Multilevel Attention Network with Semi-supervised Domain Adaptation for Drug-Target Prediction

Zhousan Xie[1], **Shikui Tu**[1*], **Lei Xu**[1,2*]

[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[2]Guangdong Institute of Intelligence Science and Technology, Zhuhai, Guangdong 519031, China
{waduhek, tushikui, leixu}@sjtu.edu.cn

Abstract

Prediction of drug-target interactions (DTIs) is a crucial step in drug discovery, and deep learning methods have shown great promise on various DTI datasets. However, existing approaches still face several challenges, including limited labeled data, hidden bias issue, and a lack of generalization ability to out-of-domain data. These challenges hinder the model's capacity to learn truly informative interaction features, leading to shortcut learning and inferior predictive performance on novel drug-target pairs. To address these issues, we propose MlanDTI, a semi-supervised domain adaptive multilevel attention network (Mlan) for DTI prediction. We utilize two pre-trained BERT models to acquire bidirectional representations enriched with information from unlabeled data. Then, we introduce a multilevel attention mechanism, enabling the model to learn domain-invariant DTIs at different hierarchical levels. Moreover, we present a simple yet effective semi-supervised pseudo-labeling method to further enhance our model's predictive ability in cross-domain scenarios. Experiments on four datasets show that MlanDTI achieves state-of-the-art performances over other methods under intra-domain settings and outperforms all other approaches under cross-domain settings. The source code is available at https://github.com/CMACH508/MlanDTI.

Introduction

The process of drug discovery and development is characterized by its high costs and time-intensive nature. Bringing a first-in-class drug to the market typically requires several decades and substantial investments amounting to billions of dollars. Predicting drug-target interactions (DTIs) is an essential task in drug discovery and drug repurposing (Paul et al. 2010), which hold significant value in the field of biomedicine (Agamah et al. 2020; Ezzat et al. 2019). While traditional techniques like high-throughput screening, proteomics, and genomics remain prevalent, they suffer from time and cost constraints due to the vast chemical space involved (Broach, Thorner et al. 1996; Bakheet and Doig 2009).

In order to expedite the drug discovery process and reduce costs, virtual screening (VS) techniques have been developed to aid *in silico* (Rifaioglu et al. 2019). Molecular docking and molecular simulations have shown great success in drug discovery (Cheng et al. 2012), but they are limited for being computationally resource-intensive and relying on the availability of 3D structure data. Methods including machine learning approaches (Faulon et al. 2008; Wang et al. 2021; Meng et al. 2017) perform well in predicting DTIs for known drug-target pairs, while their performance tends to deteriorate when applied to unknown structures.

With the accumulation of a large volume of labeled DTI data in recent years, numerous end-to-end deep learning methods have been employed for predicting DTIs. From the perspective of the input data, DTI prediction models can be categorized into three groups. The first category is sequence-based models, where drugs are represented as Simplified Molecular Input Line Entry System (SMILES) or Extended-Connectivity Fingerprints (ECFP) and proteins are treated as amino acid sequences. These models commonly utilize 1D-CNN (Öztürk, Özgür, and Ozkirimli 2018; Lee, Keum, and Nam 2019; Zhao et al. 2022; Bai et al. 2023) or transformer architectures (Chen et al. 2020; Huang et al. 2022). Secondly, drug molecules can be represented as graphs (Nguyen et al. 2021; Tsubaki, Tomii, and Sese 2019; Huang et al. 2022) or images (Qian, Wu, and Zhang 2022). Similarly, protein distance maps can serve as a 2D abstraction of their 3D structural information (Zheng et al. 2020), enabling the use of Graph Neural Networks (GNNs) (Scarselli et al. 2008), Graph Convolutional Networks (GCNs) (Kipf and Welling 2016), and Convolutional Neural Networks (CNNs). Thirdly, the incorporation of 3D structural data such as protein pockets (Yazdani-Jahromi et al. 2022) or molecular dynamics simulation data (Wu et al. 2022) undoubtedly improves model performance and reduces computational complexity compared to those directly using the whole 3D data as input (Wallach, Dzamba, and Heifets 2015; Stepniewska-Dziubinska, Zielenkiewicz, and Siedlecki 2018). Nonetheless, they are still constrained by the limited availability of 3D structural data.

Despite these remarkable development, deep learning methods still face several challenges. The first challenge is the restriction of limited labeled data. Previous works have primarily concentrated on utilizing the available labeled data and learn interactions on a few thousands labeled drug-target pairs (Öztürk, Özgür, and Ozkirimli 2018; Lee, Keum, and Nam 2019; Tsubaki, Tomii, and Sese 2019; Nguyen et al.

*Corresponding authors

2021; Huang et al. 2022; Zhao et al. 2022; Bai et al. 2023). However, these approaches often overlook the enormous amount of unlabeled biomedical data, which hinders the models from fully leveraging the chemical structures and interactions of drugs and proteins. Consequently, the models struggle to extract truly informative features, leading to limited generalization ability.

The second challenge is the hidden bias and shortcut learning. The issue of hidden bias has been reported on the DUD-E and MUV datasets (Sieg, Flachsenberg, and Rarey 2019). It has been observed that models trained on the DUD-E dataset (Chen et al. 2019) and other datasets (Chen et al. 2020), tend to rely predominantly on drug patterns when making predictions, rather than capturing the comprehensive interaction between drugs and targets. This lead to a gap between theoretical modeling and practical application. We further identify two main reasons for this issue: 1) The presence of a greater variety and quantity of drug molecules in datasets than proteins; 2) The inherent ease of feature extraction for drug molecules compared to proteins. These factors result in shortcut learning, where the model tends to prioritize learning features from the larger and easier-to-learn drug molecule data, rather than focusing on the features of proteins. Consequently, the model struggles to effectively capture the interaction features between drugs and proteins.

The third challenge lies in the model's ability to generalize and make predictions on out-of-domain data, which is closely related to the previous two challenges. Developing a first-in-class drug often involves predicting interactions with a completely new target using novel compounds, which may have a distribution that differs significantly from the data on which the model was trained. Thus, the model needs to be capable of cross-domain generalization (Abbasi et al. 2020; Bai et al. 2023; Kao et al. 2021). Currently, most models are trained on limited labeled data and fail to address the issue of shortcut learning, resulting in limited ability to predict interactions between completely new drugs and proteins.

To tackle the three challenges, we propose MlanDTI, a semi-supervised domain adaptive multilevel attention network for DTI prediction. We utilize two pre-trained BERT models to acquire bidirectional embeddings of protein and SMILES (drug) sequences from millions of unlabeled data. Inspired by the least mean squared error reconstruction (Lmser) network (Xu 1993; Huang, Tu, and Xu 2022), we then devise a variant of transformer with a multi-level attention mechanism with drug and protein embeddings as input. It enables the joint extraction of both drug and target features with reduced hidden bias and facilitates the learning of multi-level interactions. Moreover, we incorporates a simple yet effective semi-supervised pseudo-labeling method to further enhance our model's predictive ability in cross-domain scenarios. Experiments on four datasets demonstrate that MlanDTI achieves state-of-the-art performances over other methods under intra-domain settings and outperforms all other approaches under cross-domain settings.

The main contributions are three-fold as follows:

- To leverage massive unlabeled biomedical data, we employed two pre-trained BERT models to acquire representations that possess better robustness and generalization capabilities. We observed that the representations obtained by the BERT models significantly enhance the accuracy of pseudo-labeling.
- We propose a novel multi-level attention mechanism which enables effective feature extraction by allowing the model to dynamically focus on different aspects of proteins and drugs during the learning process. The attention mechanism mitigates the shortcut learning problem and reduces the impact of hidden bias on predictions.
- We propose a simple yet effective pseudo-label domain adaptation method, which significantly reduces the noise of pseudo-labels.

Related Work

Leveraging Additional Data

One of the key to DTI prediction is how to represent drug molecules and proteins that allows the model to learn useful features. Learning from 3D structural information (Wallach, Dzamba, and Heifets 2015; Stepniewska-Dziubinska, Zielenkiewicz, and Siedlecki 2018) is undoubtedly the most direct approach, but it is limited by the high computational costs and model complexity. Another indirect approach is to provide additional data containing 3D structural information, such as molecular dynamics simulations (Wu et al. 2022) and protein pocket data (Yazdani-Jahromi et al. 2022). While the aforementioned methods are limited by the availability of a finite amount of 3D structural data, Moltrans (Huang et al. 2021), in contrast, leverages a vast amount of unlabeled protein and drug sequences by using Frequent Consecutive Sub-sequence (FCS) algorithm to extract high-quality substructures and enhances the representations using transformers. However, FCS has certain limitations in its ability to comprehensively extract information from sequence data, and the quantity of unlabeled data utilized is also insufficient. In this paper, we utilize two pre-trained BERT (Devlin et al. 2018) models learned on a large amount of unlabeled data to obtain rich representations of proteins and drug sequences with powerful generalization abilities.

Learning Interactions

Proteins and drugs are two fundamentally different types of data, and the task of DTI prediction requires the model to learn their interaction features. The simplest approach is to concatenate the features (Öztürk, Özgür, and Ozkirimli 2018; Lee, Keum, and Nam 2019; Zheng et al. 2020; Nguyen et al. 2021) and pass them through a Fully-Connected Network (FCN) to obtain the prediction results. Another approach (Qian, Wu, and Zhang 2022) is to overlap the feature maps and use CNN to extract interaction features. However, these methods lack interpretability and overlook the inherent structure of interactions. Recently, attention mechanisms have been demonstrated effective in capturing intricate interactions between proteins and drugs. Multi-head attention (Bian et al. 2023; Chen et al. 2020) and other attention variations (Bai et al. 2023; Zhao et al. 2022) have been widely applied in DTI prediction. However, (Chen et al. 2020) found that the hidden bias in some datasets that

led models to rely mainly on drug patterns rather than the interactions for prediction. We further observed that this issue was prevalent in existing models. To address this issue, we proposed a multi-level attention mechanism.

Domain Generalization in DTI Predictions

In previous works (Huang et al. 2021; Yazdani-Jahromi et al. 2022; Zhao et al. 2022), the evaluation of model generalization was often conducted through the partitioning of datasets into "unseen drug" or "unseen protein" scenarios, where drugs or proteins were only present in the test set. However, such evaluations still fall into the intra-domain setting, different from real-world applications. Currently, there is limited research on domain generalization in DTI prediction. DrugBAN (Bai et al. 2023) addresses this challenge by utilizing Conditional Domain Adversarial Network (CDAN) to transfer the learned knowledge from the source domain to the target domain, thereby enhancing the model's performance in cross-domain settings. Here, we leverage pseudo-labeling techniques to mitigate the distribution discrepancy between the target and source domains. Through the integration of an auxiliary classifier and the powerful representational capacity of BERT models, our approach significantly improves the accuracy of pseudo-labeling. Under the cross-domain setting, our method demonstrates remarkable performance surpassing that of DrugBAN.

Method

Problem Formulation

The task of DTI prediction aims to determine whether a drug compound and a target protein will interact. For drug compounds, most existing deep learning methods utilize the SMILES strings to represent the drugs. Specifically, a drug is represented as $\mathbf{D} = (d_1, ..., d_m)$, where d_i is a SMILES symbol with chemical meanings such as atoms, m is the length. As for target proteins, each protein sequence is represented as $\mathbf{T} = (a_1, ..., a_n)$, where a_i corresponds to one of the 23 amino acids, n is the length of the protein sequence.

Given a drug SMILES sequence $\mathbf{D}$ and a protein sequence $\mathbf{T}$, the objective is to train a model to assign an interaction probability score $P \in [0, 1]$ by mapping the joint feature representation space $\mathbf{D} \times \mathbf{T}$.

The Proposed Framework

An overview of MlanDTI is depicted in Figure 1. It commences by encoding the drug and target sequences into vector embeddings via pre-trained BERT models, i.e., ChemBERTa-2 (Ahmad et al. 2022) and ProtTrans (Elnaggar et al. 2021). Subsequently, these embeddings are passed through the encoder and decoder of a modified transformer architecture with a multi-level attention module to extract interaction features. The classifier comprises a bilinear attention module and a max pooling layer, followed by a FCN for prediction. For cross-domain prediction, we employ an auxiliary classifier that directly accepts BERT outputs. It aids in learning implicit distributional information from BERT representations, thereby enhancing pseudo-label accuracy. After training the two classifiers on labeled source domain data, the model predicts on unlabeled target domain data to obtain pseudo-labels. The pseudo-label learning process consists learning high-confidence pseudo-labels and minimizing conflicting predictions.

Encoder for Protein Sequence We build the encoder by adopting a modification on the transformer similar to TransformerCPI (Chen et al. 2020). Instead of using the self-attention module, we utilize a 1D-CNN and GLU (gated linear unit) (Dauphin et al. 2017) as alternatives. The hidden layers $h_0, ..., h_L$ in the encoder are computed as:

$$h_i(\mathbf{X}_T) = (\mathbf{X}_T \mathbf{W}_{i1} + \mathbf{s}) \otimes \sigma(\mathbf{X}_T \mathbf{W}_{i2} + \mathbf{t}), \quad (1)$$

where $\mathbf{X}_T \in \mathbb{R}^{n \times m_1}$ is the input of layer h_i, $\mathbf{W}_{i1} \in \mathbb{R}^{k \times m_1 \times m_2}, \mathbf{s} \in \mathbb{R}^{m_2}, \mathbf{W}_{i2} \in \mathbb{R}^{k \times m_1 \times m_2}, \mathbf{t} \in \mathbb{R}^{m_2}$ are parameters, n is the input sequence length, k is the patch size, m_1, m_2 are the dimensions of input and hidden vectors, σ is the sigmoid function, and $\otimes$ is the element-wise product.

Since the length of a protein sequence may range in the thousands or even tens of thousands, the self-attention module in transformers poses a significant computational and memory burden with $\mathcal{O}(n^2)$ time and space complexity, and is prone to overfitting when working on small datasets. The above modification by Eq. (1) mitigates the computational and storage burden on long protein sequences and remedies overfitting on small datasets.

Multilevel Cross-Attention For the task of DTI prediction, the most crucial ability for the model is to learn the interaction patterns between drugs and targets. It involves aligning the features of proteins with the features of drugs in a shared feature subspace. However, extracting multi-level features from proteins is more challenging than extracting features from drugs, because protein sequences are notably long, with intricate multi-level structures, while drugs are often small chemical molecules. This difference contributes to the hidden bias in DTI models (another is inherent dataset bias). Aligning protein features with drug features also requires a multi-level process, but the model may not capture the multi-level features of proteins well and effectively align them with drug features. Thus, the existing models tend to learn a shortcut by relying on the features of drug molecules to predict drug-target interactions.

In an early literature (Xu 1993), Lmser network was proposed to enhance the representation learning by building bidirectional skip connections on every levels of layers between encoder and decoder. It was first demonstrated in a deep CNN implementation to be robust and effective on image processing (Huang, Tu, and Xu 2022; Xu 2019), and then Lmser-transformer was developed to improve the molecular representation learning by adding hierarchical connections to the original transformer (Qian et al. 2022). Inspired by these works, we propose a multi-level cross-attention mechanism to address this issue, as illustrated in Figure 1(a). In the vanilla transformer, the encoder uses the protein features from the last layer of the encoder as the Key and Value for the cross-attention layer of the decoder, aligning them with the drug features in the decoder. However, the protein features obtained from the encoder's output do not fully capture the expression of the multi-level structural

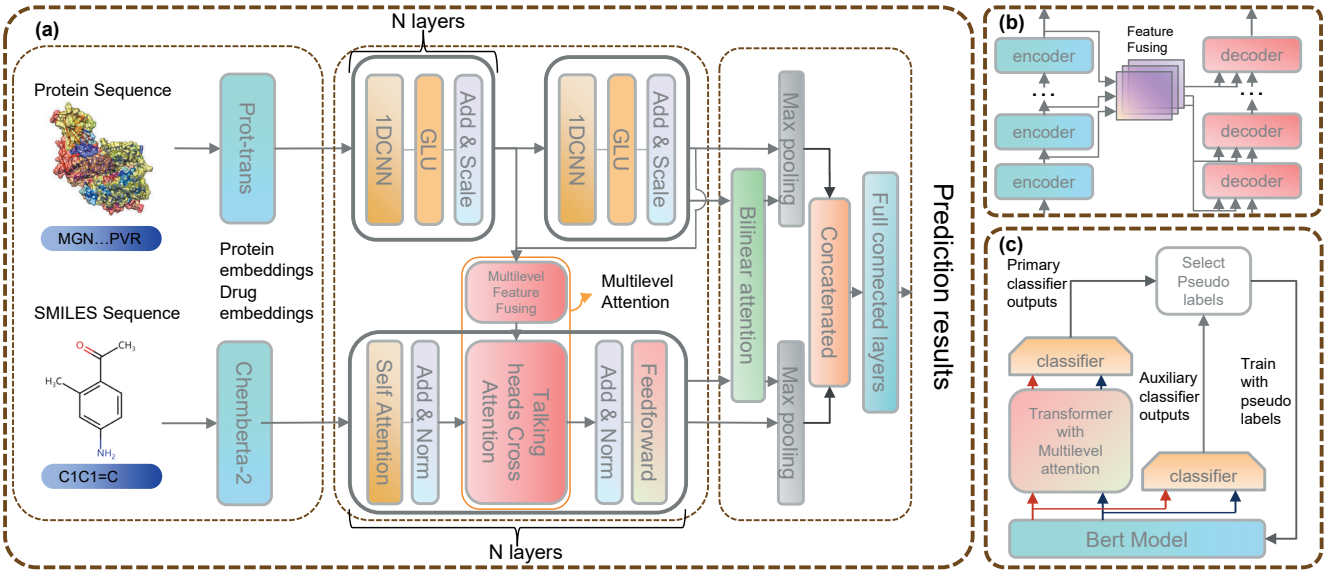

Figure 1: (a). The overall framework of MlanDTI, it consists of two pre-trained BERT models that convert SMILES and amino acid sequences into vector embeddings. The Encoder and Decoder are connected by a multilevel attention module, and the final output is processed through the classifier with a bilinear attention module and a max pooling layer before being fed into FCN to generate the prediction results. (b). The detail of Multilevel attention. (c). Training with pseudo labeling with an auxiliary classifier.

information of proteins, and they do not align with drug features at different levels in the decoder.

Here, we develop the multi-level attention mechanism by two steps: 1) the multi-level feature fusing step and 2) the cross-attention feature aligning step. Suppose the protein feature matrices of each encoder layers are $T_0, ..., T_n \in \mathbb{R}^{m\times d}$, where n is the number of transformer layers, m is the size of the protein sequence and d is the vector dimension. For the ℓ-th decoder layer, we concatenate the protein feature matrices from the preceding ℓ layers to form $T_{cat_\ell} = [T_0, ..., T_\ell] \in \mathbb{R}^{\ell\times m\times d}$. Then, we perform a cross-layer feature aggregation by applying a fusion matrix $F_\ell \in \mathbb{R}^{\ell\times 1}$. This results in multi-level fused protein feature matrix $T'_\ell = F_\ell^T T_{cat_\ell}$. To summarize, we compute all T'_ℓ as:

$$\text{diag}(T'_0, ..., T'_n) = \mathbf{F} \cdot \text{diag}(T_{cat_0}, ..., T_{cat_n}), \quad (2)$$

where $\mathbf{F}$ is a learnable diagonal matrix with each diagonal element being F_ℓ from each layer, i.e., $\ell = 0, ..., n$. Then, the query, key, and value for the multi-level cross-attention mechanism at the ℓ-th layer are respectively computed by

$$Q = D_\ell W_q, \quad K = T'_\ell W_k, \quad V = T'_\ell W_v, \quad (3)$$

where D_ℓ is the drug feature matrix which has passed the self-attention module, and T'_ℓ the multi-level protein feature matrix given by Eq. (2).

To enhance the extraction capabilities of attention heads for multi-level interactions, we incorporate the talking-heads attention mechanism (Shazeer et al. 2020) for feature alignment. This variation of multi-head attention in the transformer introduces two additional linear projections. These projections transform the attention logits and the attention weights, respectively, allowing the flow of information across different attention heads and improving the overall performance of the model, i.e.,

$$\text{Attention}(Q, K, V) = \text{softmax}\left(P_\ell \frac{QK^T}{\sqrt{d_k}}\right) P_w V, \quad (4)$$

where Q, K, V are given by Eq. (3), and $P_\ell \in \mathbb{R}^{h_k\times h_k}$, $P_w \in \mathbb{R}^{h_k\times h_v}$ are the two additional linear projections. h_k represents the number of attention heads for keys and queries, and h_v denotes the number of attention heads for values, and they can optionally differ in size.

The advantages of the proposed multi-level attention mechanism are briefly summarized below.

- **Encourage multi-level feature learning**: By fusing protein features, drug features are derived to interact with relevant characteristics, which thereby captures multi-level interaction features, leading to a more comprehensive understanding of drug-target interactions.
- **Alleviate hidden bias and reduce overfitting**: Multi-level attention encourages the model to focus more on hierarchical interaction features, the model becomes less prone to biased representations that might emerge from focusing solely on specific patterns, and thus the model is less likely to overfit to noisy patterns of the data.
- **Improve generalization abilities**: Multi-level attention enables the model to learn domain-invariant interaction features. These representations exhibit robustness and enhance transferability across different data domains.

The Classifier The classifier consists of the bilinear attention module from hyperattentionDTI (Zhao et al. 2022)

to further extract bidirectional interaction features. Subsequently, we utilize a multi-layer FCN with each layer followed by a leaky ReLU activation function (He et al. 2015) to combine these features and generate prediction results. Since it is a binary classification problem, we utilize the binary cross-entropy loss function to train the model.

$$\mathcal{L}_{CE} = -[y \log \hat{y} + (1-y)\log(1-\hat{y})], \quad (5)$$

where y is the ground truth label, $\hat{y}$ is the classifier's output.

Pseudo Label Learning for Domain Adaptation Pseudo-labeling (Lee et al. 2013) is a semi-supervised learning (SSL) method that utilizes a model trained on labeled source domain data to generate pseudo-labels for unlabeled target domain data. By incorporating these pseudo-labels into the training process, the model can adapt to target domain, which is particularly suitable for DTI predictions where labeled data are limited and unlabeled data are massive. However, in other domains, pseudo-labeling based SSL methods often suffer from poor model performance due to the presence of noisy pseudo-labels (Rizve et al. 2021). Here, we propose a simple yet effective approach that significantly improves the accuracy of generated labels and reduces the impact of noisy labels on the model.

Our method consists of two steps. In the first step, we perform the selection and learning of high-confidence pseudo-labels. To achieve this, we introduce an auxiliary classifier for co-training, which is essentially the classifier mentioned earlier but directly takes BERT representations as input. Let $\mathbf{P_1} = \{p_1^{(i)}\}_{i=1}^N, \mathbf{P_0} = \{p_0^{(i)}\}_{i=1}^N$ and $\mathbf{P}_{1,aux} = \{p_{1,aux}^{(i)}\}_{i=1}^N, \mathbf{P}_{0,aux} = \{p_{0,aux}^{(i)}\}_{i=1}^N$ be the probability outputs of the model and the auxiliary classifier for target domain data $X_t = \{x^{(i)}\}_{i=1}^N$, respectively, such that $p_0^{(i)}, p_{0,aux}^{(i)}$ is the probability of no interaction for sample $x^{(i)}$ and $p_1^{(i)}, p_{1,aux}^{(i)}$ is the probability the sample interact. Rather than selecting thresholds, which we observed may lead to unbalanced pseudo-labels, we sort $(\mathbf{P_1} + \mathbf{P}_{1,aux})$ and $(\mathbf{P_0}+\mathbf{P}_{0,aux})$ in descending order and select the top M positive and negative sample pairs based on their probabilities to assign pseudo-labels:

$$Y_1 = \{\hat{y}_1^{(i)} = 1 | p_1^{(i)} + p_{1,aux}^{(i)} \in \text{top}_M(\mathbf{P}_1 + \mathbf{P}_{1,aux})\}, \quad (6)$$

$$Y_0 = \{\hat{y}_0^{(i)} = 0 | p_0^{(i)} + p_{0,aux}^{(i)} \in \text{top}_M(\mathbf{P}_0 + \mathbf{P}_{0,aux})\}, \quad (7)$$

where Y_1, Y_0 represent the sets of pseudo-labels for positive and negative samples, respectively, and M is the number of selected samples which grows with the number of iterations.

The auxiliary classifier focuses on learning the latent relationships between target domain and source domain data within the BERT representations, while the main model prioritizes learning domain-invariant DT interaction features. This leads classifier discrepancy in nature, enabling higher accuracy for pseudo-label with high confidence on both classifiers. After generating pseudo-labels, we employ the cross-entropy loss to train model, i.e.,

$$\mathcal{L}_{pseudo} = -\frac{1}{2M}\sum_{i=1}^{2M}[y \log \hat{y}^{(i)} + (1-y)\log(1-\hat{y}^{(i)})]. \quad (8)$$

The second step is to penalize conflict predictions. Let X_d be the set of samples for which the two classifiers exhibit conflicting classifications, i.e.,

$$X_d = \{x^{(i)} | x^{(i)} \in X_t, \text{argmax}\, \mathbf{p}^{(i)} \neq \text{argmax}\, \mathbf{p}_{aux}^{(i)}\}, \quad (9)$$

where $\mathbf{p}^{(i)} = (p_0^{(i)}, p_1^{(i)}), \mathbf{p}_{aux}^{(i)} = (p_{0,aux}^{(i)}, p_{1,aux}^{(i)})$.

We randomly select a subset, X_d', of size M', from X_d, where the value of M' increases with the number of model iterations. We utilize a modified binary cross entropy loss to augment the prediction uncertainty for conflicting samples between the two classifiers, i.e.,

$$\mathcal{L}_{conf} = -\frac{1}{M'}\sum_{i=1}^{M'}[y \log 0.5 + (1-y)\log(1-0.5)]. \quad (10)$$

Both steps enable the model to acquire pseudo-labels with reduced noise for training, consequently enhancing the model's performance in the target domain.

Experiments

Datasets

We evaluated our model on the human dataset, Caenorhabditis elegans dataset (Tsubaki, Tomii, and Sese 2019), bindingdb dataset (Liu et al. 2007), and Biosnap dataset (Huang et al. 2021). Specifically, we conducted both intra-domain and cross-domain tests on the BindingDB and Biosnap datasets. For the intra-domain evaluation, we randomly split the dataset into training, validation, and test sets with a ratio of 8:1:1 in smaller human and C.elegans datasets, and 7:1:2 in larger BindingDB and Biosnap datasets. We also conducted cold pair split experiments on BindingDB and Biosnap datasets. We select 70% of drugs/proteins randomly, and all related DT pairs were collected as the training set. Subsequently, DT pairs in the remaining 30% were split into a 3:7 ratio, as validation set and test set. This ensures that all drugs and proteins in the test set are unseen to model.

For the cross-domain evaluation, we followed the clustering-based split strategy used in DrugBAN. We applied the ECFP4 and PSC algorithms to cluster drugs and proteins, respectively. Then, we randomly selected 60% of the drug and protein clusters and used all drug-protein pairs belonging to these clusters as the source domain data. The drug-protein pairs in the remaining 40% clusters were used as the target domain data. This data partitioning ensured that the target domain and source domain data were from disjoint distributions, making the evaluation more challenging and enabling a true assessment of the model's ability to predict interactions for unknown proteins and molecules.

For the domain adaptation setting, we used all labeled source domain data and 80% of the unlabeled target domain data as the training set. This 80% of the target domain data was also used as the validation set, while the remaining 20% of labeled data from the target domain served as the test set.

Baselines and Implementation Details

We conducted a comparison between our proposed method and eight baseline approaches: Support Vector Machine

methods	human			C.elegans			BindingDB			BioSNAP		
	AUC	AUPR	F1	AUC	AUPR	F1	AUC	AUPR	F1	AUC	AUPR	F1
SVM	0.910	–	**0.967**	0.894	–	0.801	0.939	0.928	0.787	0.862	0.864	0.762
RF	0.940	–	0.878	0.902	–	0.832	0.942	0.921	0.858	0.860	0.886	0.808
GraphDTA	0.960	0.959	0.897	0.974	0.975	0.919	0.951	0.934	0.867	0.887	0.890	0.789
DeepConvDTI	0.967	0.964	0.922	0.983	0.985	0.944	0.945	0.925	0.859	0.886	0.890	0.797
MolTrans	0.974	0.976	0.944	0.982	0.985	**0.966**	0.952	0.936	0.865	0.895	0.897	0.824
TransformerCPI	0.973	0.975	0.920	0.988	0.986	0.952	0.943	0.925	0.855	0.889	0.893	0.798
HyperAttDTI	0.984	0.984	0.946	0.989	0.990	0.958	**0.959**	**0.948**	**0.887**	0.901	0.902	0.838
DrugBAN	0.981	0.983	0.940	0.986	0.988	0.949	**0.959**	0.947	0.881	0.903	0.902	0.832
Ours	**0.988**	**0.990**	0.961	**0.990**	**0.992**	0.962	0.945	0.926	0.857	**0.909**	**0.912**	**0.841**

Table 1: The results of the proposed model and baslines on four datasets (5 random runs), Metric: AUROC (AUC), AUPRC (AUPR), F1-score (F1), The best results are indicated by bold. ”–” means no result for this metric.

methods	cold						cross-domain					
	BindingDB			BioSNAP			BindingDB			BioSNAP		
	AUC	AUPR	F1	AUC	AUPR	F1	AUC	AUPR	F1	AUC	AUPR	F1
Moltrans	0.595	0.522	0.511	0.672	0.697	0.437	0.537	0.476	0.389	0.632	0.635	0.401
TransformerCPI	0.656	0.594	0.566	0.680	0.708	0.523	0.568	0.450	0.410	0.656	0.693	0.432
HyperAttDTI	0.661	0.598	0.582	0.732	0.760	0.539	0.545	0.462	0.376	0.654	0.685	0.395
DrugBAN	0.655	**0.600**	0.542	0.651	0.667	0.449	0.578	0.471	0.484	0.608	0.606	0.438
DrugBAN$_{CDAN}$	NA	NA	NA	NA	NA	NA	0.616	0.512	0.426	0.673	0.706	0.542
Ours	**0.671**	0.594	**0.601**	**0.782**	**0.801**	**0.653**	0.657	0.537	0.489	0.728	0.759	0.604
Ours (with PL)	NA	NA	NA	NA	NA	NA	**0.687**	**0.579**	**0.564**	**0.749**	**0.770**	**0.629**

Table 2: In-domain (cold pair split: unseen drugs & proteins) and cross-domain (clustering-based split) comparison on the BindingDB and BioSNAP datasets (5 random runs). 1) Underlined values explanation: We chose a threshold of 0.5 (the same one as in MolTrans) to calculate the F1-score of DrugBAN. This is to ensure a fair comparison and to avoid ineffective classification caused by overly low thresholds in DrugBAN. Further information is provided in the appendix. 2) NA, not applicable to this study. 3) The term “with PL” within parentheses refers to our method that incorporates the pseudo-labeling module.

(SVM) (Cortes and Vapnik 1995), Random Forest (RF) (Ho 1995), GraphDTA (Nguyen et al. 2021), DeepConv-DTI (Lee, Keum, and Nam 2019), MolTrans (Huang et al. 2021), TransformerCPI (Chen et al. 2020), HyperattentionDTI (Zhao et al. 2022), and DrugBAN (Bai et al. 2023). These baselines encompass both classic machine learning methods and the current state-of-the-art deep learning approaches, ensuring a comprehensive comparison. All deep learning methods were employed with their default configurations as provided by their respective authors. Our proposed method in implemented in PyTorch, utilizing the Adam optimizer with an initial learning rate of 0.001. Detailed hyperparameter settings are provided in the appendix.

Intra-domain Experiments

Table 1 displays the comparison on the human and C.elegans datasets. These two datasets are relatively small, with balanced positive and negative samples, enabling us to evaluate the model’s predictive ability within the same distribution. Our method outperforms all deep learning baselines in terms of AUROC and AUPRC, and it also exhibits competitive performance in terms of F1-score.

We also conducted comparisons on the larger datasets, BindingDB and BioSNAP. In the random split tests, our model achieved state-of-the-art performance on the BioSNAP dataset, but its performance on the BindingDB dataset was not particularly competitive. This discrepancy was due to the hidden bias issue present in the BindingDB dataset.

The BindingDB dataset contains 14643 drugs and 2623 proteins, which results in an extremely imbalanced drug-to-protein ratio compared to the other datasets (BioSNAP: 4510 / 2181, human: 2726 / 2001, C.elegans: 1767 / 1876). Compared to the other three datasets, deep learning models even struggle to outperform traditional machine learning methods (AUC: RF 0.942, deepConv-DTI 0.945) on the BindingDB dataset. Previous studies (Bai et al. 2023) have also reported that the performance in the BindingDB dataset under unseen-drug setting shows minimal decline compared to random splits. This phenomenon is attributed to the presence of a large number of highly similar molecules in the dataset, which makes it challenging for the naive unseen-drug setting to distinguish between them. The excessive number of highly similar drug samples causes baseline models to lean towards learning drug patterns rather than drug-target interactions for prediction. As a result, deep learning and machine learning methods exhibit similar performance levels. However, this shortcut learning approach contradicts the original intent of DTI prediction and cannot be considered reliable in practical applications.

However, our model focuses more on learning the multi-level interactions between proteins and drugs. In the cold split setting in Table 2, the model can only learn drug-target

Ablation	BindingDB			BioSNAP		
	AUC	AUPR	F1	AUC	AUPR	F1
	0.687	0.579	0.564	0.749	0.770	0.629
-BERT	0.573	0.455	0.413	0.648	0.671	0.499
-MLA	0.628	0.511	0.523	0.731	0.753	0.585
-PL	0.657	0.537	0.489	0.728	0.759	0.604
-Aux Cls	0.626	0.486	0.503	0.739	0.776	0.633

Table 3: Ablation study on BindingDB and BioSNAP datasets (cross-domain,five random runs)

interaction features, due to the lack of sufficiently similar drug and protein molecules as references. Our model outperforms other baselines on the BindingDB dataset, while on the more balanced BioSNAP dataset, our model achieves a superior performance compared to the baselines.

Overall, the challenges posed by the hidden bias issue on the BindingDB dataset highlight the importance of our model's ability to capture multi-level drug-target interactions, which allows it to perform well in scenarios where other baselines struggle to maintain effectiveness.

Cross-Domain Experiments

Table 2 presents a comparison of model performance on the BindingDB and BioSNAP datasets under the cross-domain setting. Compared to the intra-domain setting, the majority of models experience significant performance drop due to the differences in data distributions. Particularly, for the BindingDB dataset, the clustering-based strategy ensures that there are no similar drugs or proteins between the training and testing sets, preventing the models from relying on drug patterns. This breaks the false high-performance illusion observed in the intra-domain scenario, and some models even show no better performances than random guessing (AUC: 0.5). Among all baselines, DrugBAN$_{CDAN}$, which leveraged a conditional domain adversarial network (CDAN) for domain adaptation, achieved the best performance. However, DrugBAN$_{CDAN}$ did not surpass our vanilla model with out pseudo labeling, and our model with pseudo labeling significantly outperformed all state-of-the-art models, including DrugBAN with domain adaptation module. Specifically, our model outperformed DrugBAN$_{CDAN}$ by 11.52% and 11.29% (AUROC) on the BindingDB and BioSNAP datasets, respectively.

Ablation Studies

We conducted ablation studies in Table 3 under the cross-domain setting on the BindingDB and BioSNAP datasets to analyze the effectiveness of modules in our proposed model.

Effectiveness of BERT Embeddings We replaced BERT with Word2Vec and GCN as used in TransformerCPI (Chen et al. 2020) to obtain embeddings for drugs and proteins. As shown in Table 3, the performance of the model experienced a notable decline. This outcome can be attributed to the auxiliary classifier's inability to effectively capture the implicit relationship between the source and target domains through the representations. As a result, in Figure 2 the accuracy of pseudo-labels exhibited a significant drop, introducing a

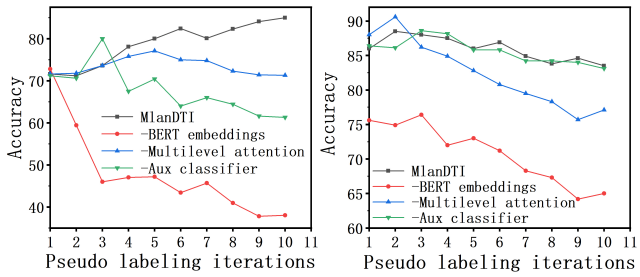

Figure 2: Ablation experiments of Pseudo labeling accuracy on (a) BindingDB dataset (b) BioSNAP dataset

substantial amount of noisy pseudo-labels that deteriorated the model's performance.

Effectiveness of Multilevel Attention We replaced the multilevel attention (MLA) mechanism with the original Transformer multi-head attention. However, on both datasets, the model exhibited performance drop in varying degrees. With an increase in training iterations, a significant decline in the accuracy of pseudo-labels was observed. It turns out that the multilevel attention mechanism is better equipped to capture domain-invariant drug-target interaction features, thereby enhancing the model's performance in the target domain.

Effectiveness of Pseudo Labeling and Auxiliary Classifier Pseudo-labeling (PL) proves effective in enhancing the model's performance within the target domain. Concurrently, auxiliary classifiers contribute to reducing the noise within these pseudo-labels. This effect is particularly pronounced in BindingDB dataset, which exhibits substantial disparities in domain distributions. The absence of auxiliary classifiers exacerbates the noise present within the pseudo-labels, leading to the insufficiency of the pseudo-labeling approach in enhancing the model's performance.

Conclusion

In this paper, we proposed MlanDTI, a semi-supervised domain adaptive multilevel attention network that leverages a large amount of unlabeled data to obtain enriched bidirectional representations of drugs and proteins from a pre-trained BERT model. Additionally, we introduced the multilevel-attention mechanism to capture domain-invariant interaction features between proteins and drugs at different levels and depths. Finally, we incorporated a simple yet effective pseudo labeling method to further enhance our model's generalization ability. Our model demonstrated excellent domain generalization capabilities, making it well-suited for predicting interactions between new drugs and targets in drug development. Through comprehensive comparisons with state-of-the-art models, we establish a substantial performance superiority over prior methodologies.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62172273) and Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102). Shikui Tu and Lei Xu are co-corresponding authors.

References

Abbasi, K.; Razzaghi, P.; Poso, A.; Amanlou, M.; Ghasemi, J. B.; and Masoudi-Nejad, A. 2020. DeepCDA: deep cross-domain compound–protein affinity prediction through LSTM and convolutional neural networks. *Bioinformatics*, 36(17): 4633–4642.

Agamah, F. E.; Mazandu, G. K.; Hassan, R.; Bope, C. D.; Thomford, N. E.; Ghansah, A.; and Chimusa, E. R. 2020. Computational/in silico methods in drug target and lead prediction. *Briefings in bioinformatics*, 21(5): 1663–1675.

Ahmad, W.; Simon, E.; Chithrananda, S.; Grand, G.; and Ramsundar, B. 2022. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*.

Bai, P.; Miljković, F.; John, B.; and Lu, H. 2023. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. *Nature Machine Intelligence*, 5(2): 126–136.

Bakheet, T. M.; and Doig, A. J. 2009. Properties and identification of human protein drug targets. *Bioinformatics*, 25(4): 451–457.

Bian, J.; Zhang, X.; Zhang, X.; Xu, D.; and Wang, G. 2023. MCANet: shared-weight-based MultiheadCrossAttention network for drug–target interaction prediction. *Briefings in Bioinformatics*, 24(2): bbad082.

Broach, J. R.; Thorner, J.; et al. 1996. High-throughput screening for drug discovery. *Nature*, 384(6604): 14–16.

Chen, L.; Cruz, A.; Ramsey, S.; Dickson, C. J.; Duca, J. S.; Hornak, V.; Koes, D. R.; and Kurtzman, T. 2019. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PloS one*, 14(8): e0220113.

Chen, L.; Tan, X.; Wang, D.; Zhong, F.; Liu, X.; Yang, T.; Luo, X.; Chen, K.; Jiang, H.; and 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.

Cheng, F.; Zhou, Y.; Li, J.; Li, W.; Liu, G.; and Tang, Y. 2012. Prediction of chemical–protein interactions: multitarget-QSAR versus computational chemogenomic methods. *Molecular BioSystems*, 8(9): 2373–2384.

Cortes, C.; and Vapnik, V. 1995. Support-vector networks. *Machine learning*, 20: 273–297.

Dauphin, Y. N.; Fan, A.; Auli, M.; and Grangier, D. 2017. Language modeling with gated convolutional networks. In *International conference on machine learning*, 933–941. PMLR.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; et al. 2021. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10): 7112–7127.

Ezzat, A.; Wu, M.; Li, X.-L.; and Kwoh, C.-K. 2019. Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. *Briefings in bioinformatics*, 20(4): 1337–1357.

Faulon, J.-L.; Misra, M.; Martin, S.; Sale, K.; and Sapra, R. 2008. Genome scale enzyme–metabolite and drug–target interaction predictions using the signature molecular descriptor. *Bioinformatics*, 24(2): 225–233.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, 1026–1034.

Ho, T. K. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, 278–282. IEEE.

Huang, K.; Xiao, C.; Glass, L. M.; and Sun, J. 2021. MolTrans: molecular interaction transformer for drug–target interaction prediction. *Bioinformatics*, 37(6): 830–836.

Huang, L.; Lin, J.; Liu, R.; Zheng, Z.; Meng, L.; Chen, X.; Li, X.; and Wong, K.-C. 2022. CoaDTI: multi-modal co-attention based framework for drug–target interaction annotation. *Briefings in Bioinformatics*, 23(6): bbac446.

Huang, W.; Tu, S.; and Xu, L. 2022. Deep CNN based Lmser and strengths of two built-in dualities. *Neural Processing Letters*, 54(5): 3565–3581.

Kao, P.-Y.; Kao, S.-M.; Huang, N.-L.; and Lin, Y.-C. 2021. Toward drug-target interaction prediction via ensemble modeling and transfer learning. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2384–2391. IEEE.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 896. Atlanta.

Lee, I.; Keum, J.; and Nam, H. 2019. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS computational biology*, 15(6): e1007129.

Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; and Gilson, M. K. 2007. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research*, 35(suppl_1): D198–D201.

Meng, F.-R.; You, Z.-H.; Chen, X.; Zhou, Y.; and An, J.-Y. 2017. Prediction of drug–target interaction networks

from the integration of protein sequences and drug chemical structures. *Molecules*, 22(7): 1119.

Nguyen, T.; Le, H.; Quinn, T. P.; Nguyen, T.; Le, T. D.; and Venkatesh, S. 2021. GraphDTA: Predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8): 1140–1147.

Öztürk, H.; Özgür, A.; and Ozkirimli, E. 2018. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17): i821–i829.

Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; and Schacht, A. L. 2010. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature reviews Drug discovery*, 9(3): 203–214.

Qian, H.; Lin, C.; Zhao, D.; Tu, S.; and Xu, L. 2022. AlphaDrug: protein target specific de novo molecular generation. *PNAS nexus*, 1(4): pgac227.

Qian, Y.; Wu, J.; and Zhang, Q. 2022. CAT-CPI: Combining CNN and transformer to learn compound image features for predicting compound-protein interactions. *Frontiers in Molecular Biosciences*, 9: 963912.

Rifaioglu, A. S.; Atas, H.; Martin, M. J.; Cetin-Atalay, R.; Atalay, V.; and Doğan, T. 2019. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Briefings in bioinformatics*, 20(5): 1878–1912.

Rizve, M. N.; Duarte, K.; Rawat, Y. S.; and Shah, M. 2021. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*.

Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1): 61–80.

Shazeer, N.; Lan, Z.; Cheng, Y.; Ding, N.; and Hou, L. 2020. Talking-heads attention. *arXiv preprint arXiv:2003.02436*.

Sieg, J.; Flachsenberg, F.; and Rarey, M. 2019. In need of bias control: evaluating chemical data for machine learning in structure-based virtual screening. *Journal of chemical information and modeling*, 59(3): 947–961.

Stepniewska-Dziubinska, M. M.; Zielenkiewicz, P.; and Siedlecki, P. 2018. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics*, 34(21): 3666–3674.

Tsubaki, M.; Tomii, K.; and Sese, J. 2019. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, 35(2): 309–318.

Wallach, I.; Dzamba, M.; and Heifets, A. 2015. AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv preprint arXiv:1510.02855*.

Wang, X.-r.; Cao, T.-t.; Jia, C. M.; Tian, X.-m.; and Wang, Y. 2021. Quantitative prediction model for affinity of drug–target interactions based on molecular vibrations and overall system of ligand-receptor. *BMC bioinformatics*, 22(1): 1–18.

Wu, F.; Jin, S.; Jiang, Y.; Jin, X.; Tang, B.; Niu, Z.; Liu, X.; Zhang, Q.; Zeng, X.; and Li, S. Z. 2022. Pre-Training of Equivariant Graph Matching Networks with Conformation Flexibility for Drug Binding. *Advanced Science*, 9(33): 2203796.

Xu, L. 1993. Least mean square error reconstruction principle for self-organizing neural-nets. *Neural networks*, 6(5): 627–648.

Xu, L. 2019. An overview and perspectives on bidirectional intelligence: Lmser duality, double IA harmony, and causal computation. *IEEE/CAA Journal of Automatica Sinica*, 6(4): 865–893.

Yazdani-Jahromi, M.; Yousefi, N.; Tayebi, A.; Kolanthai, E.; Neal, C. J.; Seal, S.; and Garibay, O. O. 2022. AttentionSiteDTI: an interpretable graph-based model for drug-target interaction prediction using NLP sentence-level relation classification. *Briefings in Bioinformatics*, 23(4): bbac272.

Zhao, Q.; Zhao, H.; Zheng, K.; and Wang, J. 2022. HyperAttentionDTI: improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism. *Bioinformatics*, 38(3): 655–662.

Zheng, S.; Li, Y.; Chen, S.; Xu, J.; and Yang, Y. 2020. Predicting drug–protein interaction using quasi-visual question answering system. *Nature Machine Intelligence*, 2(2): 134–140.