# Blend the Separated: Mixture of Synergistic Experts for Data-Scarcity Drug-Target Interaction Prediction

**Xinlong Zhai[1], Chunchen Wang[1], Ruijia Wang[2], Jiazheng Kang[1], Shujie Li[1], Boyu Chen[1], Tengfei Ma[3], Zikai Zhou[1], Cheng Yang[1], Chuan Shi[1]*,**

[1]Beijing University of Posts and Telecommunications, [2]China Telecom Cloud Computing Research Institute
[3]College of Computer Science and Electronic Engineering, Hunan University
{zhaijojo, wangchunchen, kjz, shujielie, chenbys4, 2020211668, yangcheng, shichuan}@bupt.edu.cn,
wangrj12@chinatelecom.cn, tfma@hnu.edu.cn

## Abstract

Drug-target interaction prediction (DTI) is essential in various applications including drug discovery and clinical application. There are two perspectives of input data widely used in DTI prediction: Intrinsic data represents how drugs or targets are constructed, and extrinsic data represents how drugs or targets are related to other biological entities. However, any of the two perspectives of input data can be scarce for some drugs or targets, especially for those unpopular or newly discovered. Furthermore, ground-truth labels for specific interaction types can also be scarce. Therefore, we propose the first method to tackle DTI prediction under input data and/or label scarcity. To make our model functional when only one perspective of input data is available, we design two separate experts to process intrinsic and extrinsic data respectively and fuse them adaptively according to different samples. Furthermore, to make the two perspectives complement each other and remedy label scarcity, two experts synergize with each other in a mutually supervised way to exploit the enormous unlabeled data. Extensive experiments on 3 real-world datasets under different extents of input data scarcity and/or label scarcity demonstrate our model outperforms states of the art significantly and steadily, with a maximum improvement of 53.53%. We also test our model without any data scarcity and it also outperforms current methods. The codes are available at https://github.com/BUPT-GAMMA/MoseDTI.

## 1 Introduction

The task of drug-target interaction (DTI) prediction is crucial across various biological fields, particularly within the pharmacology (Lukačišin and Bollenbach 2019; Bredel and Jacoby 2004; Lee et al. 2019; Zhang, Zang, and Zhao 2024). In this task, a drug (molecule) and a target (a gene or the encoded protein of a gene) are input and output is the probability of them interacting.

There has been a surge in the development of diverse neural networks for DTI prediction, which significantly reduces the need for domain knowledge and has demonstrated superior results. Generally, there are two perspectives of data which can be utilized in these methods, which are illustrated in Fig. 1. The first perspective of data is how molecules or proteins are composed, like molecule structures and amino acid residue sequences of proteins. We denote this perspective of data as intrinsic data. DeepDTA (Öztürk, Özgür, and Ozkirimli 2018) uses two separate CNNs to encode the SMILES representation of molecule structures and amino acid residue sequences of proteins respectively. The second perspective of data is the relations between various biological entities, such as diseases, side effects and symptoms, besides the drugs and targets. We denote this perspective of data as extrinsic data. Entities and relations can be formed as graphs so various graph embedding methods can be applied (Su et al. 2024; Wang et al. 2022). It is natural to consider utilizing both perspectives of data to achieve better prediction performance, and more recently, there have emerged a few methods to realise it. MDTips (Xia et al. 2023) uses a ConvE to embed extrinsic data, a GAT and a transformer to embed intrinsic data of drugs and targets, and then concatenate them to predict interaction.

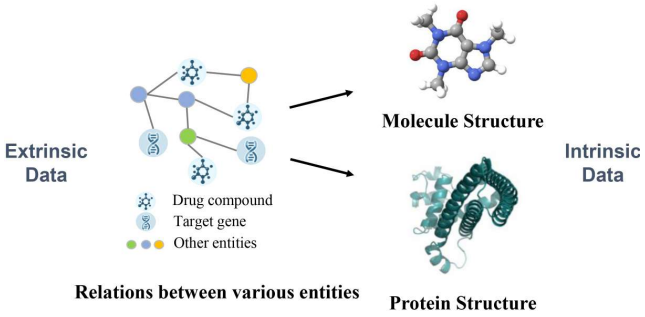



Figure 1: Illustration of intrinsic and extrinsic data.

However, there are two forms of data scarcity that limit the usage of all current works: (1) Intrinsic or extrinsic input data scarcity. With regard to intrinsic data, for example, the acquisition of the most accurate and precise structure of proteins still relies on wet experiments with expensive equipment like cryo-electron microscopes, causing the scarcity of precise protein structures. For extrinsic data, though there has been massive relation data between biology entities accumulated, newly discovered or unpopular drugs or targets could still have few connections with other entities. (2) Interaction label scarcity. The interactions between drugs and targets have diverse specific types. Though there are abun-

*Corresponding author

dant binary labels of whether they interact, the labels for a specific interaction type are still limited. For example, though there are about 210k DTI labels in the DRKG (Ioannidis et al. 2020) dataset, some specific interaction types, like "positive allosteric modulator", could only have dozens of labels, which are insufficient for common deep learning methods.

The main research goal of this work is to propose a method that exploits both intrinsic and extrinsic data effectively, while still functional under input data and/or interaction label scarcity. This requires us to address the following two challenges: (1) *How to fuse intrinsic and extrinsic data flexibly and substantially.* Models with a direct fusion of intrinsic and extrinsic data, like concatenating embeddings from two perspectives, cannot predict when one perspective of data is absent. Furthermore, when predicting without one data perspective, how could we still utilize the knowledge learnt from data of this perspective during training? (2) *How to optimize efficiently with limited interaction labels.* Caused by the huge divergence between different specific relations, we cannot transfer the knowledge learnt in the general interaction to specific interactions to remedy the interaction label scarcity, which is demonstrated experimentally in Appendix B. Moreover, intrinsic data contains composition information of the drugs and targets themselves while extrinsic data contains higher-level semantic information between drugs and targets. Hence it also remains to be explored how to optimize models more label-efficiently by exploiting the complementarity between the two data perspectives.

In this paper, we propose a novel method MoseDTI, i.e., **m**ixture **o**f **s**ynergistic **e**xperts for data-scarcity **d**rug-**t**arget **i**nteraction prediction, which performs well under any or both of these two types of data scarcity. We propose a novel model architecture called the mixture of synergistic experts to address the two challenges unitedly and organically. To address the first challenge, two heterogeneous experts are designed to predict DTI interaction probabilities according to intrinsic and extrinsic data respectively. Then a gating model is applied to adaptively adopt their output according to whether the intrinsic or extrinsic data of a sample is more reliable. The two experts are synergized, i.e., one expert supervises the other during training to inject knowledge from one perspective into the other expert. If intrinsic or extrinsic data is absent when predicting, one of the experts can still predict normally. To address the second challenge, the two experts are designed to generate pseudo labels for each other as the supervision method. The pseudo labels generated effectively enlarge the training samples for the two experts and the gating model, and adequately exploit the complementarity between the two perspectives of data.

Elaborate experiments on three real-world datasets under any or both data scarcity of different extents show that our method outperforms state-of-the-art steadily and significantly, with a maximum improvement of 53.53%. We also test our method on two real-world datasets without data scarcity and it still outperforms other methods, which proves the generality of our method.

## 2 Related work

In this section, we classify all current DTI works into intrinsic methods, extrinsic methods and hybrid methods according to which data perspective they use, which is elaborated in Sec. 1, and roughly review them.

**Intrinsic methods.** Many works utilizing various deep neural networks have achieved excellent performance for drug-target interaction prediction to encoder intrinsic data of drugs and targets (Tsubaki, Tomii, and Sese 2019; Li et al. 2020; Chen et al. 2020, 2023; Nguyen et al. 2021; Öztürk, Özgür, and Ozkirimli 2018; Karimi et al. 2019). An end-to-end deep learning framework named GNN-CPI (Tsubaki, Tomii, and Sese 2019) that applied the GNN to embed the compound represented by molecular graph is an early work. MONN (Li et al. 2020) was proposed to jointly predict both non-covalent interactions and binding affinities between compounds and proteins. TransformerCPI2.0 (Chen et al. 2023) introduces a sequence-to-drug concept, employing end-to-end differentiable learning based on protein sequences. These methods only use the local features of drugs and targets themselves ignoring there are abundant extrinsic data between biology entities and cannot predict a specific interaction type with few-shot labels. From another perspective, they can also be seen as orthogonal to our work because the drug and target encoder in our model can be easily replaced by the encoders presented in these works.

**Extrinsic methods.** Some studies on DTI prediction apply extrinsic data and resolve a link prediction task on a graph or a heterogeneous information network (Mohamed, Nováček, and Nounu 2020; Su et al. 2024; Ezzat et al. 2016; Wan et al. 2019; Peng et al. 2021; Wang et al. 2022; Li et al. 2021). For example, TriModel (Mohamed, Nováček, and Nounu 2020) adopted KG embedding to learn the representations of drugs and targets for DTI prediction. AMGDTI (Su et al. 2024) introduces an adaptive meta-graph learning approach and automates semantic information aggregation from heterogeneous networks for DTI prediction. However, these works ignore the intrinsic data, i.e., the local features of nodes themselves, which is fatal when we need to predict newly discovered or unpopular drugs or targets with few connections to other biological entities.

**Hybrid methods.** We also noticed that there are attempts to utilize both intrinsic and extrinsic data for better DTI prediction performance (Zhou et al. 2021; Ma et al. 2022; Xia et al. 2023; Li et al. 2023; Dong et al. 2023). KG-MTL (Ma et al. 2022) tries to merge knowledge graph and molecule graph via multi-task learning, which employs a shared unit to jointly maintain drug entity semantics and compound structural relations in both graphs. MDTips (Xia et al. 2023) predicts DTI using multi-modal data, integrating knowledge graphs, gene expression profiles, and structural details of drugs and targets. However, They are not designed to handle DTI prediction with limited labels and are hard to predict when extrinsic or intrinsic data is absent for some samples.

## 3 Preliminaries

**Extrinsic data.** We consider extrinsic data as a knowledge graph (KG) as $\mathcal{KG} = (E, R, O)$ that provides abundant re-

lation information between different kinds of biological entities, where E is a set of entities, R is a set of relations, and O is a set of observed $(\mathsf{h}, \mathsf{r}, \mathsf{t})$ triples. In a triple, $\mathsf{h}, \mathsf{r}, \mathsf{t}$ represents the head entity, relation, and tail entity respectively. The entity set $E$ contains various biological entities such as diseases, side-effects and symptoms, and the drug and target sets are subsets of the entity set: $D, T \subset E$. To prevent label data leakage, we remove all direct connections of drugs and targets from $\mathcal{KG}$, i.e. remove all $(\mathsf{h}, \mathsf{r}, \mathsf{t})$ from $O$ which satisfies $\mathsf{h} \in D, \mathsf{t} \in T$ or $\mathsf{h} \in T, \mathsf{t} \in D$.

**Intrinsic Data.** For a drug $d_i \in D$, we use the SMILES sequence $SM_{d_i}$, i.e., *simplified molecular-input line-entry system* sequence as its intrinsic feature, which is a specification in the form of a line notation for describing the structure of chemical substance using short ASCII strings. For a target gene $t_i \in T$, we use the UniProt database to obtain the amino acid sequence of the protein it encodes as its intrinsic feature, denoted as $AS_{t_i}$.

**DTI Task.** In the drug-target interaction task, we aim to estimate the interaction probability $p_{ij}$ of a drug-target pair $(d_i, t_j)$ under a specific interaction type, where $d_i \in D, t_j \in T$. Such a DTI dataset can be described as $(X^p, X^n, \mathcal{KG}, SM_D, AS_T)$, where $X^p$ or $X^n$ is $\{(d_i, t_j)\}$ which indicates these drug-target pairs have or do not have this type of interaction, and $SM_D$ and $AS_T$ denote intrinsic data of all drugs and targets respectively.

# 4 Methodology

In this section, we describe the proposed MoseDTI model for drug-target interaction prediction with data scarcity specifically, as illustrated in Fig. 2. We first introduce the model structure of two heterogeneous experts and a gating model, and then elaborate on how we optimize them.

## Model Architecture

Our model consists of three components: an extrinsic expert, an intrinsic expert and a gating model. The two experts take extrinsic and intrinsic data as input respectively and output interaction probabilities. The gating model takes intrinsic and extrinsic representations of both the drug and target and output weight to determine whether the intrinsic or extrinsic expert is more reliable for the current sample. This architecture utilizes the extrinsic and intrinsic data adaptively according to specific samples, and the two experts can work alone if one perspective of data is absent when predicting.

**Extrinsic Expert** The extrinsic expert is to predict DTI based on relation data between biological entities. We first use the massive unlabeled association data between biological entities to pretrain embeddings of drugs and targets, and then train a classifier with labels to output the interaction probabilities from the extrinsic perspective.

**Knowledge graph embedding.** The knowledge graph $\mathcal{KG}$ contains various association data between different biological entities, in which drugs and targets are connected to other types of entities, like diseases, side effects, and symptoms. To leverage the abundant semantic information it implies, we first use the KG embedding method to pretrain the $d$-dimensional drug extrinsic embedding $\mathbf{h}_{d_i}^{ex} \in \mathcal{R}^d$ for drug $d_i$ and target extrinsic embedding $\mathbf{h}_{t_j}^{ex} \in \mathcal{R}^d$ for target $t_j$. We do not introduce any labelled drug-target interaction data into the pretrain, so the embeddings can be used for different specific DTI datasets without retraining.

**Extrinsic classifier.** After that, given a specific interaction dataset and its ground-truth samples $(X^p, X^n)$, there is an extrinsic classifier $g_r$ to predict the interaction probability for $(d_i, t_j)$:

$$p_{ij}^{ex} = g^{ex}(\mathbf{h}_{d_i}^{ex}, \mathbf{h}_{t_j}^{ex}) \tag{1}$$

In the experiment, we implement the $g_r$ as a simple MLP because a simpler $g_r$ with fewer parameters can be more easily trained by limited samples.

**Intrinsic Expert** The intrinsic expert is to predict DTI based on the structure data of drugs and targets. We use a drug encoder and a target encoder to encode drug SMILES sequence and target amino acid residue sequences respectively. Then, an intrinsic classifier is applied to output the interaction probability intrinsically.

**Drug encoder.** For a drug $d_i$, its SMILES sequence $SM_{d_i}$ is first translated to a molecule graph $\mathcal{MG}_{d_i}$ with RDKit (Landrum 2006). $\mathcal{MG}_{d_i} = (\mathcal{V}_{d_i}, \mathcal{E}_{d_i})$, where $\mathcal{V}_{d_i}$ denotes the set of nodes, i,e., atoms and $\mathcal{E}_{d_i}$ is the set of edges between atoms, i.e., chemical bonds. A node $v \in \mathcal{V}_{d_i}$ has its embedding initialized as $\mathbf{h}_v^{(0)}$ with the method proposed in (Quan et al. 2019). We utilize a graph neural network (GNN) to obtain the final embedding of each node (Gilmer et al. 2017; Quan et al. 2019):

$$\mathbf{m}_{u\to v}^{(l)} = Message^{(l)}\left(\mathbf{h}_v^{(l-1)}, \mathbf{h}_v^{(l-1)}\right) \tag{2}$$

$$\mathbf{m}_v^{(l)} = Reduce_{u\in\mathcal{N}(v)}\mathbf{m}_{u\to v}^{(l)} \tag{3}$$

$$\mathbf{h}_v^{(l)} = Update^{(l)}\left(\mathbf{h}_v^{(l-1)}, \mathbf{m}_v^{(l)}\right), \tag{4}$$

where $Message$, $Reduce$ and $Update$ are three functions specified by the selected GNN, like GCN (Kipf and Welling 2017) or GAT (Velickovic et al. 2017). The superscript $(l), l = 1, 2, ..., L$ indicates a certain GNN layer, and the $\mathcal{N}(v)$ denotes the nodes connected to $v$ by edges. $\mathbf{m}_{u\to v}^{(l)}$ and $\mathbf{m}_v^{(l)}$ indicate the message from $u$ to $v$ and the overall message $v$ received at layer $(l)$ respectively. Then, the embedding of the whole molecule graph $\mathbf{h}_{d_i}^{in}$ is calculated by a multi-layer perceptron (MLP) from the $L-th$ layer embeddings and a max readout function:

$$\mathbf{h}_{d_i}^{in} = max(\{MLP(\mathbf{h}_v^{(L)}) | v \in V\}). \tag{5}$$

**Target encoder.** For a target $t_j$, its amino acid residues sequence $AS_{t_j}$ is first input to a pre-trained protein language model ESM-MSA-1b (Rives et al. 2021) and embeddings $\{\mathbf{e}_m^{(0)} | m = 1, ..., M\}$for every amino acid residue are output, where $M$ is the length of the sequence. Then, a $K$ layer 1-dimensional CNN with adaptive max pooling is applied to obtain the final embedding for the sequence (Ma et al. 2022):

$$\mathbf{h}_{t_j}^{in} = AMP(Conv(\{\mathbf{e}_j^{(0)} | k = 1, ..., K\})), \tag{6}$$

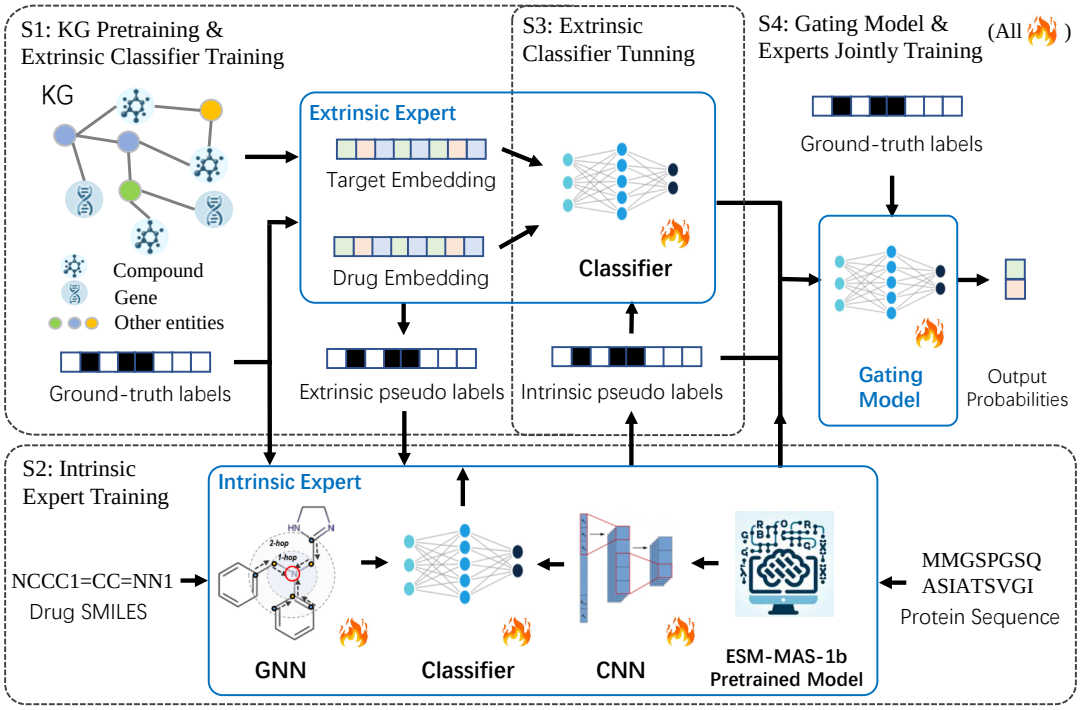



Figure 2: The framework of our MoseDTI. The three components are surrounded in blue rectangles and the first three training steps S1 to S3 are surrounded in dotted black rectangles. For the last step S4, all components with little flames are jointly trained.

where $Conv = Conv^{(1)} \circ \cdots \circ Conv^{(K)}$ and each convolution layer contains a 1-dimensional convolution operation and a ReLU activation function, and $AMP$ represents adaptive max pooling.

**Intrinsic classifier.** Subsequently, given a specific interaction dataset, an intrinsic classifier $g^{in}$ is trained to predict the interaction probability:

$$p_{ij}^{in} = g^{in}(\mathbf{h}_{d_i}^{in}, \mathbf{h}_{t_j}^{in}). \tag{7}$$

Experimentally, like the extrinsic classifier $g^{ex}$, we also implement $g^{in}$ as a MLP.

**Gating model** To exploit both extrinsic and intrinsic adaptively, we fuse the output of two experts according to specific samples. We design a gating model (Jacobs et al. 1991) that accepts the hidden embeddings of the two experts to generate a weight $w_{ij}$:

$$w_{ij} = Gating(\mathbf{h}_{d_i}^{ex}, \mathbf{h}_{t_j}^{ex}, \mathbf{h}_{d_i}^{in}, \mathbf{h}_{t_j}^{in}). \tag{8}$$

We implement $Gating$ as an MLP and a softmax function. Then we use $w_{ij}$ to blend the two models as the final output of our entire model:

$$p_{ij} = w_{ij} p_{ij}^{ex} + (1 - w_{ij}) p_{ij}^{in}. \tag{9}$$

## Optimization

In this subsection, we first introduce how experts are synergized and then elaborate on the entire training procedures.

**Synergizing experts** Inspired by the self-training optimization strategy, we design a novel expert synergizing mechanism, in which experts generate pseudo labels from massive unlabeled samples for each other. Compared to the self-training, the exchange of pseudo labels could bring in high-confidence samples from another perspective, preventing the model from over-fitting to the samples similar to the offered ground-truth samples. Here we elaborate on how one expert called $ExpertA$ generates pseudo labels with two-stage sampling to train another expert $ExpertB$.

First, we sample a portion of the cartesian set of the drug set $D$ and the target set $T$ as a candidate set:

$$Cand_A = Sample(D \times T, \lfloor \alpha_A L_{ca} \rfloor), \tag{10}$$

where $\times$ denotes the cartesian product of two sets, $L_{ca} = |D||T|$ is length of the cartesian product set and $\alpha_A$ is the sampling rate . The length of $Cand_A$ is denoted as $L_{Cand_A} = \lfloor \alpha_A L_{ca} \rfloor$. Next, we use expert A to predict on $Cand_A$:

$$P_{Cand_A} = ExpertA(Cand_A), \tag{11}$$

where $P_{Cand_A} = \{p_s^A | s = 1, ..., L_{Cand_A}\}$ and $p_s^A$ is the probability of interaction predicted by $ExpertA$ for drug-target pair $s$. Then we sort $Cand_A$ according to $P_{Cand_A}$ and select top samples in $Cand_A$ as pseudo positive samples, denoted as $X_A^p$:

$$X_A^p = Top(Cand_A, P_{Cand_A}, \lfloor \beta_A L_{Cand_A} \rfloor), \tag{12}$$

where $X_A^p = \{(d_i, t_j)\}$, its length $L_{X_A^p} = \lfloor \beta_A L_{Cand_A} \rfloor$ and $\beta_A$ is a choosing rate. Denoting the ground-truth positive

and negative samples as $X^p$ and $X^n$, the expert B is trained with loss:

$$\mathcal{L}_B = -\left(\sum_{(d_i,t_j)\in\gamma_A X^p \cup X_A^p} \log(p_{ij}^B) + \sum_{(d_i,t_j)\in X^n \cup X_A^n} \log(1-p_{ij}^B)\right) \quad (13)$$

, where $X_A^n$ is pseudo negative samples and $\gamma_A$ is an integer to amplify the weight of true positive samples versus pseudo positive samples. Considering that most samples in $Cand_A$ are actually negative samples, to improve the diversity of negative samples, we select the bottom samples from $Cand_A$ according to $P_{Cand_A}$ with a larger length $|X_A^n|$ which satisfies

$$|X_A^n| = \gamma_A |X^p| + |X_A^p| - |X^n| \quad (14)$$

to keep the label balanced.

**Training procedures** Due to the scarcity of interaction labels, in the first step S1 in Fig. 2, we first pretrain the KG embedding with classical methods like TransE (Bordes et al. 2013) or RotatE (Sun et al. 2019), exploiting all the observed triples in the knowledge graph. Then we only train the extrinsic classifier in Equ.1 with relatively much fewer parameters above the frozen pre-trained embeddings, using the ground-truth labels:

$$\mathcal{L}_{s1} = -\left(\sum_{(d_i,t_j)\in X^p} \log(p_{ij}^{ex}) + \sum_{(d_i,t_j)\in X^n} \log(1-p_{ij}^{ex})\right). \quad (15)$$

In the second step S2, the extrinsic expert is the $ExpertA$, the intrinsic expert is the $ExpertB$ and the intrinsic expert is trained with Equ.13. In the third step S3, the two models exchange their positions, and the extrinsic expert is tuned with Equ. 13. In the last step S4, we jointly train the gating model, intrinsic, and extrinsic expert with ground-truth and pseudo labels from two experts:

$$\mathcal{L}_g = -\left(\sum_{(d_i,t_j)\in\gamma_g X^p \cup X_A^{pos'} \cup X_B^{pos'}} \log(p_{ij}) + \sum_{(d_i,t_j)\in X^n \cup X_A^{neg'} \cup X_B^{neg'}} \log(1-p_{ij})\right). \quad (16)$$

Similar to Equ.12, the positive pseudo samples $X_A^{pos'}$ and $X_B^{pos'}$ are also selected from the top of $Cand_A$ and $Cand_B$ with a shared rate $\beta_g$ and tailing samples of the longer one are trimmed to keep their length equal. $X_A^{neg'}$ and $X_B^{neg'}$ are also selected from the bottom of $Cand_A$ and $Cand_B$ according to $P_{Cand_A}$ and $P_{Cand_B}$ respectively, with their lengths

$$|X_A^{neg'}| = |X_B^{neg'}| = (\gamma_g |X^p| + |X_A^p| + |X_B^p| - |X^n|)/2, \quad (17)$$

to balance the total positive and negative samples.

# 5 Experiments

In this section, we first introduce on the datasets and baselines and then show model results. The goal of our experiments is to answer the following research questions (RQs).

1. Can MoseDTI effectively confront input data scarcity (including intrinsic or extrinsic data scarcity) and/or interaction label scarcity? (**RQ1**)
2. If there is no data scarcity, can MoseDTI still perform well? (**RQ2**)
3. Are the MOE architecture and the synergizing mechanism beneficial? (**RQ3**)

We also conduct experiments of hyper-parameters, case studies and few-shot learning on general interaction datasets. Please see Appendix F, G and H respectively.

## Experimental Setup

**Datasets** We conduct experiments on 5 datasets. There are 3 few-shot datasets of specific interactions including DGIDB::BLOCKER (blocker), DGIDB::AGONIST (agonist) (Griffith et al. 2013) and GNBR::E- (e-) (Percha and Altman 2018). Each dataset presents a specific interaction type and only contains 10 positive ground-truth samples for training. There are also two normal datasets of general DTI interaction including DrugBank (Wishart et al. 2018) and DrugCentral (Ursu et al. 2016), which contain 18480 and 18066 samples respectively and are partitioned into train, valid and test with a ratio of 6:2:2. All of the datasets use the DRKG (Ioannidis et al. 2020) as the common extrinsic data and all connections between drugs and targets are removed to prevent data leakage. The SMILES of drugs are also from DrugBank (Ursu et al. 2016). We collect the amino acid residue sequences of proteins coded by targets from UniProt[1]. More details of datasets and the evaluation protocol are in Appendix C and D.

**Baselines** We use 8 baselines and classify them into intrinsic methods, extrinsic methods and hybrid methods according to whether they only use the intrinsic or extrinsic data, or use both perspectives of data, which is elaborated in Sec. 1. Intrinsic methods include GNNCPI (Tsubaki, Tomii, and Sese 2019), TransformerCPI (Chen et al. 2020) and TransformerCPI2.0 (Chen et al. 2023). Extrinsic methods include TransE (Bordes et al. 2013), RotatE (Sun et al. 2019), TriModel (Mohamed, Nováček, and Nounu 2020), AMGDTI (Su et al. 2024). Hybrid methods include KG-MTL (Ma et al. 2022) and MDTips (Xia et al. 2023). For the implementation details of our model and baselines, see Appendix E.

## Performance under data scarcity (RQ1)

To validate that MoseDTI is effective under different scenarios of data scarcity, we conduct exhaustive experiments with the following two orthogonal scarcity settings: (1) For intrinsic or extrinsic data scarcity, there are 3 different settings of data availability when inference: only intrinsic data

[1]https://www.uniprot.org/

Table 1: Model performance on three 10-shot datasets of specific interaction comparing three variants of our model including MoseDTI, Mose-intr and Mose-extr with nine baselines. The best performance is boldfaced.

| | | DGIDB::AGONIST | | | DGIDB::BLOCKER | | | GNBR::E- | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | ACC | AUC | AUPR | ACC | AUC | AUPR | ACC | AUC | AUPR |
| Intrinsic Method | GNNCPI | 51.10±1.03 | 61.22±2.77 | 61.89±1.85 | 57.63±3.63 | 70.23±5.73 | 73.51±5.49 | 54.39±3.29 | 64.71±3.84 | 65.81±3.44 |
| | TFCPI | 58.55±3.42 | 66.11±6.33 | 66.37±4.67 | 56.27±5.07 | 68.15±1.72 | 63.94±3.05 | 63.10±3.20 | 78.14±0.37 | 78.06±0.45 |
| | TFCPI2.0 | 50.79±0.10 | 53.05±0.11 | 50.77±0.12 | 36.66±0.25 | 33.53±0.42 | 38.85±0.16 | 42.07±0.07 | 40.49±0.12 | 41.90±0.06 |
| | Mose-intr | 64.50±6.31 | 73.70±6.40 | 72.47±6.91 | 88.48±4.64 | 93.17±3.64 | 94.34±2.34 | 70.64±4.32 | 80.43±3.75 | 78.83±4.32 |
| Extrinsic Method | TransE | 50.00±0.00 | 50.22±0.08 | 50.65±0.08 | 50.00±0.00 | 53.96±0.23 | 55.22±0.48 | 50.00±0.00 | 48.59±0.04 | 48.62±0.06 |
| | RotatE | 50.00±0.00 | 50.41±0.07 | 50.49±0.05 | 50.00±0.00 | 51.24±0.40 | 50.98±0.52 | 50.00±0.00 | 48.84±0.07 | 49.91±0.08 |
| | TriModel | 50.19±0.15 | 48.34±1.23 | 50.21±1.29 | 50.00±0.00 | 33.85±5.38 | 42.09±2.57 | 50.01±0.04 | 49.59±1.81 | 50.08±0.98 |
| | AMGDTI | 57.91±3.72 | 66.48±2.87 | 69.42±6.84 | 80.89±2.42 | 96.14±2.35 | 96.08±0.36 | 61.49±8.87 | 63.84±1.24 | 64.28±1.69 |
| | Mose-extr | 64.88±3.54 | 75.55±3.94 | 75.41±4.48 | 80.28±2.18 | 98.11±0.29 | 97.14±0.31 | 70.87±4.08 | 88.03±4.52 | **87.24±4.74** |
| Hybrid Method | KG-MTL | 56.76±1.02 | 56.06±2.71 | 50.22±0.89 | 65.83±5.69 | 73.57±4.09 | 76.07±3.16 | 54.06±0.71 | 57.65±1.08 | 57.61±1.16 |
| | MDTips | 64.23±4.13 | 73.51±2.91 | 72.37±2.87 | 91.25±2.34 | 97.27±0.86 | 96.90±0.79 | 69.61±3.43 | 82.35±3.89 | 80.78±4.86 |
| | MoseDTI | **67.27±1.90** | **75.82±2.90** | **75.57±3.62** | **92.85±2.79** | **98.76±0.82** | **97.21±0.64** | **78.17±4.86** | **88.71±4.34** | 86.62±5.31 |

is available; only extrinsic data is available; both perspectives of data are available. We take the intrinsic expert in our method as Mose-intr, the extrinsic expert as Mose-extr, after the entire training procedure, which can predict when only one data perspective is available. (2) For interaction label scarcity, there are also three scarcity extents regarding to the number of labelled positive samples for training, i.e., 10-shots, 20-shots and 40-shots.

we compare MoseDTI and its variants with state-of-the-art methods on three real-world datasets under totally 9 ($3 \times 3$) differnet scarcity settings. The results of 10-shots are shown in Tab. 1. The results of 20 and 40 shots are shown in Appendix I. Generally, our method significantly outperforms other methods.

The first four methods can be applied when intrinsic data is absent and only extrinsic data is available for prediction. GNN-CPI, which applies a GNN to encode molecular graphs of compounds and a CNN to obtain chemical features of proteins, demonstrates its stable but limited ability to confront few-shot settings. TransformerCPI2.0 (TFCPI2.0) claims its excellent performance for generalizing to new compounds and proteins, while it fails to predict specific interaction types, probably owing to the substantial difference between specific and general DTI interaction. Our Mose-intr makes use of a pre-trained protein language model and massive unlabeled drug-target pairs when training, outperforming them steadily by a large margin.

The following five methods can be applied when extrinsic data is absent and only extrinsic data is available for prediction. The three KG-embedding-based methods, i.e., TransE, RotatE and TriModel all fail to perform well with limited labelled interactions. It is conceivable that the few-shot labels are not enough for them to optimize their free embeddings of entities and relations associated with a certain interaction type. AMGDTI automatically aggregates semantic information from KG by training an adaptive meta-path and performs pretty well in the blocker dataset. However, our Mose-extr, which also accommodates label scarcity via the designing of the pre-trained entity embedding plus a simple MLP layer and trained with unlabeled potential interaction pairs, performs generally better on the three datasets

The last three methods can be applied when both extrinsic and intrinsic data are available for prediction. KG-MTL fails to perform well, probably because the complex model architecture for handling two related tasks together needs sufficient labelled samples. MDTips fuses the embeddings from KG embeddings, drug embeddings and target embeddings and performs remarkably across three datasets. Thanks to the MOE architecture and the usage of unlabeled samples, our MoseDTI outperforms it obviously and steadily.

## Performance without data scarcity (RQ2)

We also test the performance of our model compared with all the baselines above in two datasets of normal label scale for general DTI prediction task. General DTI prediction task regards all kinds of specific interactions as a whole and does not distinguish them, and hence there is abundant labelled interaction data accumulated.

We can observe that GNNCPI, AMGDTI, KG-MTL and TFCPI all perform well, which proves general DTI prediction with plentiful labels is a relatively simple task compared to predicting a specific interaction with a limited amount of labels. Furthermore, taking advantage of blending two experts adaptively according to the relative importance of intrinsic and extrinsic data of each sample, our MoseDTI even performs better than all other methods.

## Ablation Study (RQ3)

To investigate how our synergizing mechanism and the design of MOE improve the performance for DTI prediction, we conduct the ablation study with the following variants on few-shot specific DTI prediction: 1) True-intr: training the intrinsic model with only the ground-truth labels. 2) True-extr: training the extrinsic expert with only the ground-truth labels. 3) True-all: training the entire model with only the ground-truth labels. For convenience of comparison, we also

Table 2: Model performance on general DTI datasets. The best performance is boldfaced.

| | | DrugCentral | | | DrugBank | | |
|---|---|---|---|---|---|---|---|
| | | ACC | AUC | AUPR | ACC | AUC | AUPR |
| Intrinsic Method | GNNCPI | 72.64 ± 0.51 | 78.44 ± 0.12 | 80.20 ± 0.18 | 73.82 ± 0.14 | 80.65 ± 0.11 | 81.99 ± 0.10 |
| | TFCPI | 80.97 ± 1.34 | 88.94 ± 0.46 | 88.67 ± 0.48 | 81.50 ± 0.53 | 90.69 ± 0.46 | 90.28 ± 0.39 |
| | TFCPI2.0 | 57.58 | 58.67 | 60.60 | 54.51 | 57.19 | 60.93 |
| Extrinsic Method | TransE | 55.89 ± 0.59 | 74.83 ± 1.02 | 75.11 ± 1.02 | 57.86 ± 0.42 | 74.44 ± 0.23 | 76.62 ± 0.46 |
| | RotatE | 54.34 ± 0.18 | 63.09 ± 0.74 | 60.22 ± 1.30 | 60.62 ± 9.72 | 68.17 ± 11.99 | 68.66 ± 12.69 |
| | TriModel | 52.89 ± 0.40 | 63.41 ± 0.60 | 61.41 ± 0.70 | 54.25 ± 0.28 | 65.08 ± 0.49 | 64.35 ± 0.87 |
| | AMGDTI | 80.70 ± 3.32 | 89.27 ± 1.84 | 89.48 ± 2.70 | 83.80 ± 0.45 | 90.03 ± 0.62 | 92.57 ± 0.54 |
| Hybrid Method | KG-MTL | 81.60 ± 0.71 | 88.69 ± 0.30 | 88.98 ± 0.73 | 80.31 ± 0.46 | 87.50 ± 0.54 | 89.71 ± 0.24 |
| | MDTips | 88.10 ± 0.30 | 94.62 ± 0.12 | **95.32 ± 0.21** | 87.75 ± 0.58 | 94.38 ± 0.15 | 94.02 ± 0.23 |
| | MoseDTI | **88.40 ± 0.49** | **95.11 ± 0.19** | **95.32 ± 0.22** | **88.15 ± 0.78** | **94.90 ± 0.14** | **95.23 ± 0.28** |

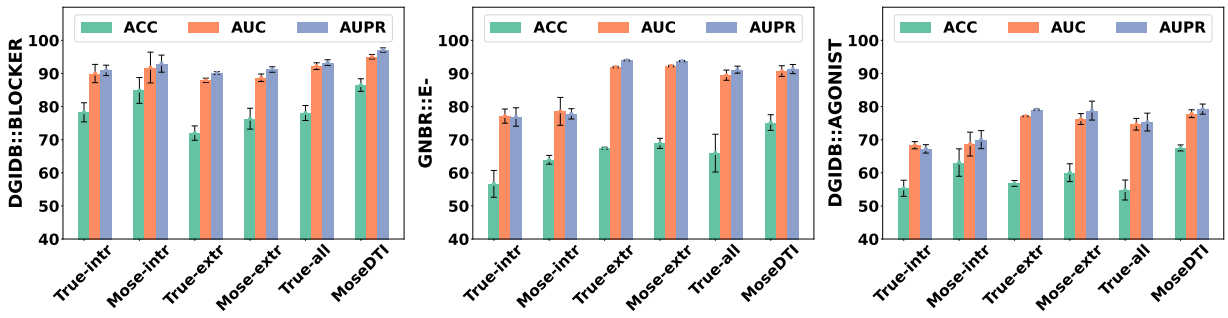



Figure 3: Ablation study on three real-world datasets. The standard deviations are shown in small black lines on top of each bar.

add the results of Mose-intr, Mose-extr and MoseDTI to Fig.3. The difference between True-intr and Mose-intr is whether trained with the synergizing mechanism, and the same with True-extr and Mose-extr. We have the following observations:

**Effect of synergizing mechanism** The effect of the synergizing mechanism to augment one expert is affected by the performance of the other. For example, on the agonist dataset, the performance of Mose-extr is not obviously better than True-extr, while the improvement is more obvious on the blocker dataset. It is more than likely due to the intrinsic data on the blocker dataset is more easily to be utilized for the intrinsic model to produce pseudo labels with higher quality according to the better performance of True-intr on the blocker dataset compared with the agonist dataset. We also find our synergizing mechanism is robust enough to avoid a negative impact, even if there is a huge gap between the performance of True-intr and True-extr, like on the e- dataset. Furthermore, the performance promotion is even larger when we compare the whole model True-all and MoseDTI on all three datasets, thanks to the synergizing mechanism also providing additional information for the joint training of the whole model.

**Effect of mixture of experts** The better performance of MoseDTI versus Mose-intr and Mose-extr on all three datasets demonstrates the design of the mixture of experts can enhance both experts while keeping their independence. However, direct training the whole model with few-shot true labels may not surely promote the performance, according to the comparison of True-all against True-intr and True-extr on the e- and agonist datasets.

## 6 Conclusion

This work confronts the problem of effective DTI prediction under input data scarcity (including intrinsic or extrinsic data scarcity) and/or interaction label scarcity. We propose a model architecture: the mixture of synergized experts, which utilizes two synergized heterogeneous experts to process different perspectives of data, which supervise each other mutually with pseudo labels generated from unlabelled samples. The framework solves both forms of data scarcity organically and exhaustive experiments under various data scarcity settings prove its superiority over states of the art.

**Limitations and Broader Impact.** Despite the encouraging results, there could be more modalities to be incorporated to promote drug-target interaction prediction, such as textual data describing biological entities. Our future work will address these limitations. This research may inspire the AI4Science community to pay more attention to the separation and synergizing of different components when designing their models and make them robust with data scarcity in real-world applications.

## Acknowledgments

This work is supported by the National Key Research and Development Program of China (2023YFF0725103), the National Natural Science Foundation of China (U22B2038, 62192784), Young Elite Scientists Sponsorship Program (No.2023QNRC001) by CAST and the China Scholarship Council (CSC).

## References

Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Bredel, M.; and Jacoby, E. 2004. Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nature Reviews Genetics*, 5(4): 262–275.

Chen, L.; Fan, Z.; Chang, J.; Yang, R.; Hou, H.; Guo, H.; Zhang, Y.; Yang, T.; Zhou, C.; Sui, Q.; et al. 2023. Sequence-based drug design as a concept in computational drug design. *Nature Communications*, 14(1): 4217.

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.

Dong, W.; Yang, Q.; Wang, J.; Xu, L.; Li, X.; Luo, G.; and Gao, X. 2023. Multi-modality attribute learning-based method for drug–protein interaction prediction based on deep neural network. *Briefings in bioinformatics*, 24(3): bbad161.

Ezzat, A.; Zhao, P.; Wu, M.; Li, X.-L.; and Kwoh, C.-K. 2016. Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM transactions on computational biology and bioinformatics*, 14(3): 646–656.

Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural message passing for quantum chemistry. In *International conference on machine learning*, 1263–1272. PMLR.

Griffith, M.; Griffith, O. L.; Coffman, A. C.; Weible, J. V.; McMichael, J. F.; Spies, N. C.; Koval, J.; Das, I.; Callaway, M. B.; Eldred, J. M.; et al. 2013. DGIdb: mining the druggable genome. *Nature methods*, 10(12): 1209–1210.

Ioannidis, V. N.; Song, X.; Manchanda, S.; Li, M.; Pan, X.; Zheng, D.; Ning, X.; Zeng, X.; and Karypis, G. 2020. DRKG - Drug Repurposing Knowledge Graph for Covid-19. https://github.com/gnn4dr/DRKG/.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1): 79–87.

Karimi, M.; Wu, D.; Wang, Z.; and Shen, Y. 2019. DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics*, 35(18): 3329–3338.

Kipf, T.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. *ArXiv*, abs/1609.02907.

Landrum, G. 2006. RDKit: Open-source cheminformatics.

Lee, Y.-s.; Krishnan, A.; Oughtred, R.; Rust, J.; Chang, C. S.; Ryu, J.; Kristensen, V. N.; Dolinski, K.; Theesfeld, C. L.; and Troyanskaya, O. G. 2019. A computational framework for genome-wide characterization of the human disease landscape. *Cell systems*, 8(2): 152–162.

Li, J.; Wang, J.; Lv, H.; Zhang, Z.; and Wang, Z. 2021. IMCHGAN: inductive matrix completion with heterogeneous graph attention networks for drug-target interactions prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(2): 655–665.

Li, S.; Wan, F.; Shu, H.; Jiang, T.; Zhao, D.; and Zeng, J. 2020. MONN: A Multi-objective Neural Network for Predicting Pairwise Non-covalent Interactions and Binding Affinities Between Compounds and Proteins. In *Research in Computational Molecular Biology*, 259–260.

Li, X.; Yang, Q.; Luo, G.; Xu, L.; Dong, W.; Wang, W.; Dong, S.; Wang, K.; Xuan, P.; and Gao, X. 2023. SAGDTI: self-attention and graph neural network with multiple information representations for the prediction of drug–target interactions. *Bioinformatics Advances*, 3(1): vbad116.

Lukačišin, M.; and Bollenbach, T. 2019. Emergent gene expression responses to drug combinations predict higher-order drug interactions. *Cell Systems*, 9(5): 423–433.

Ma, T.; Lin, X.; Song, B.; Philip, S. Y.; and Zeng, X. 2022. Kg-mtl: knowledge graph enhanced multi-task learning for molecular interaction. *IEEE Transactions on Knowledge and Data Engineering*.

Mohamed, S. K.; Nováček, V.; and Nounu, A. 2020. Discovering protein drug targets using knowledge graph embeddings. *Bioinformatics*, 36(2): 603–610.

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.

Peng, J.; Wang, Y.; Guan, J.; Li, J.; Han, R.; Hao, J.; Wei, Z.; and Shang, X. 2021. An end-to-end heterogeneous graph representation learning-based framework for drug–target interaction prediction. *Briefings in bioinformatics*, 22(5): bbaa430.

Percha, B.; and Altman, R. B. 2018. A global network of biomedical relationships derived from text. *Bioinformatics*, 34(15): 2614–2624.

Quan, Z.; Guo, Y.; Lin, X.; Wang, Z.-J.; and Zeng, X. 2019. GraphCPI: Graph neural representation learning for compound-protein interaction. In *IEEE International Conference on Bioinformatics and Biomedicine*, 717–722.

Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15): e2016239118.

Su, Y.; Hu, Z.; Wang, F.; Bin, Y.; Zheng, C.; Li, H.; Chen, H.; and Zeng, X. 2024. AMGDTI: drug–target interaction prediction based on adaptive meta-graph learning in heterogeneous network. *Briefings in Bioinformatics*, 25(1): bbad474.

Sun, Z.; Deng, Z.-H.; Nie, J.-Y.; and Tang, J. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197*.

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.

Ursu, O.; Holmes, J.; Knockel, J.; Bologa, C. G.; Yang, J. J.; Mathias, S. L.; Nelson, S. J.; and Oprea, T. I. 2016. DrugCentral: online drug compendium. *Nucleic Acids Research*, 45(D1): D932–D939.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y.; et al. 2017. Graph attention networks. *stat*, 1050(20): 10–48550.

Wan, F.; Hong, L.; Xiao, A.; Jiang, T.; and Zeng, J. 2019. NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug-target interactions. *Bioinformatics*, 35(1): 104–111.

Wang, H.; Huang, F.; Xiong, Z.; and Zhang, W. 2022. A heterogeneous network-based method with attentive meta-path extraction for predicting drug–target interactions. *Briefings in Bioinformatics*, 23(4): bbac184.

Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46: D1074–D1082.

Xia, X.; Zhu, C.; Zhong, F.; and Liu, L. 2023. MDTips: a multimodal-data-based drug–target interaction prediction system fusing knowledge, gene expression profile, and structural data. *Bioinformatics*, 39(7): btad411.

Zhang, C.; Zang, T.; and Zhao, T. 2024. KGE-UNIT: toward the unification of molecular interactions prediction based on knowledge graph and multi-task learning on drug discovery. *Briefings Bioinform.*, 25(2).

Zhou, D.; Xu, Z.; Li, W.; Xie, X.; and Peng, S. 2021. MultiDTI: drug–target interaction prediction based on multimodal representation learning to bridge the gap between new chemical entities and known heterogeneous network. *Bioinformatics*, 37(23): 4485–4492.

# A Details of Notations

We elaborate the key notations we use in Tab. 1.

# B Preliminary Experiments of Transfering from General to Specific Interaction Datasets

To investigate whether it is beneficial to directly transfer from the general interaction datasets to specific interaction datasets, we pre-train the MDTips model (Xia et al. 2023) using the Drugbank (Wishart et al. 2018) general interaction dataset and fine-tune it on the three specific interaction datasets of 10-shots. The results are shown in Table 2, compared with results of direct training on specific datasets.

The results demonstrate that direct transferring cannot bring in a positive effect, probably because the data patterns of general interactions have a significant variance with that of specific interactions. Hence, we choose to design a method which can be effectively trained with limited specific interaction labels.

# C Dataset Details

The whole data of all our experiments can be divided into two parts: input data (i.e., intrinsic and extrinsic data) and interaction label data. Every dataset shares common input data. The extrinsic input data is the knowledge graph DRKG (Ioannidis et al. 2020) with 97238 entities and 11326934 triples. The intrinsic input data is the 11439 SMILES of drugs and 57063 amino acid sequences of targets we collect from Uniprot and Drugbank(Wishart et al. 2018). The following elaborates on the interaction label data, i.e., datasets of labelled DTI pairs.

**Datasets of specific interactions.** We split every dataset to train, validation and test set. The length of the positive train sets is set to 10, 20 or 40, the length of positive valid sets is fixed to 20, and the rest of the positive samples are considered positive test sets. By convention, we randomly sample pairs from drugs and targets of positive samples to generate the same number of negative samples. DGIDB::BLOCKER means antagonist interactions including alpha blockers, beta blockers, and calcium channel blockers, which contains 253 positive samples. GNBR::E- means decreasing expression or production, which contains 1401 positive samples. DGIDB::AGONIST means a drug binds to a target receptor and activates the receptor to produce a biological response, which contains 1338 positive samples.

**Datasets of general interactions.** We adopt the datasets used in (Ma et al. 2022) and remove the samples for which we cannot find their corresponding intrinsic data. The Drugbank and Drugcentral datasets contain 18480 and 18066 samples respectively. The split ratio of train, valid and test set is 6:2:2.

# D Evaluation Protocol

We report accuracy (ACC), AUC (area under the receiver operating characteristic curve), and AUPR (area under the precision-recall curve) as the performance metrics. For experiments under data scarcity (RQ1), we construct 5 different cross-validation variants of each dataset and report the means and standard deviations of performance across the 5 variants. For other experiments, we report the means and standard deviations of performance across 5 independent runs with different random seeds on fixed data. TransformerCPI2.0(Chen et al. 2023) only provides its model for inference instead of the training code, so its results are deterministic without a standard deviation value except for the RQ1 experiments.

# E Implementation Details

We keep the model architecture and hyper-parameters invariant across all datasets except that for general interaction datasets, due to the sufficient labelled samples, we do not use unlabelled samples to generate pseudo labels. Due to the input length limitation of the ESM-MAS-1b model, we truncate long protein sequences to take their middle 1022 amino acid residues as input. We use the TransE (Bordes et al. 2013) method to pretrain our KG embeddings and GCN (Kipf and Welling 2017) to encode molecule graphs. We use the Adam optimizer to optimize our model. The learning rate is set to 1e-2 for joint training and 1e-3 otherwise. The train and test batch size is 16. $\alpha_{extr}, \alpha_{intr}, \beta_{extr}, \beta_{intr}\beta_g$ are 2e-2, 2e-4, 3e-5, 3e-4, 3e-3 respectively. The $\alpha_{extr}$ is much larger than $\alpha_{intr}$ since the inference of extrinsic expert is much faster than that of intrinsic expert. $\gamma_{extr}, \gamma_{intr}, \gamma_g$ are set to make the ratios of the total length of ground-truth labels versus pseudo labels 0.3, 1, and 1 respectively. We conduct experiments on NVIDIA 3090 graphic cards and Intel Xeon Gold 6348 CPU.

# F Hyper-parameter Sensitivity Analysis

In this experiment shown in Fig. 2, we test the impact of the major hyper-parameters of MoseDTI, which include the sample selecting rate for the extrinsic expert $\beta_{extr}$, sample selecting rate for the intrinsic expert $\beta_{intr}$, sample selecting rate for the joint training $\beta_g$. We can observe that, for all these three parameters, performance grows as the selecting rate grows. Subsequently, performance plateaus and then decreases as the selection rate grows. This is probably because experts need enough pseudo labels to be trained while an overly high choosing rate may corrupt the quality of pseudo labels. The optimal $\beta$ of the intrinsic expert differs from those of the extrinsic expert by a factor of 100 because the two $\alpha$ have a reverse difference of 100.

# G Discovering New Patterns by Synergizing Experts

To illustrate how synergizing experts can improve model performance, we conduct an experiment to observe whether experts can discover new data patterns after being synergized. After the KG pretraining, we use only one drug-target pair in the blocker dataset as the positive training set to train the extrinsic expert. The paths between the drug and the target of the training pair are shown in Fig.1(a), which is a frequent pattern in the whole dataset. It describes if a drug

Table 1: Descriptions of key notations.

| Notations | Descriptions |
|---|---|
| $D$ | The drug set. |
| $T$ | The target set. |
| $\{(d_i, t_j)\}$ | A set of DTI pairs. $d_i \in D$ and $t_j \in T$ denote the i-th drug and the t-th target respectively. |
| $(X^p, X^n)$ | A DTI dataset. $X^p$ and $X^n$ denote the postitive and negative drug-target pairs respectively. |
| $p_{ij}, p_{ij}^{in}, p_{ij}^{ex}$ | The probabilities estimated of the entire model, the intrinsic expert and the extrinsic expert respectively. |
| $\mathcal{KG} = (E, R, O)$ | The definition of a knowledge graph with an entity set $E$, a relation set $R$ and a set of observed triplets. |
| $(\mathsf{h}, \mathsf{r}, \mathsf{t})$ | A triple consists of a head entity $mathsfh$, a relation $mathsfr$, and a tail entity $mathsft$. |
| $SM_{d_i}$ | SMILES of drug $d_i$. |
| $AS_{t_i}$ | amino acid sequence of target $t_i$. |
| $\mathbf{h}_{d_i}^{ex}, \mathbf{h}_{t_i}^{ex}, \mathbf{h}_{d_i}^{in}, \mathbf{h}_{t_i}^{in}$ | Embedding of drug $d_i$ or target $t_i$ generated by extrinsic or intrinsic expert. |
| $g^{ex}, g^{in}$ | The extrinsic or intrinsic classifier (a part of the extrinsic or intrinsic expert). |
| $\mathcal{MG}_{d_i} = (\mathcal{V}_{d_i}, \mathcal{E}_{d_i})$ | The definition of a molecule graph of drug $d_i$, with its node set $\mathcal{V}_{d_i}$ and edge set $\mathcal{E}_{d_i}$. |
| $\{\mathbf{e}_m^{(0)} \mid m = 1, ..., M\}$ | The feature set for target $t_j$ of length M generated by ESM-MSA-1b. |
| $w_{ij}$ | The blending weight output by the gating model. |

Table 2: Performance comparison between direct training and transfering from general to specific interaction datasets.

| | DGIDB::AGONIST | | | DGIDB::BLOCKER | | | GNBR::E- | | |
|---|---|---|---|---|---|---|---|---|---|
| | ACC | AUC | AUPR | ACC | AUC | AUPR | ACC | AUC | AUPR |
| MDTips | 64.23±4.13 | 73.51±2.91 | 72.37±2.87 | 91.25±2.34 | 97.27±0.86 | 96.90±0.79 | 69.61±3.43 | 82.35±3.89 | 80.78±4.86 |
| MDTips-transfer | 53.00±12.28 | 54.55±17.23 | 59.16±14.00 | 83.29±3.38 | 91.17±2.02 | 90.61±2.08 | 67.87±3.79 | 75.07±6.08 | 73.06±6.70 |

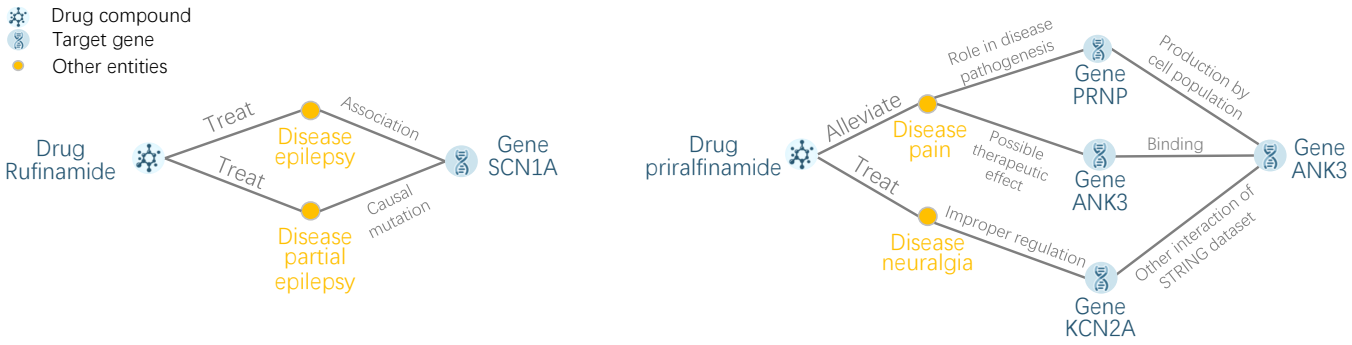
treats a disease and the disease is associated with or caused by mutation of a gene, then the drug may be a blocker of the gene. However, the extrinsic expert may not identify some drug-target pairs with relatively rare path patterns in the positive testing set. For example, a testing pair may have a totally different path pattern illustrated in Fig.1(b). In this pattern, the drug and target are connected by at least 3 hops with complicated drug-disease-gene-gene paths. This testing pair is given a score of only 0.0457 by the extrinsic expert, which means being trained with the drug-target pattern (a), the extrinsic expert cannot identify a different drug-target pattern (b). However, the intrinsic expert assigns pair (b) a score of 0.9999, which means this sample is regarded as a simple positive case from the perspective of intrinsic data. After being trained with the pseudo labels generated by the intrinsic expert, the extrinsic expert assigns a score of 0.8418 for pair (b), which indicates the extrinsic expert has learned to recognize this pair pattern with the aid of the intrinsic expert.

# H Performance on General Interaction Datasets with Limited Labels

There have been massive labelled samples accumulated for general DTI interaction in both the drugcentral and drugbank datasets, so we regard few-shot learning for specific interaction as more meaningful. To validate that our method can also handle few-shot scenarios on general interaction datasets, here we also evaluate our method on the general datasets with 10 positive training samples, compared with the competitive baseline MDTips(Xia et al. 2023), shown in Tab. 3. From the result, our models still outperform the competitive baseline by a large margin.

# I Results of 20 and 40 Shots of Experiment RQ1

Due to the page limit of the paper, we display the results of 20 and 40 shots of experiment RQ1 here, see Tab. 4




(a) (b)

Figure 1: Illustration for path patterns between pairs. (a) is the path pattern in the training set and (b) is a totally different pattern in the test set.

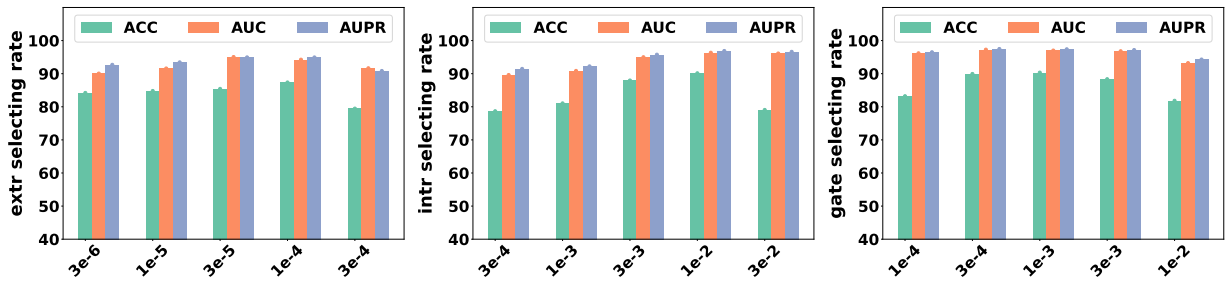



Figure 2: Hyper-parameter sensitivity analysis on the blocker datasets.

Table 3: Model performance on general DTI datasets with 10 positive labels.

| | DrugBank | | | DrugCentral | | |
|---|---|---|---|---|---|---|
| | ACC | AUC | AUPR | ACC | AUC | AUPR |
| MDTips | 55.94±3.34 | 62.47±4.70 | 61.96±3.35 | 56.89±4.16 | 61.31±6.36 | 62.50±5.48 |
| Mose-intr | 62.51±1.66 | 63.70±1.27 | 66.00±2.61 | 62.14±1.22 | 65.41±2.02 | 67.28±2.55 |
| Mose-extr | 61.12±1.52 | 63.52±4.39 | 64.72±1.13 | 62.83±1.43 | 66.65±3.12 | 66.67±2.39 |
| MoseDTI | 61.76±1.84 | 64.89±1.00 | 66.25±2.38 | 63.64±1.05 | 67.10±0.68 | 68.25±1.51 |

Table 4: Model performance of 20 and 40 shots in experiment RQ1.

| | | | DGIDB::AGONIST | | | DGIDB::BLOCKER | | | GNBR::E- | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | ACC | AUC | AUPR | ACC | AUC | AUPR | ACC | AUC | AUPR |
| 20 shots | Intrinsic Method | GNNCPI | 55.26±4.63 | 60.01±4.99 | 58.86±2.79 | 68.54±7.19 | 74.97±8.57 | 76.77±8.64 | 51.23±3.12 | 60.38±1.89 | 58.93±2.72 |
| | | TFCPI | 60.56±4.26 | 68.56±3.42 | 68.73±2.49 | 51.79±3.33 | 68.67±2.63 | 64.34±3.93 | 62.40±1.99 | 77.46±2.12 | 77.60±1.58 |
| | | TFCPI2.0 | 50.92±0.24 | 53.27±0.32 | 51.06±0.27 | 37.16±0.48 | 33.99±0.88 | 39.05±0.34 | 42.17±0.14 | 40.65±0.19 | 41.99±0.11 |
| | | Mose-intr | **70.89±1.46** | 77.18±2.39 | 75.43±4.71 | 89.34±3.07 | 93.42±2.14 | 94.34±1.64 | 68.32±1.94 | 78.00±1.99 | 77.65±2.06 |
| | Extrinsic Method | TransE | 50.00±0.00 | 50.15±0.10 | 50.67±0.10 | 50.00±0.00 | 54.84±0.68 | 56.24±0.88 | 50.00±0.00 | 48.80±0.10 | 48.89±0.11 |
| | | RotatE | 50.00±0.00 | 50.21±0.10 | 50.35±0.06 | 50.00±0.00 | 51.08±0.30 | 50.74±0.58 | 50.00±0.00 | 48.87±0.17 | 49.97±0.12 |
| | | TriModel | 40.01±22.37 | 45.02±2.30 | 47.08±1.92 | 49.95±0.20 | 26.52±2.35 | 38.82±1.83 | 50.01±0.06 | 49.33±1.57 | 50.92±1.05 |
| | | AMGDTI | 66.18±2.98 | 75.22±2.64 | 76.58±4.99 | 86.17±2.62 | 96.65±2.45 | 96.37±0.32 | 75.37±5.20 | 84.97±1.80 | 84.72±1.17 |
| | | Mose-extr | 70.54±1.88 | 80.32±2.72 | 80.17±3.08 | 88.86±2.79 | 98.13±0.25 | 97.45±0.29 | 81.93±3.10 | 92.34±2.22 | 92.71±2.49 |
| | Hybrid Method | KG-MTL | 56.81±1.27 | 57.74±5.56 | 51.71±4.07 | 73.31±4.12 | 81.56±3.32 | 84.31±3.14 | 59.16±2.16 | 65.65±2.06 | 64.47±3.06 |
| | | MDTips | 68.41±3.56 | 80.00±3.17 | 79.13±4.27 | 89.12±4.64 | 98.11±0.58 | 98.04±0.64 | 80.10±4.72 | 90.85±2.05 | 90.47±2.09 |
| | | MoseDTI | 70.28±2.36 | **80.82±3.59** | **80.95±5.02** | **91.15±4.45** | **98.46±0.62** | **98.04±0.39** | **81.98±4.76** | **92.56±1.71** | **93.96±1.84** |
| 40 shots | Intrinsic Method | GNNCPI | 66.13±1.72 | 73.04±2.20 | 71.32±2.85 | 73.40±3.82 | 82.95±2.48 | 84.83±1.70 | 61.97±4.15 | 70.67±3.31 | 70.35±3.71 |
| | | TFCPI | 63.83±2.03 | 71.19±1.03 | 70.32±0.69 | 69.72±2.39 | 78.47±2.46 | 76.74±2.31 | 70.55±2.64 | 77.50±2.15 | 76.79±2.08 |
| | | TFCPI2.0 | 50.92±0.22 | 53.20±0.36 | 50.97±0.25 | 36.91±0.48 | 33.73±0.71 | 38.97±0.30 | 42.10±0.15 | 40.64±0.18 | 41.97±0.10 |
| | | Mose-intr | 70.51±7.23 | 81.68±2.50 | 79.88±3.13 | 93.97±1.17 | 96.96±1.32 | 96.24±1.36 | 70.88±4.23 | 78.39±4.25 | 77.97±4.29 |
| | Extrinsic Method | TransE | 50.00±0.00 | 50.16±0.10 | 50.69±0.06 | 50.00±0.00 | 54.52±0.31 | 56.10±0.58 | 50.00±0.00 | 48.60±0.27 | 48.76±0.22 |
| | | RotatE | 50.00±0.00 | 50.24±0.26 | 50.42±0.18 | 50.00±0.00 | 50.65±0.52 | 50.68±0.56 | 50.00±0.00 | 48.95±0.19 | 49.92±0.18 |
| | | TriModel | 50.02±0.04 | 39.88±0.57 | 44.86±0.64 | 49.97±0.21 | 22.31±3.05 | 36.30±1.52 | 50.00±0.00 | 43.22±3.11 | 46.60±2.09 |
| | | AMGDTI | 72.37±2.81 | 81.98±2.47 | 82.19±3.58 | 93.45±2.68 | 97.67±2.64 | 97.18±0.35 | 83.41±2.38 | 91.76±1.24 | 91.35±1.97 |
| | | Mose-extr | **76.88±2.72** | 85.97±1.00 | 85.21±1.14 | 95.95±1.56 | **98.78±0.25** | 98.14±0.44 | **85.44±0.86** | **93.93±0.75** | 94.19±0.84 |
| | Hybrid Method | KG-MTL | 59.24±1.00 | 63.50±2.08 | 58.39±2.53 | 84.12±1.30 | 90.63±1.25 | 92.22±1.30 | 62.66±2.99 | 68.38±4.60 | 68.77±3.36 |
| | | MDTips | 73.91±3.97 | 86.46±1.36 | **85.94±1.44** | 84.42±11.57 | 98.53±0.56 | 98.34±0.72 | 77.98±6.49 | 92.11±1.43 | 91.94±1.53 |
| | | MoseDTI | 76.07±3.82 | **86.71±3.10** | 85.78±3.73 | **96.42±6.82** | 98.13±0.53 | **98.39±0.59** | 84.83±4.11 | 93.42±3.79 | **94.29±3.63** |

## References


Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Chen, L.; Fan, Z.; Chang, J.; Yang, R.; Hou, H.; Guo, H.; Zhang, Y.; Yang, T.; Zhou, C.; Sui, Q.; et al. 2023. Sequence-based drug design as a concept in computational drug design. *Nature Communications*, 14(1): 4217.

Ioannidis, V. N.; Song, X.; Manchanda, S.; Li, M.; Pan, X.; Zheng, D.; Ning, X.; Zeng, X.; and Karypis, G. 2020. DRKG - Drug Repurposing Knowledge Graph for Covid-19. https://github.com/gnn4dr/DRKG/.

Kipf, T.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. *ArXiv*, abs/1609.02907.

Ma, T.; Lin, X.; Song, B.; Philip, S. Y.; and Zeng, X. 2022. Kg-mtl: knowledge graph enhanced multi-task learning for molecular interaction. *IEEE Transactions on Knowledge and Data Engineering*.

Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46: D1074–D1082.

Xia, X.; Zhu, C.; Zhong, F.; and Liu, L. 2023. MDTips: a multimodal-data-based drug–target interaction prediction system fusing knowledge, gene expression profile, and structural data. *Bioinformatics*, 39(7): btad411.