MocFormer: A Two-Stage Pre-training-Driven Transformer for Drug-Target Interactions Prediction

Yi-Lun Zhang[1*], Wen-Tao Wang[2*], Jia-Hui Guan[1], Deepak Kumar Jain[3], Tian-Yang Wang[4], Swalpa Kumar Roy[5]

[1*]School of Medicine, The Chinese University of Hong Kong (Shenzhen), Longcheng Street, Shenzhen, 518172, Guangdong, China.
[2]School of Automation, Chongqing University of Posts and Telecommunications, Nanshan Street, Chongqing, 400065, Chongqing, China.
[3]Key Laboratory of lntelligent Control and Optimization for Industrial Equipment of Ministry of Education, Dalian University of Technology, Linggong Road, Dalian, 116024, Liaoning, China.
[4]Department of Computer Science, University of Alabama at Birmingham, 10th Ave S, Birmingham, 35294, AL, USA.
[5]Department of Computer Science and Engineering, Alipurduar Government Engineering and Management College, Alipurduar, 736206, West Bengal, India.

*Corresponding author(s). E-mail(s): yilunzhang@link.cuhk.edu.cn; 2020212707@stu.cqupt.edu.cn;
Contributing authors: jiahuiguan@link.cuhk.edu.cn; dkj@ieee.org; tw2@uab.edu; swalpa@ieee.org;

Abstract

Drug-target interactions (DTIs) is essential for advancing pharmaceuticals. Traditional drug-target interaction studies rely on labor-intensive laboratory techniques. Still, recent advancements in computing power have elevated the importance of deep learning methods, offering faster, more precise, and cost-effective screening and prediction. Nonetheless, general deep learning methods often yield low-confidence results due to the complex nature of drugs and proteins, bias, limited labeled data, and feature extraction challenges. To address

these challenges, a novel two-stage pre-trained framework is proposed for DTIs prediction. In the first stage, pre-trained molecule and protein models develop a comprehensive feature representation, enhancing the framework's ability to handle drug and protein diversity. This also reduces bias, improving prediction accuracy. In the second stage, a transformer with bilinear pooling and a fully connected layer (FCN) enables predictions based on feature vectors. Comprehensive experiments were conducted using public datasets from DrugBank and Dissimilar-compound-split to evaluate the framework's effectiveness. The results demonstrate that the proposed framework outperforms the state-of-the-art methods regarding accuracy, area under the ROC curve (AUC), recall, and the area under the precision-recall curve (AUPRC).

Keywords: Drug-target interactions, Pre-training, Transformer, Bilinear pooling, FCN

1 Introduction

Predicting the drug-target interactions (DTIs) could be applied in multiple fields, for example, the drug discovery [1], drug repositioning [2], and the prediction of drug side effect [3]. The critical drug discovery process is identifying DTIs among numerous candidates[4]. Although conventional measurement in vitro experimental testing can verify DTIs, it suffers from extremely long time and monetary costs. To reduce the wet-lab-based verification procedure's expensive workload, computational approaches are adopted to efficiently filter potential DTIs from a large number of candidates for subsequent biological experiments[5]. The traditional in-silico computational methods could generally be classified into three categories: ligand-based, target-based, and chemogenomic approches [6]. The ligand-based approaches predict the DTIs by the similarities of the ligands. They search for similar compounds verified to interact with the particular target. Target-based methods claim that a target with a similar 3D structure can interact with the same drug. However, both methods rely on powerful computing resources, running time, and accurate 3D protein structures. With the development of machine learning and deep learning in recent years, chemogenomic methods have begun to be widely used. These use target and ligand characters simultaneously to make the interaction predictions. These computational approaches rely on machine learning techniques to build a prediction model to accurately estimate undiscovered interactions based on the chemogenomic space that incorporates drug and target information.

Recently, various deep models have shown encouraging performance in DTI predictions. Initially, researchers usually only used manual annotation to label proteins and small molecules with manual descriptors in limited datasets. Then, researchers proposed a CNN-based model [7], which utilizes multi-scale one-dimensional convolutional neural network [8] to obtain targets features and Extended Connectivity Fingerprints [9] to get compounds features. Also, the attention mechanism is introduced into DTIs prediction [10, 11]. At the same time, with the further development of deep learning [12–16], transformer [17] and GNN [18] were proposed, and attempts

were made to encode and decode molecules and proteins separately through transformer [19]. Encoding and decoding [19] 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 DTIs, 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 [20]. HyperAttentionDTI's attention mechanism can infer the interactions of each amino acid atom pair but also control the characteristics on the channel [21]. DrugBAN proposed a bilinear attention network with domain adaptation to explicitly learn pairwise local interactions between drugs and targets and has a specific generalization ability. Both methods can represent local interactions to some extent through improved attention mechanisms. In recent years, with the development of molecule and protein language pre-training models, people have tried to encode the smiles and protein sequences of molecules into vectors to represent their physical and chemical functions and structural information [22–26]. These pre-trained models are trained on an extensive unlabeled molecule and protein data set so these vectors can better represent their physical and chemical characteristics. DeepLPI [27] and AI-Bind [28] utilize these pre-trained models to make the DTI predictions.

Despite these efforts, the following challenges are still open. 1) The complex nature of drugs and proteins presents a formidable challenge. These molecules exhibit various structural variations, chemical interactions, and biological functions, making them difficult to predict accurately. 2) Inherent bias in the data can introduce significant uncertainties. Biomedical datasets are often collected from specific populations or experimental conditions, which may not fully represent the diversity of biological systems. This bias can result in models that perform well in particular scenarios but struggle when applied to more diverse or real-world situations. 3) A limited availability of labeled data poses a substantial challenge. Supervised learning methods rely on labeled examples for training and require substantial quantities of accurately annotated data. In drug discovery and protein analysis, obtaining large, high-quality labeled datasets is expensive and time-consuming. Consequently, models may not be sufficiently trained to handle the full spectrum of potential inputs, leading to lower confidence in their predictions. 4) Feature extraction remains a persistent challenge. Identifying and selecting relevant features from complex biological data is a non-trivial task. Inaccurate feature representation or the omission of crucial information can significantly impact the performance of predictive models, contributing to the uncertainty in their results.

To overcome the above issues, a two-stage framework is proposed for accurate and rubust drug-target interactions (DTIs) prediction, as shown in Fig. 1. In the first stage, pre-trained molecule and protein foundation models are applied to encode the drug and protein sequences into comprehensive feature vectors. The advantages of the pre-training are twofold. Firstly, pre-training foundation models for molecule and protein structures provide a powerful starting point for feature representation. This pre-trained model addresses the limitation of having a limited number of labeled protein-drug pairs when using deep learning methods to predict drug-target interactions (DTIs).

Compared to previous deep learning methods which builds embeddings from a limited number of proteins and molecules using relatively simple methods from scratch, the approach of using transfer learning with pre-trained models for molecule and protein representation allows training on a much larger dataset of unlabeled molecules and proteins, thereby avoiding overfitting and obtaining more accurate features. In the second stage, a transformer with bilinear pooling and a fully connected layer further processes the feature vector acquired from the first stage and outputs the final prediction result. It enhances the grasp of drug-target relationships and interaction prediction accuracy, spanning various scales like molecule structures.

In summary, this paper presents the following contributions: 1) To the best of our knowledge, a pre-training driven transformer framework is proposed for the first time, termed MocFormer, to achieve drug and target interactions prediction based on transfer learning. 2) The first stage obtains a comprehensive vector representation of molecule and protein features through fine-tuning and transfer learning. The second stage enhances the grasp of drug-target relationships and the accuracy of interaction prediction through transformer, bilinear pololing and FCN. 3) Experimental results show that our method outperforms the most recent state-of-the-art drug-target interactions methods on two public benchmarks, demonstrating the effectiveness of our method and its potential applicability in clinical practice.

The structure of this paper is outlined as follows: In Section 2, we provide an overview of the framework by presenting its workflow. Section 3 presents the experimental setup and comprehensive experimental results. In Section 4, we summarize the entire paper and give priorities for future work.

2 Methods

Figure 1 offers a comprehensive illustration of the framework for identifying drug–target interactions (DTIs) through the utilization of drug SMILES strings and protein amino acid sequences. The framework is divided into two primary stages. In the first stage, we employ pre-trained foundational models to process molecule and protein data. In the subsequent second stage, we utilize a pre-training-driven transformer, bilinear pooling, and a fully connected layer (FCN) to further refine the DTIs prediction. This two-stage process is pivotal in achieving accurate and reliable predictions of drug–target interactions.

2.1 Molecule Pre-trained Module

Uni-Mol is an advanced framework designed for 3D molecule representation learning with three key components. 1) The foundation of Uni-Mol relies on a transformer-based backbone. This backbone effectively processes input data consisting of individual atoms and atom pairs, integrating the SE(3) methodology to condense the intricate 3D structure of molecules effectively; 2) To ensure robustness and comprehensive learning, Uni-Mol undergoes training on a vast dataset, encompassing an impressive 209 million molecules and 3 million proteins. This extensive training dataset equips the model with a broad understanding of molecule structures and their relationships; 3) Uni-Mol's capabilities are further enhanced through fine-tuning various downstream tasks.

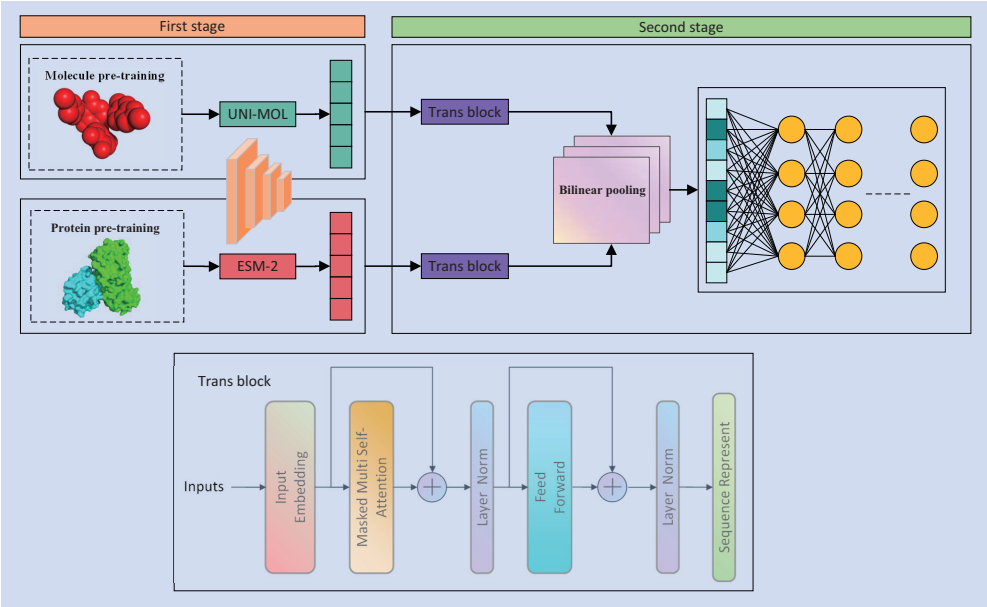

Fig. 1 The overarching workflow of the proposed framework encompasses three pivotal constituents: a data representation driven by pre-training, a transformer influenced by pre-trained models, and the dissemination of results.

These tasks include predicting drug-target interaction sites, distinguishing between correct and incorrect binding sites, and predicting the corresponding 3D structures. Fine-tuning refines the model's abilities to make precise predictions and contributes to its overall versatility in molecule analysis.

In the MocFormer pipeline, the grid search method was employed to fine-tune the pre-trained model provided by Uni-Mol on the DrugBank dataset. The pre-trained model from the DrugBank 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. Following the preprocess of the Molecule Pre-trained Module, we obtain the drug's embedding matrix, which is denoted as f_D. This computation can be concisely expressed using Equation (1), where f denotes the size of the embeddings for drug strings, and we've set this dimension to 512. This equation encapsulates the transformation of drug data into a meaningful embedded representation.

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

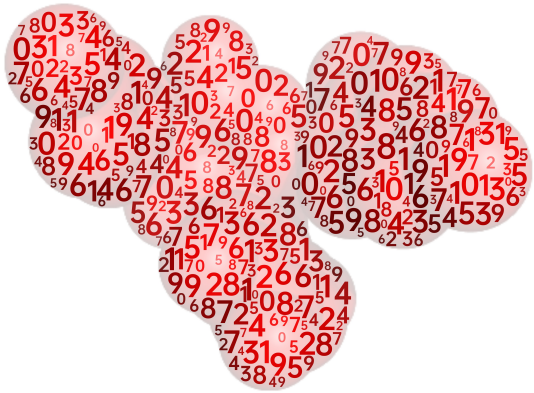

Fig. 2 Word cloud visualization of molecule pre-training results.

2.2 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. A fine-tuning process is meticulously designed to enhance its capabilities further and adapt it for the Drug-Target Interactions (DTIs) task. 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 distinct sets of vector representations are derived. These sets are then utilized as input for the subsequent model in the pipeline. The goal is to identify the set of representation vectors that consistently delivers the best performance, ensuring that the final model is optimized for the DTIs task. 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.

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

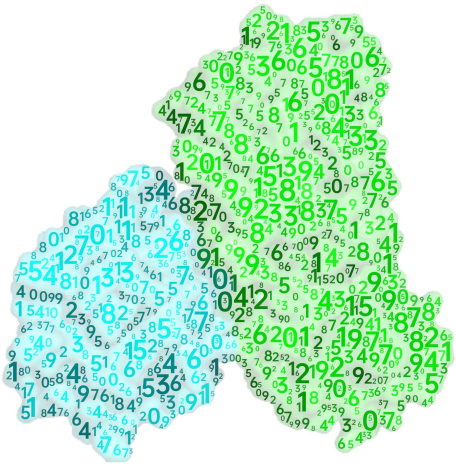

Fig. 3 Word cloud visualization of protein pre-training results.

2.3 Transformer Module

In this pipeline, transformer modules utilize a multi-attention mechanism to calculate the feature vector of molecule and protein acquired from the first stage. This mechanism assigns weights to dimensions from the 512-dimensional drug vectors and the 2,560-dimensional protein vectors. Moreover, the multi-head attention mechanism within the transformer further improves this process, ensuring that the more critical vector dimensions are focused on. The allocation of weights facilitates MocFormer in learning the intrinsic patterns associated with drug-target interactions. By focusing on the most pertinent vector dimensions, MocFormer learns and captures the intrinsic relationships and nuances inherent in drug-target interactions.

The computation can be summarized using Equation 3–6. Where Q represents the query, K represents the key, and V represents the value. 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 $W_O \in \mathbb{R}^{d_{512} \times d_{512}}$ is utilized.

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

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

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 similar process is used for protein embedding vectors.

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

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

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

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.

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

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

Also, the network comprises two dense layers, each followed by a ReLU activation function, allowing for nonlinear transformations. This can be summarized using Equation 15. 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.

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

2.4 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. The specific calculation process can be expressed by Equation 16–20.

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

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

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

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

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

3 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 3.1. The implementation details of the experiments will be discussed in Section 3.2. In addition, Section 3.3 will present the results of the ablation study, while Section 3.4 will provide a comprehensive comparison with the current state of the art.

3.1 Dataset and Evaluation Metrics

DrugBank dataset: The experimental dataset for this study was derived by extracting drug and target data from the DrugBank database [29], as presented in Table 1. 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 [30]. 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 [19, 31]. This approach allowed for the generation of negative samples, resulting in a balanced dataset for analysis.

Table 1 Summary of the DrugBank dataset

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

Dissimilar-compound-split dataset: This dataset is based on protein family-specific datasets (Large-scale) [32], further constructed by applying a strategy that only considers compound similarities while distributing bioactivity data points into train-test splits, as presented in Table 2. Compounds in train and test splits are

dissimilar (Tanimoto score < 0.5). Therefore, similar compounds cannot participate in both train and test splits. This strategy makes the prediction task more challenging and realistic than random splitting. It partly prevents the model from memorizing bioactivities over identical or highly similar compound fingerprints shared between train and test folds. In previous experiments of a similar nature, researchers segregated distinct molecules into the training and test sets. In other words, they ensured that the molecules in the test set were not present in the training set and vice versa. However, in our current study, we have taken a different approach by including entirely dissimilar types of molecules. Specifically, molecule A is part of the test set, while molecule B is found in the training set. These molecules exhibit a substantial dissimilarity, indicated by a Tanimoto score of less than 0.5. This implies that their degree of similarity is exceedingly low, akin to the difference between humans and dogs, as opposed to mere similarity, such as that between men and women. The score quantifies the degree of similarity. The term "Pchem value" denotes the experimental measurement of the interaction between the target and ligand. In this context, we selected a threshold of 6. If the Pchem value exceeds 6, it indicates the presence of an interaction, and the corresponding label is set to 1.

Table 2 Summary of the Dissimilar-compound-split dataset

Datasets	Protein	Drug	Interaction	Positive	Negative
Dissimilar-compound-split	113	10120	17757	8834	8923

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.

3.2 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.

3.3 Ablation Study

To assess the effectiveness of each component in our method, a series of ablation experiments were conducted, as presented in Table 3 and Figure 4. These experiments

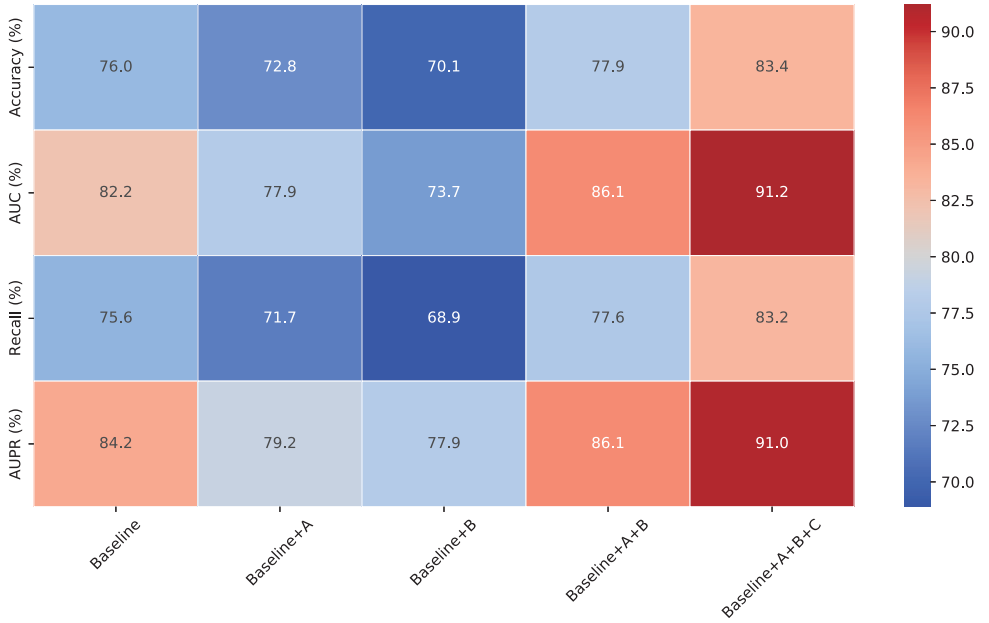

Fig. 4 The heatmap shows that adding conditions "A" and "B" individually initially reduces performance, but combining them yields a significant positive impact ("Baseline+A+B"). Moreover, introducing condition "C" ultimately allows the model to achieve the best performance.

Table 3 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**

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 (C) was incorporated after combining the molecule and protein pre-trained modules with the baseline. **Baseline**: 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**: Baseline+A will replace the original word2vec representation for molecule in baseline with the pre-trained model (fine-tuned) of Uni_mol to characterize small molecules vectorially. At the same time, proteins are still processed

using word2vec. Other settings are the same as the baseline. **Baseline+B**: Baseline+B will replace the original word2vec representation for protein in baseline with the pre-trained model (fine-tuned) of ESM-2 to generate embeddings. At the same time, molecules are still processed using word2vec. Other settings are the same as the baseline. **Baseline+A+B**: Although Uni_mol and ESM-2 are known as powerful molecule characterization models, using word2vec-generated vector representations as input might cause the model to rely on topology for predictions, thereby lacking practical biochemical meaning. This issue influences interactions between the single-side (Molecule or Protein) word2vec module and the pre-trained molecule characterization module (Protein or Molecule), causing them to learn an incorrect paradigm and ultimately resulting in weakened results. Therefore, better performance can be achieved by simultaneously pre-training encoding for both molecules and proteins. **Baseline+A+B+C**: Our final framework is Baseline+A+B+C. In this framework, Uni_mol and ESM-2 generate the embeddings, which are then input into the MLP after passing through the transformer and bilinear pooling layers, ultimately yielding the prediction results.

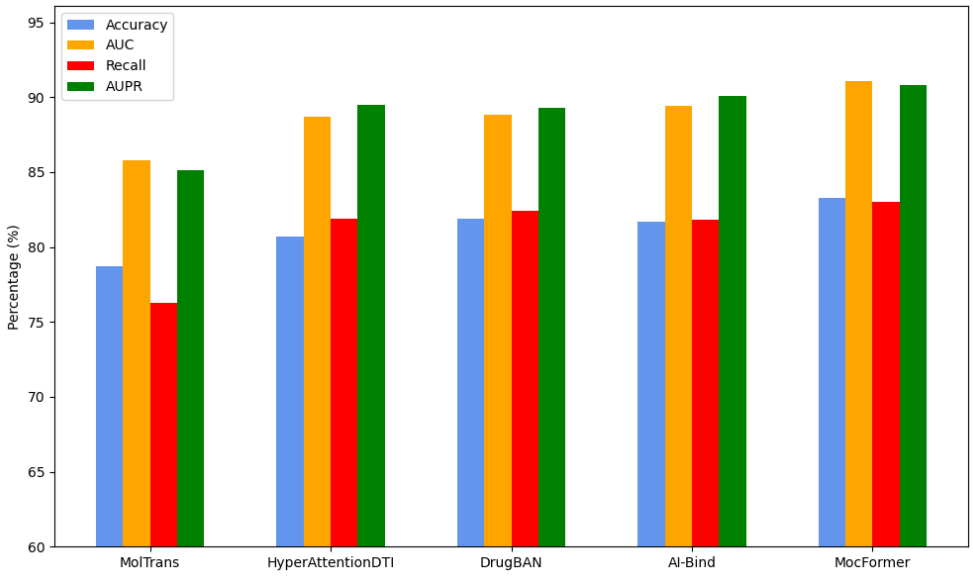

Fig. 5 Bar chart visualization of quantitative comparisons on DrugBank dataset.

3.4 Comparison with the State-of-the-Art

To establish the superiority of our proposed method, we conducted comprehensive comparison experiments, pitting it against two attention-based networks (DrugBAN and HyperAttentionDTI), one transformer-based network (Moltrans), and one

Table 4 Results of quantitative comparisons on DrugBank dataset.

Settings	Acc (%)	AUC (%)	Recall (%)	AUPR (%)
MolTrans (2020) [19]	78.7	85.8	76.3	85.1
HyperAttentionDTI (2021) [21]	80.7	88.7	81.9	89.5
DrugBAN (2023) [33]	81.9	88.8	82.4	89.3
AI-Bind (2023) [34]	81.7	89.4	81.8	90.1
Ours	**83.3**	**91.1**	**83.0**	**90.8**

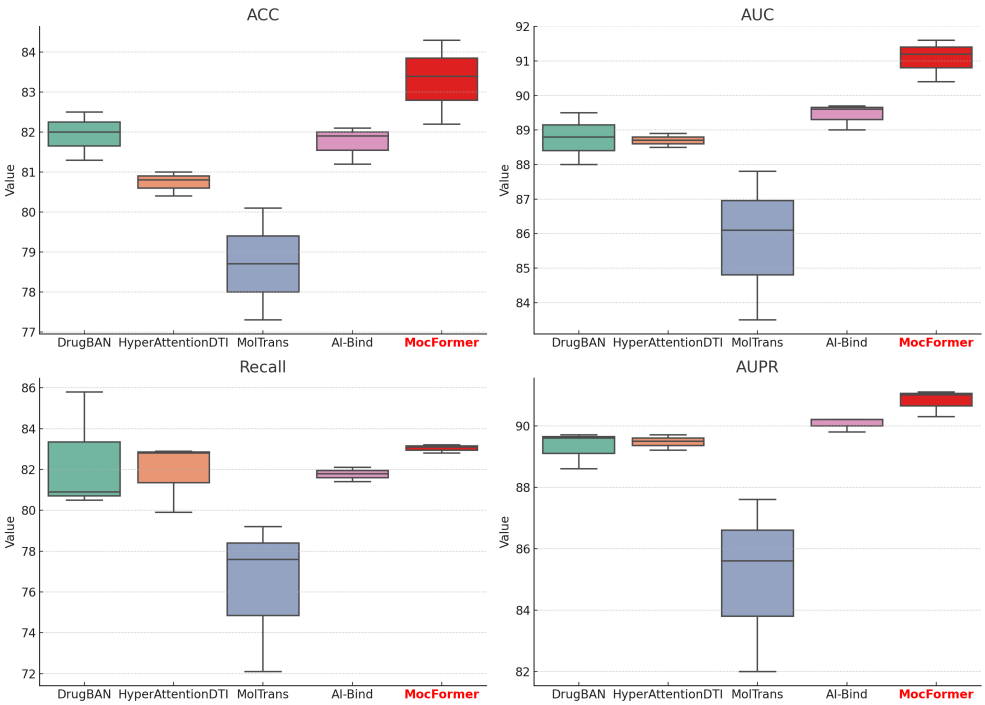

Fig. 6 Box chart visualization of quantitative comparisons on DrugBank dataset.

transfer-learning-based network (AI-Bind). These experiments were carried out using the DrugBank dataset. The results demonstrate that our method surpasses previous methods, achieving state-of-the-art (SOTA) performance across multiple critical evaluation metrics, including accuracy, AUC, recall, and AUPR. These findings are detailed in the data presented within table 4. For further clarity, our method's dominance across these metrics is visually emphasized in the bar graph displayed in figure 5. This graphical representation vividly illustrates how our approach consistently outperforms other methods in terms of their average scores. Furthermore, the box plot presented in figure 6 underscores our method's robustness and stability across these four key performance indicators.

Table 5 Results of quantitative comparisons on Dissimilar-compound-split dataset.

Settings	Acc (%)	AUC (%)	Recall (%)	AUPR (%)
MolTrans (2020) [19]	55.2	63.0	16.4	58.4
HyperAttentionDTI (2021) [21]	58.9	64.5	36.6	61.0
DrugBAN (2023) [33]	54.6	53.0	13.7	53.3
AI-Bind (2023) [34]	54.7	63.5	58.4	55.0
Ours	**59.6**	**66.1**	**60.9**	**64.5**

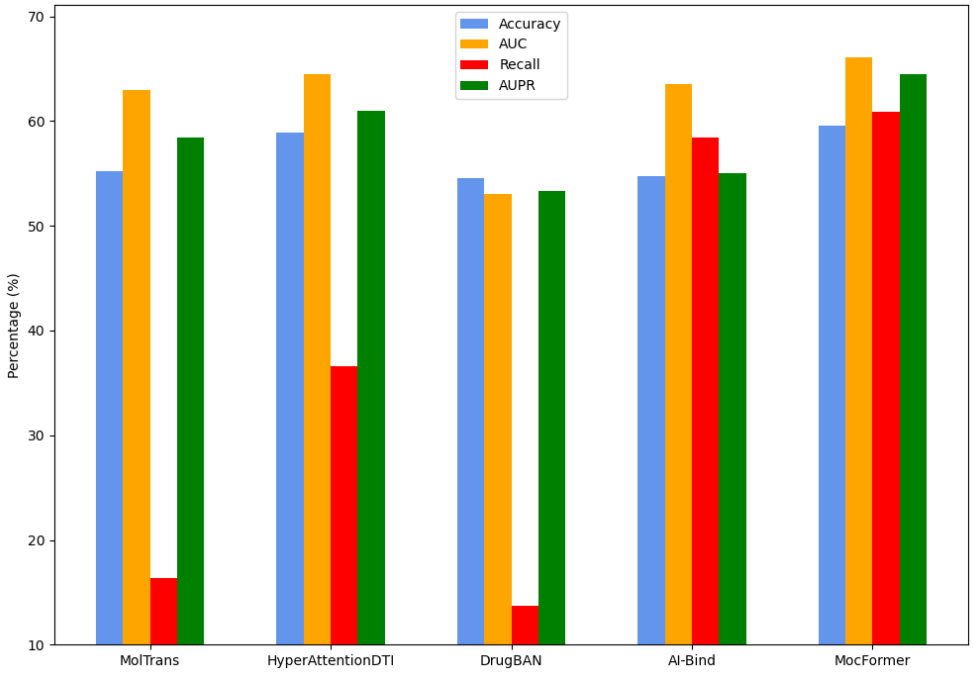

Fig. 7 Bar chart visualization of quantitative comparisons on Dissimilar-compound-split dataset.

To further showcase the remarkable generalization capabilities of our proposed approach, we conducted supplementary experiments utilizing the Dissimilar-compound-split dataset. These additional experiments' outcomes are presented in tabel 5. The bar chart figure 7, similar to the experiments conducted on the DrugBank dataset, is a compelling testament to our method's consistently superior performance. The data showcased herein reflects the mean performance of each method across an array of evaluation metrics. Furthermore, the box chart figure 8 underscores the robust nature of our approach. This is evidenced by the minimized fluctuations in the metrics, affirming the stability and reliability of our method. This body of evidence firmly

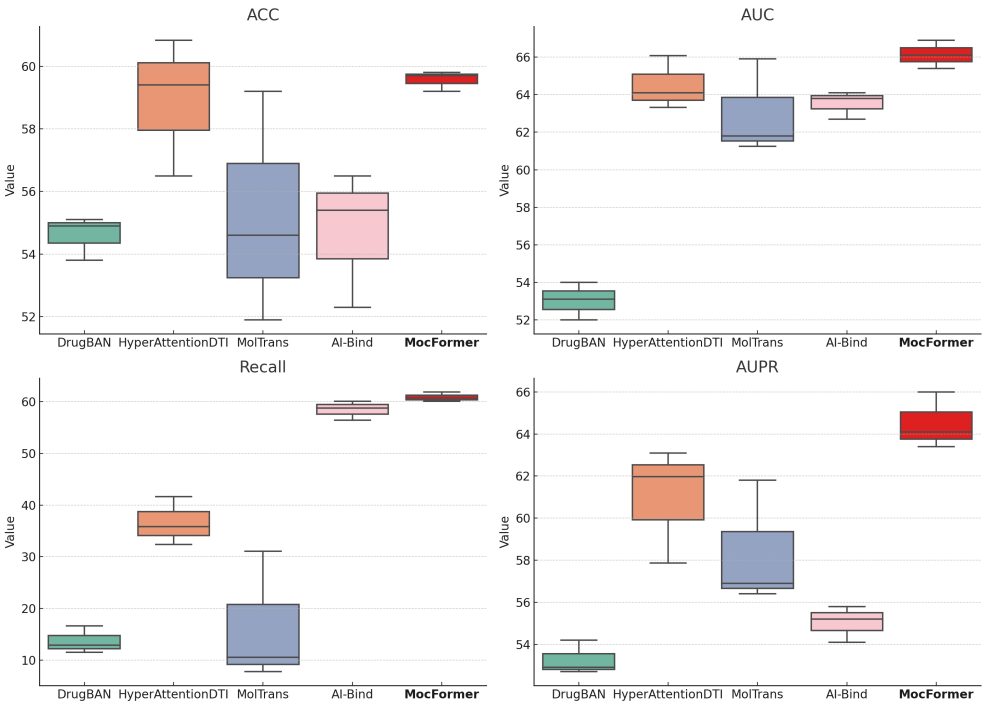

Fig. 8 Box chart visualization of quantitative comparisons on Dissimilar-compound-split dataset.

establishes that our method consistently demonstrates formidable predictive capabilities, even when exposed to previously unencountered feature data, when compared against alternative methods.

4 Conclusion

This paper introduces a two-stage pre-training driven transformer, 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 bias in the data. Quantitative and qualitative evaluations on the DrugBank and Dissimilar-compound-split databases demonstrate that our framework significantly improves accuracy and robustness, achieving state-of-the-art performance. In future work, we would like to investigate the End-to-End Learning Paradigm to optimize the final objective function directly, making it better adapt to the intricacies and complexities of the task process.

References

[1] Hopkins, A.L.: Predicting promiscuity. Nature **462**(7270), 167–168 (2009)

[2] Dudley, J.T., Deshpande, T., Butte, A.J.: Exploiting drug–disease relationships for computational drug repositioning. Briefings in Bioinformatics **12**(4), 303–311 (2011) https://doi.org/10.1093/bib/bbr013

[3] Lounkine, E., Keiser, M.J., Whitebread, S., Mikhailov, D., Hamon, J., Jenkins, J.L., Lavan, P., Weber, E., Doak, A.K., Côté, S., *et al.*: Large-scale prediction and testing of drug activity on side-effect targets. Nature **486**(7403), 361–367 (2012)

[4] Himmat, M., Salim, N., Al-Dabbagh, M.M., Saeed, F., Ahmed, A.: Adapting document similarity measures for ligand-based virtual screening. Molecules **21**(4), 476 (2016)

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems, vol. 30 (2017)

[6] Sachdev, K., Gupta, M.K.: A comprehensive review of feature based methods for drug target interaction prediction. Journal of Biomedical Informatics **93**, 103159 (2019) https://doi.org/10.1016/j.jbi.2019.103159

[7] Lee, I., Keum, J., Nam, H.: Deepconv-dti: Prediction of drug-target interactions via deep learning with convolution on protein sequences. PLoS computational biology **15**(6), 1007129 (2019)

[8] Li, Z., Liu, F., Yang, W., Peng, S., Zhou, J.: A survey of convolutional neural networks: Analysis, applications, and prospects. IEEE Transactions on Neural Networks and Learning Systems **33**(12), 6999–7019 (2022) https://doi.org/10.1109/TNNLS.2021.3084827

[9] Rogers, D., Hahn, M.: Extended-connectivity fingerprints. Journal of Chemical Information and Modeling **50**(5), 742–754 (2010) https://doi.org/10.1021/ci100050t

[10] Tsubaki, M., Tomii, K., Sese, J.: Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. Bioinformatics **35**(2), 309–318 (2018) https://doi.org/10.1093/bioinformatics/bty535

[11] Chen, W., Chen, G., Zhao, L., Chen, C.Y.-C.: Predicting drug–target interactions with deep-embedding learning of graphs and sequences. The Journal of Physical Chemistry A **125**(25), 5633–5642 (2021) https://doi.org/10.1021/acs.jpca.1c02419

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural

information processing systems **30** (2017)

[13] Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)

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

[15] Ni, Z.-L., Bian, G.-B., Zhou, X.-H., Hou, Z.-G., Xie, X.-L., Wang, C., Zhou, Y.-J., Li, R.-Q., Li, Z.: Raunet: Residual attention u-net for semantic segmentation of cataract surgical instruments. In: Proceedings of International Conference on Neural Information Processing, pp. 139–149 (2019). Springer

[16] Liu, M., Zou, W., Wang, W., Jin, C.-B., Chen, J., Piao, C.: Multi-conditional constraint generative adversarial network-based mr imaging from ct scan data. Sensors **22**(11), 4043 (2022) https://doi.org/10.3390/s22114043

[17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems, vol. 30 (2017)

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

[19] Huang, K., Xiao, C., Glass, L.M., Sun, J.: MolTrans: Molecular Interaction Transformer for drug–target interaction prediction. Bioinformatics **37**(6), 830–836 (2020)

[20] Yazdani-Jahromi, M., Yousefi, N., Tayebi, A., Kolanthai, E., Neal, C.J., Seal, S., Garibay, O.O.: AttentionSiteDTI: an interpretable graph-based model for drug-target interaction prediction using NLP sentence-level relation classification. Briefings in Bioinformatics **23**(4), 272 (2022)

[21] Zhao, Q., Zhao, H., Zheng, K., Wang, J.: HyperAttentionDTI: improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism. Bioinformatics **38**(3), 655–662 (2021)

[22] Jaeger, S., Fulle, S., Turk, S.: Mol2vec: Unsupervised machine learning approach with chemical intuition. Journal of Chemical Information and Modeling **0**(ja), (0) https://doi.org/10.1021/acs.jcim.7b00616

[23] Zhou, G., *et al.*: Uni-mol: A universal 3d molecular representation learning framework. In: Proceedings of The Eleventh International Conference on Learning Representations (2023)

[24] Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., Das, P.: Large-scale chemical language representations capture molecular structure and properties. Nature Machine Intelligence **4**(12), 1256–1264 (2022) https://doi.org/10.1038/s42256-022-00580-7

[25] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., Rives, A.: Evolutionary-scale prediction of atomic-level protein structure with a language model. Science **379**(6637), 1123–1130 (2023)

[26] Dallago, C., Schütze, K., Heinzinger, M., Olenyi, T., Littmann, M., Lu, A.X., Yang, K.K., Min, S., Yoon, S., Morton, J.T., Rost, B.: Learned embeddings from deep learning to visualize and predict protein sets. Current Protocols **1**(5), 113 (2021) https://doi.org/10.1002/cpz1.113

[27] Wei, B., Zhang, Y., Gong, X.: Deeplpi: a novel deep learning-based model for protein–ligand interaction prediction for drug repurposing. Scientific reports **12**(1), 18200 (2022)

[28] Chatterjee, A., Walters, R., Shafi, Z., Ahmed, O.S., Sebek, M., Gysi, D., Yu, R., Eliassi-Rad, T., Barabási, A.-L., Menichetti, G.: Improving the generalizability of protein-ligand binding predictions with ai-bind. Nature Communications **14**(1), 1989 (2023)

[29] Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., Woolsey, J.: Drugbank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Research **34**$(\text{suppl}_1), 668 - -672$(2006)

[30] Landrum, G., et al.: Rdkit: Open-source cheminformatics software (2016)

[31] Wen, M., Zhang, Z., Niu, S., Sha, H., Yang, R., Yun, Y., Lu, H.: Deep-learning-based drug–target interaction prediction. Journal of Proteome Research **16**(4), 1401–1409 (2017)

[32] Atas Guvenilir, H., Doğan, T.: How to approach machine learning-based prediction of drug/compound–target interactions. Journal of Cheminformatics **15**(1), 1–36 (2023)

[33] Bai, P., Miljković, F., John, B., Lu, H.: Interpretable bilinear attention network with domain adaptation improves drug-target prediction. Nature Machine Intelligence (2023) https://doi.org/10.1038/s42256-022-00605-1

[34] Chatterjee, A., Walters, R., Shafi, Z., Ahmed, O.S., Sebek, M., Gysi, D., Yu, R., Eliassi-Rad, T., Barabási, A.-L., Menichetti, G.: Improving the generalizability of protein-ligand binding predictions with AI-bind. Nature Communications **14**(1) (2023) https://doi.org/10.1038/s41467-023-37572-z