKit Python package [24]. After this filtering process, 6,655 drugs, 4,294 proteins, and 17,511 positive drug-target interactions (DTIs) remained in the dataset.

To create a balanced dataset with equal positive and negative samples, unlabeled drug-protein pairs were sampled following a common practice [11], [25]. This approach allowed for the generation of negative samples, resulting in a balanced dataset for analysis.

Four key metrics were considered for a comprehensive performance analysis: Accuracy, AUC, Recall, and Area Under the Precision-Recall Curve (AUPRC). Accuracy assesses overall correctness, AUC evaluates the model's ability to rank positive and negative samples correctly, Recall measures the model's effectiveness in identifying positive samples, and AUPRC evaluates the model's performance in classifying imbalanced datasets.

TABLE I
Summary of the benchmark datasets

Datasets	Protein	Drug	Interaction	Positive	Negative
DrugBank	4294	6655	35022	17511	17511

B. Implementation Details

The framework used in this study is built on the PyTorch platform and utilizes an NVIDIA Tesla V100S GPU. The entire dataset was divided into training, validation, and testing sets, with proportions of 70%, 20%, and 10%, respectively. Each experiment employed a 5-fold cross-validation approach. The AdamW optimizer optimized the model with an initial learning rate of 0.001 and a weight decay 0.001. Additionally, a learning rate schedule based on ReduceROnPlateau was implemented. This schedule had a patience of 5, meaning that if the model's validation loss did not decrease after five epochs, the learning rate would decay to 10% of the previous rate.

C. Ablation Study

To assess the effectiveness of each component in our method, a series of ablation experiments were conducted, as presented in Table II. These experiments progressively enhanced the baseline network by applying the following configurations: 1) Adding only the molecule pre-trained module (A) to the baseline. 2) Adding only the protein pre-trained module (B) to the baseline. 3) Simultaneously adding the molecule and protein pre-trained modules to the baseline. 4) A transformer with bilinear pooling was incorporated After combining the molecule and protein pre-trained modules with the baseline (C).

The baseline will be to perform the encoding and decoding process for small molecules and protein sequences using the two word2vec functions in gensim, respectively, using average pooling to connect the two types of vectors and pass them to the multilayer perceptron MLP (consisting of multiple fully-connected layers). Baseline+A, on the other hand, Baseline+A will replace the word2vec representation in baseline with the pre-trained model (fine-tuned) of Uni_mol, using the word2vec process for proteins. Baseline+A, the word2vec in the baseline, is replaced by a pre-trained model of Uni_mol (fine-tuned) to characterize small molecules vectorially. At the same time, proteins are still processed using word2vec and input to the MLP using average pooling. Baseline+B, the vectorial characterization of proteins, is replaced by a pre-trained model of ESM2 (fine-tuned). At the same time, small molecules are still processed using Baseline+A+B, and the embeddings are generated using the pre-training strategies of Uni_mol and ESM-2, respectively, and input into the MLP after average pooling. Baseline+A+B+C is our final pipeline. Uni_mol and ESM-2 generate the embeddings and input into the MLP after the bilinear pooling layer. The embeddings are generated by Uni_mol and ESM-2, respectively, fused by the bilinear pooling layer, input to the transformer, processed by multi-head self-attention mechanism, and then entered into a fully connected layer to get the prediction result.

During the experiments, a counterintuitive discovery was made: the final performance of both Baseline+A and Baseline+B was weaker than that of Baseline. This finding was unexpected, considering that Uni_mol and ESM-2, known as powerful molecular characterization models, were expected to enhance the model's representation.

One possible explanation for this phenomenon is that using word2vec-generated vector representations as input might drive the model to rely on topology for predictions. These word2vec-generated vector representations lack true biochemical meaning when used as input. Additionally, the interactions with embeddings generated by other pre-trained molecular characterization models can potentially lead the model to

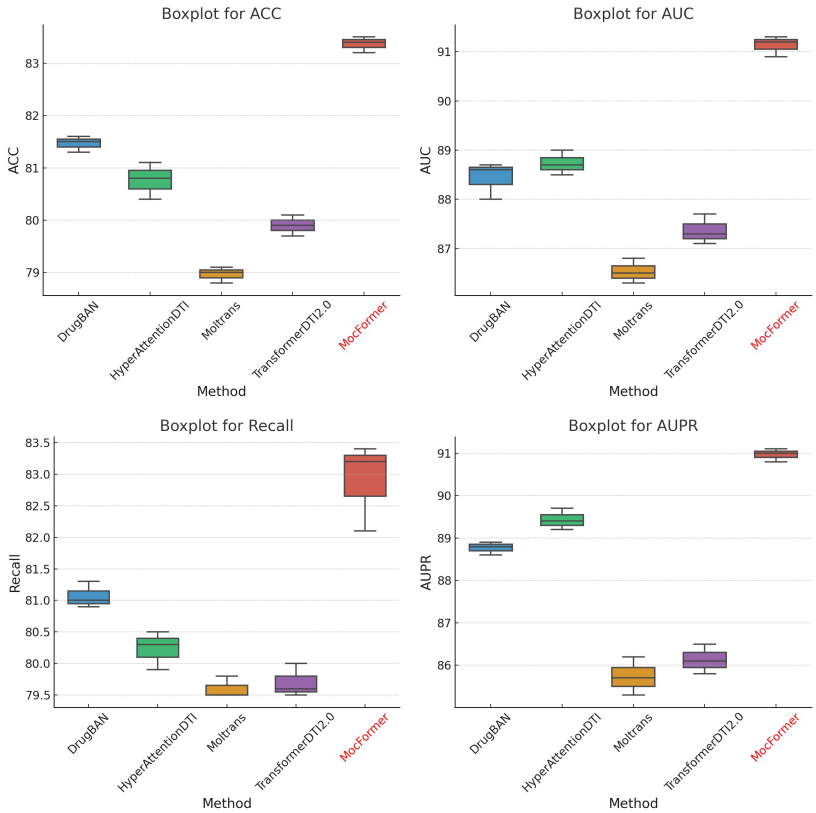

Fig. 2. Box plot of quantitative comparisons.

learn an incorrect paradigm, ultimately resulting in weakened results.

TABLE II
RESULTS OF ABLATION STUDIES

Settings	Acc (%)	AUC (%)	Recall (%)	AUPR (%)
Baseline	76.0	82.2	75.6	84.2
Baseline+A	72.8	77.9	71.7	79.2
Baseline+B	70.1	73.7	68.9	77.9
Baseline+A+B	77.9	86.1	77.6	86.1
Baseline+A+B+C	**83.4**	**91.2**	**83.2**	**91.0**

D. Comparison with Other Methods

The proposed framework was evaluated using the Davis dataset in the experiments. The results demonstrated superior performance compared to state-of-the-art methods regarding accuracy, AUC, recall, and AUPR, as shown in Figure2 and TableIII. The experimental results indicate that our method, which incorporates a robust pre-training-inspired transformer architecture, outperforms existing methods that train from scratch, thus achieving a new state-of-the-art (SOTA) result.

IV. CONCLUSION

This paper introduces the Pre-trained Inspired MocFormer, a novel framework for identifying Drug-Target Interactions (DTIs). The proposed architecture effectively addresses the challenges posed by the diversity and complexity of drugs and proteins and the presence of noise and bias in the data. The framework extracts meaningful features from limited labeled data and datasets by leveraging pre-trained molecular representation models and a pre-training-inspired transformer module. The primary objective of pre-training the model is to obtain a comprehensive feature representation capable of adequately handling the diversity and complexity of drugs and proteins. Furthermore, the pre-training stage helps mitigate noise and bias in the data by facilitating the learning of general feature representations, thereby improving prediction accuracy. The transformer mechanism is employed to extract feature matrices from drug and protein sequences, enabling the derivation of decision vectors. These decision vectors form the basis for making predictions. Experiments were conducted on the Davis dataset to evaluate our method's effectiveness. The results demonstrate that our framework significantly advances the development of accurate and general deep-learning models for DTI prediction. In future research, we plan to explore the

TABLE III
RESULTS OF QUANTITATIVE COMPARISONS

Settings	Acc (%)	AUC (%)	Recall (%)	AUPR (%)
DrugBAN [26]	81.5	88.6	81.0	88.8
HyperAttentionDTI [27]	80.8	88.7	80.3	89.4
MolTrans [11]	79.0	86.5	79.5	85.7
TransformerCPI2.0 [28]	79.9	87.3	79.6	86.1
Ours	**83.4**	**91.2**	**83.2**	**91.0**

End-to-End Learning Paradigm to enhance the performance of the identification approach further.

ACKNOWLEDGMENT

REFERENCES

[1] O. J. Wouters, M. McKee, and J. Luyten, "Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018," *JAMA*, vol. 323, no. 9, pp. 844–853, 2020.

[2] M. Himmat, N. Salim, M. M. Al-Dabbagh, F. Saeed, and A. Ahmed, "Adapting document similarity measures for ligand-based virtual screening," *Molecules*, vol. 21, no. 4, p. 476, 2016.

[3] M. Liu, W. Zou, W. Wang, C.-B. Jin, J. Chen, and C. Piao, "Multi-conditional constraint generative adversarial network-based mr imaging from ct scan data," *Sensors*, vol. 22, no. 11, 2022.

[4] R.-Q. Li, X.-L. Xie, X.-H. Zhou, S.-Q. Liu, Z.-L. Ni, Y.-J. Zhou, G.-B. Bian, and Z.-G. Hou, "A unified framework for multi-guidewire endpoint localization in fluoroscopy images," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 4, pp. 1406–1416, 2022.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proceedings of Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, 2021, pp. 8748–8763.

[8] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, B. Yin, and X. Hu, "Harnessing the power of llms in practice: A survey on chatgpt and beyond," 2023.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of Advances in Neural Information Processing Systems*, vol. 30, 2017.

[10] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020.

[11] K. Huang, C. Xiao, L. M. Glass, and J. Sun, "MolTrans: Molecular Interaction Transformer for drug–target interaction prediction," *Bioinformatics*, vol. 37, no. 6, pp. 830–836, 2020.

[12] L. Chen, Z. Fan, J. Chang, R. Yang, H. Hou, H. Guo, Y. Zhang, T. Yang, C. Zhou, Q. Sui *et al.*, "Sequence-based drug design as a concept in computational drug design," *Nature Communications*, vol. 14, no. 1, p. 4217, 2023.

[13] M. Yazdani-Jahromi, N. Yousefi, A. Tayebi, E. Kolanthai, C. J. Neal, S. Seal, and O. O. Garibay, "AttentionSiteDTI: an interpretable graph-based model for drug-target interaction prediction using NLP sentence-level relation classification," *Briefings in Bioinformatics*, vol. 23, no. 4, p. bbac272, 2022.

[14] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences*, vol. 28, no. 1, pp. 31–36, 1988.

[15] S. Wang, D. Liu, M. Ding, Z. Du, Y. Zhong, T. Song, J. Zhu, and R. Zhao, "Se-onionnet: a convolution neural network for protein–ligand binding affinity prediction," *Frontiers in Genetics*, vol. 11, p. 607824, 2021.

[16] Y. Qu, C. He, J. Yin, Z. Zhao, J. Chen, and L. Duan, "Move: Integrating multi-source information for predicting dti via cross-view contrastive learning," in *Proceedings of 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2022, pp. 535–540.

[17] J. Li, X. Li, T. Wang, S. Wang, Y. Cao, C. Xu, and D. Dou, "Improving bert fine-tuning via stabilizing cross-layer mutual information," in *Proceedings of ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[19] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2021.

[20] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[21] G. Zhou *et al.*, "Uni-mol: A universal 3d molecular representation learning framework," in *Proceedings of The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=6K2RM6wVqKu

[22] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives, "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123–1130, 2023.

[23] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey, "Drugbank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic Acids Research*, vol. 34, no. $\text{suppl}_1, pp. D668 - -D672, 2006$.

[24] G. Landrum *et al.*, "Rdkit: Open-source cheminformatics software," 2016.

[25] M. Wen, Z. Zhang, S. Niu, H. Sha, R. Yang, Y. Yun, and H. Lu, "Deep-learning-based drug–target interaction prediction," *Journal of Proteome Research*, vol. 16, no. 4, pp. 1401–1409, 2017.

[26] P. Bai, F. Miljković, B. John, and H. Lu, "Interpretable bilinear attention network with domain adaptation improves drug-target prediction," *Nature Machine Intelligence*, 2023.

[27] Q. Zhao, H. Zhao, K. Zheng, and J. Wang, "HyperAttentionDTI: improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism," *Bioinformatics*, vol. 38, no. 3, pp. 655–662, 2021.

[28] L. Chen, Z. Fan, J. Chang, R. Yang, H. Hou, H. Guo, Y. Zhang, T. Yang, C. Zhou, Q. Sui *et al.*, "Sequence-based drug design as a concept in computational drug design," *Nature Communications*, vol. 14, no. 1, p. 4217, 2023.