

## Supplementary Data

A dual-modality complex-valued fusion method for predicting side effects of drug–drug interactions based on graph neural network

Chuanze Kang<sup>1</sup>, Han Zhang<sup>1\*</sup>, Yanbin Yin<sup>2</sup>

<sup>1</sup> College of Artificial Intelligence, Nankai University, Tianjin, China

<sup>2</sup> Department of Food Science and Technology, University of Nebraska – Lincoln, Lincoln, USA

\* [zhanghan@nankai.edu.cn](mailto:zhanghan@nankai.edu.cn)

1. Parameters of baseline methods
2. Experimental results of DMCF-DDI with SC and ASC
3. Experimental results of DMCF-DDI and SOTA methods in regular and transductive situations.
4. Experiment results of five variants and DMCF-DDI.
5. Experiment results of case study for six diseases.
6. Experiment results of multi-class dataset.
7. Experiment results of contrastive learning.
8. Experiment results of PPI dataset.
9. Experiment results of multi-modality comparison.
10. Discussion about databases.
11. Supplementary discussion about using of ComplexE and RotatE for link prediction in knowledge graph.
12. Visualization of Quate.
13. New Interactions predicted by DMCF-DDI.
14. How dual-modality improve the single-modal FP performance.

# 1. Parameters of baseline methods

**Table S1: The parameters of baseline methods.**

Methods	Description	Parameters
STNN	STNN generates low-dimensional embeddings of the drugs according to the fingerprint, and views the prediction of side effects of DDIs as a 3D-tensor completion problem.	For regular situations, the epoch number is 100, the learning rate is 1e-3, and R=400. For Task A and B, the epoch number is 40, the learning rate is 5e-4, R=400.
Decagon	Decagon extracts the topological structure features from DKG by multi-relation graph convolution network, and computes prediction probability by a rank-d DEDICOM tensor decomposition.	The initial Embedding size is 256, the hidden dimension is 64, and the number of graph convolution layers is 2. For regular situations, the epoch number is 50, learning rate is 1e-3. For Task A and B, the epoch number is 30, learning rate is 5e-4.
MDNN	MDNN learns the topological feature from DKG, and extracts similarity information from heterogeneous features by PCA. Two modal features are concatenated as the final features.	For a graph convolution network, the initial embedding size is 256, hidden dimension is 64. For a fully-connected network, the number of layers is 2, hidden dimension is {256, 64}. The number of components of PCA is 64. For regular situations, the epoch number is 50, learning rate is 5e-3 For Task A and B, the epoch number is 30, learning rate is 5e-4.
MUFFIN	MUFFIN computes the cross-modal feature matrix of drug by cross-product, and uses CNN to fuse features of the molecular graph and DKG.	The feature dimension of the molecular graph is 300, and the entity dimension of the knowledge graph is 100. For the three-layer convolution neural network, the kernel size is {64, 32, 16}. For the two-layer fully-connected network classifier, the hidden dimension is {256, 128}. For regular situations, the epoch number is 198, learning rate is 1e-3 For Task A and B, the epoch number is 30, learning rate is 1e-4.
GCNMK	GCNMK adopts two DDI graph kernel encoders on fingerprint and DKG features. Then, two drug vectors are concatenated to be the vector of a drug pair.	The input of model is concatenation of fingerprint and drug-target adjacent. The first graph kernel encoder uses drug-drug interactions as input graph. The second graph kernel encoder uses associations based on drug-target-drug meta paths as input graph. Each graph kernel contains two graph convolution networks (2-layer and 1-layer). The hidden dimensions of 2-layer and 1-layer GCN are {256, 160} and {256}. The hidden dimensions of decoder are {32, 16}. For regular situations, the epoch number is 50, learning rate is 1e-3. For Task A and B, the epoch number is 30, learning rate is 5e-4.
LR-GNN	LR-GNN utilizes the information from intermediate layers of GCN-encoder to construct the DPR. LR-GNN-FP and	For LR-GNN-FP and LR-GNN-KG, the initial embedding size is 256, and the hidden dimension is {64, 64, 64}.

	LR-GNN-KG are built to encode two modal DPR.	For regular situations, the epoch number is 50, learning rate is 1e-3. For Task A and B, the epoch number is 30, learning rate is 5e-4 and 1e-3.
DMMHA-DDI	DMMHA-DDI concatenates dual-modal features of two drugs and fuses them by multi-head attention (MHA) mechanism. For a fair comparison of the fusion operators, its encoders are the same as those of DMCF-DDI.	MHA contains 8 heads with hidden dimension 8. 2-layer encoders are used to obtain the query, key, and value features. By a linear transformations and residual connection, the output of MHA is updated with input. Then, a FFNet with {64, 128, 64} and residual connection are used to update the output of MHA. For regular situations, the epoch number is 50, learning rate is 1e-3 For Task A and B, the epoch number is 30, learning rate is 5e-4 and 1e-3.
Quate	Quate uses quaternion embeddings with three imaginary parts to represent the entity and relation of the knowledge graph.	GCN-encoders of Quate are the same as that of DMCF-DDI. The real and three imaginary parts of the drug pair quaternion features are concatenated as the inputs of the classifier. The classifier is a fully connected network with a hidden dimension is 256. For regular situations, the epoch number is 50, learning rate is 1e-3. For Task A and B, the epoch number is 30, learning rate is 5e-4.

The Quate method is described in detail as follows. A quaternion vector consists  $Q \in H^k$  consists of one real part and three imaginary parts, defined as  $Q = a + bi + cj + dk$ , where  $a, b, c, d \in R^k$  are real vectors and  $i, j, k$  are imaginary units.  $i, j, k$  are square roots of -1, satisfying the Hamilton's rules:  $i^2 = j^2 = k^2 = ijk = -1$ . The quaternion multiplication between  $Q_1 = a_1 + b_1i + c_1j + d_1k$  and  $Q_2 = a_2 + b_2i + c_2j + d_2k$  is defined as:

$$\begin{aligned}
Q_1 \otimes Q_2 = & (a_1 \odot a_2 - b_1 \odot b_2 - c_1 \odot c_2 - d_1 \\
& \odot d_2) \\
& + (a_1 \odot b_2 + b_1 \odot a_2 + c_1 \odot d_2 \\
& - d_1 \odot c_2)i \\
& + (a_1 \odot c_2 - b_1 \odot d_2 + c_1 \odot a_2 \\
& + d_1 \odot b_2)j \\
& + (a_1 \odot d_2 + b_1 \odot c_2 - c_1 \odot b_2 \\
& + d_1 \odot a_2)k
\end{aligned} \tag{1}$$

Let the real part and three imaginary parts of drug quaternion representations be:

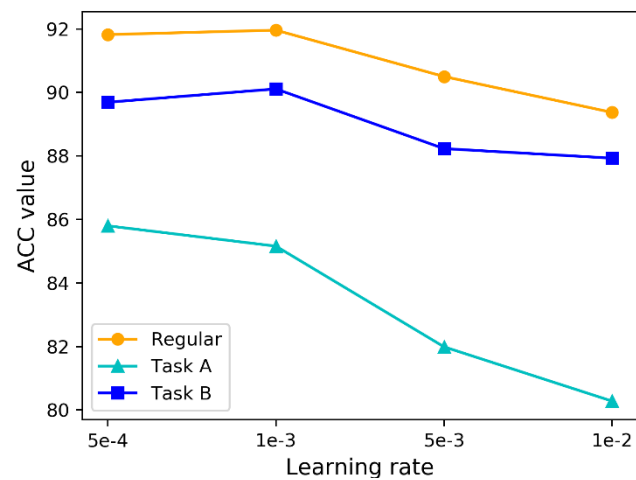
$$\begin{aligned}
Z_i &= H_{fp}[i] + H_{kg}[i]i + S_{fp}[i]j + S_{kg}[i]k, \\
Z_j &= H_{fp}[j] + H_{kg}[j]i + S_{fp}[j]j + S_{kg}[j]k.
\end{aligned} \tag{2}$$

Then, quaternion multiplication is used to fuse two drug quaternion representations:

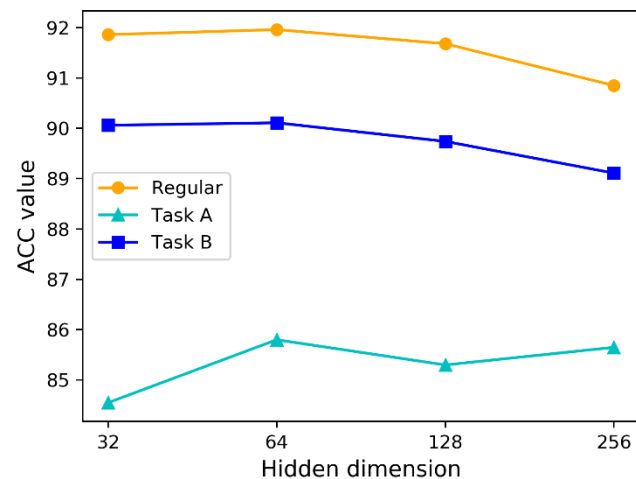
$$Z_{ij} = W_{fc} \text{Cat}(Z_i \otimes Z_j), \tag{3}$$

where  $W_{fc}$  is the weight matrix of the fully-connected network. We concatenate the real part and three imaginary parts of  $Z_i \otimes Z_j$  as the feature of the drug pair  $(i, j)$ . Then, we calculate the prediction probability of each side effect by linear classifier:

$$p_{ij}^r = \text{sigmoid}(w_r^T Z_{ij} + b_r). \tag{4}$$



**Fig. S1 The effect of the learning rate on the performance of DMCF-DDI.**



**Fig. S 2 The effect of the hidden dimension on the performance of DMCF-DDI.**

**Table S2: The number of parameters of LR-GNN, MUFFIN, Quate, and DMCF-DDI under different embedding dimensions.**

**(a) The number of parameters in each component of LR-GNN.**

	Total Parameter		
LR-GNN	5664384		
	encoder	Decoder	
	5450240	214144	
		Modal feature fusion	classifier
		90752	123392

**(b) The number of parameters in each component of DMCF-DDI.**

	Total Parameter		
DMCF-DDI	5548936		
	encoder	decoder	
	5390848	158088	
		Modal feature fusion	classifier
		32768	125320

**(c) The number of parameters in each component of methods under 32, 64, and 256 embedding dimensions.**

	Total	Encoder	Decoder	Fusion	Classifier
DMMHA-32	5421404	5328896	92508	29848	62660
LR-GNN-32	5434432	5349888	84544	22848	61696
Quate-32	5381188	5328896	52292	20480	31812
DMCF-DDI-32	5408904	5328896	80008	<b>16384</b>	63624
DMMHA-64	5696220	5390848	305372	57624	62660
LR-GNN-64	5664384	5450240	214144	90752	123392
Quate-64	5502660	5390848	111812	49152	62660
DMCF-DDI-64	5548936	5390848	158088	<b>32768</b>	125320
DMMHA-256	6228700	5999104	229596	166936	62660
LR-GNN-256	8592384	6654464	1937920	1444352	493568
DMCF-DDI-256	6625672	5999104	626568	<b>131072</b>	495496

**(d) The number of parameters in each component of methods under 32, 64, and 256 embedding dimensions.**

	Total	Encoder	Decoder	Fusion	Classifier
MUFFIN-32	665085	0	418301	418301	246784
DMCFPretrain-32	348436	0	348436	<b>283728</b>	62660
MUFFIN-100	3193665	0	2946881	2946881	246784
DMCF-DDI-64	5548936	5390848	158088	<b>32768</b>	125320

In Table S2(d), to fairly compare the fusion operators of MUFFIN and DMCF-DD, we establish the DMCFPretrain method that uses the same pre-trained features as MUFFIN. The parameter numbers of the complex-valued fusion operator in DMCFPretrain are only 25k. We use pre-trained features as trainable parameters, and their parameter numbers are 258k and have a large proportion. The overall architecture of DMCFPretrain is shown in Fig. S1:

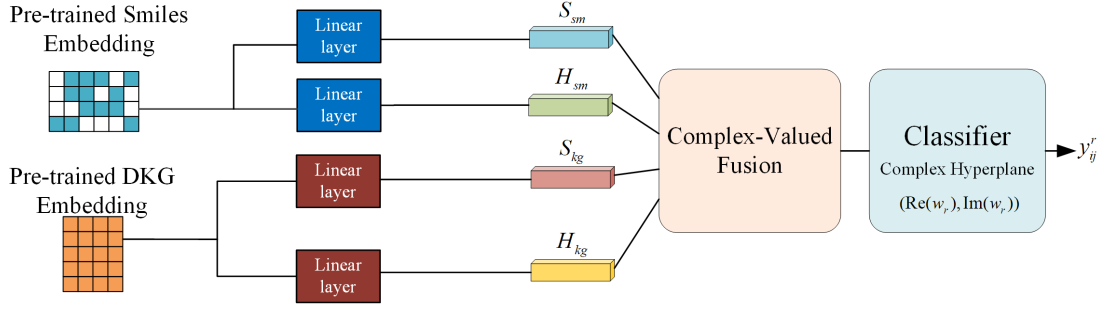


Fig. S 3 The overall architecture of DMCFPretrain.

## The pseudocode of DMCF-DDI

---

**Algorithm 1** The forward process of DMCF-DDI

---

**Input:** DDI triplets  $\{(d_i, d_j, r)\}$ , Fingerprint  $X_{fp}$ , DKG  $G_{\{kg\}} = \{(h, r, t) | h, t \in V, r \in R\}$

**Output:** Prediction Probability  $p_{ij}^r = f\left((d_i, d_j, r) | (\theta, X_{fp}, G_{kg})\right)$

Initialize all parameters

$$H_{fp}^0 = X_{fp} W_{fp}^0$$

**for**  $l = 1, \dots, L$  **do**

$$H_{fp}^l = \text{RGCN}_{fp}(\{A^{dd}\}, H_{fp}^{l-1}, \{H_{fp}^{l-1}\})$$

the randomly generated trainable word embedding  $H_d^0 \in R^{N_d \times k}$  and  $H_p^0 \in R^{N_p \times k}$

**for**  $l = 1, \dots, L$  **do**

$$H_d^l = \text{RGCN}_{kg}(\{A^{dd}, A^{dp}\}, H_d^{l-1}, \{H_d^{l-1}, H_p^{l-1}\})$$

$$H_p^l = \text{RGCN}_{kg}(\{A^{pd}, A^{pp}\}, H_p^{l-1}, \{H_d^{l-1}, H_p^{l-1}\})$$

$$S_{fp} = \sigma(\sigma(H_{fp}^0) W_{fp}^{cf})$$

$$S_{kg} = \sigma(\sigma(H_d^0) W_{kg}^{cf})$$

$$\text{Re}(H_{ij}) = \text{LN}(H_{fp}[i] + S_{fp}[j])$$

$$\text{Im}(H_{ij}) = \text{LN}(H_{kg}[i] + S_{kg}[j])$$

$$\text{Re}(H_{ji}) = \text{LN}(H_{kg}[j] + S_{kg}[i])$$

$$\text{Im}(H_{ji}) = \text{LN}(H_{fp}[j] + S_{fp}[i])$$

$$Z_{ij}^{asc} = H_{ij} \odot H_{ji}$$

$$p_{ij}^r = \text{sigmoid}\left(O(< Z_{ij}^{asc}, \overline{w_r} >)\right)$$


---

## The formula of the Y-shaped framework

$$\begin{aligned} Z_1 &= \text{encoder}_1(X_1) \\ Z_2 &= \text{encoder}_2(X_2) \end{aligned} \tag{5}$$

$$Z_{ij} = \text{fusion}(Z_1[i], Z_2[j])$$

$$p_{ij} = \text{classifier}(Z_{ij})$$

$encoder_1$  and  $encoder_2$  are the encoders of two branches.  $X_1$  and  $X_2$  are the inputs from different modalities.  $\text{fusion}$  is the fusion operator.  $Z_{ij}$  is the drug pair representation.  $\text{classifier}$  is the classifier for predicting drug-drug side effects.

## The formula of metric

**Accuracy:** average accuracy over different labels can be defined as:

$$ACC = \frac{1}{N_{label}} \sum_i \frac{(TP_i + TN_i)}{TP_i + TN_i + FP_i + FN_i} \quad (6)$$

**Area under the receiver operating characteristic (AUC):** The AUC is calculated by computing the area under the Receiver Operating Characteristic (ROC) curve. The ROC curve is a plot of the true positive rate (TPR) against the false positive rate (FPR) for various threshold values. The AUC is the area under this curve, ranging from 0 to 1, with higher values indicating better performance.

**Area under the precision-recall-curve (AUPR):** The AUPR is calculated by computing the area under the Precision-Recall (PR) curve. The PR curve is a plot of precision against recall for various threshold values. The AUPR is the area under this curve, ranging from 0 to 1, with higher values indicating better performance.

**AP@50:** The Average Precision at n (AP@n) is a performance metric used to evaluate the precision of a ranked list of items up to a certain cutoff point n. It is commonly used in information retrieval and binary classification tasks. Here's the definition and formula for AP@n:

- Let's denote  $\text{rel}(i)$  as the relevance of the  $i$ th item in the ranked list, where  $\text{rel}(i)=1$  if the item is relevant (positive) and  $\text{rel}(i)=0$  otherwise. The precision at position  $k$  ( $P@k$ ) is defined as the proportion of relevant items among the top  $k$  items:

$$P@k = \frac{\text{Number of relevant items in top } k}{k} \quad (7)$$

- Then, AP@n is calculated as the average of the precision values at each relevant position up to  $n$ , weighted by the change in recall from the previous position to the current position:

$$AP@n = \frac{1}{\text{Number of relevant items}} \sum_{k=1}^n P@k \times \text{rel}(k) \quad (8)$$

- In other words, AP@n sums the precision values at each relevant position up to  $n$ , and then divides by the total number of relevant items.

## 2. Experimental results of DMCF-DDI with SC and ASC

**Table S3: Experimental results of DMCF-DDI with SC and ASC in the regular situation.**

	ACC	AUC	AUPR	MAP
SC-MLP	91.20 $\pm$ 0.00	96.35 $\pm$ 0.00	94.91 $\pm$ 0.00	94.50 $\pm$ 0.00
SC-CIP	91.95 $\pm$ 0.00	96.77 $\pm$ 0.00	95.57 $\pm$ 0.00	95.74 $\pm$ 0.00
ASC-MLP	91.70 $\pm$ 0.00	96.70 $\pm$ 0.00	95.50 $\pm$ 0.00	95.77 $\pm$ 0.00
ASC-CIP	91.50 $\pm$ 0.00	96.67 $\pm$ 0.00	95.47 $\pm$ 0.00	95.77 $\pm$ 0.00

**Table S4: Experimental results of DMCF-DDI with SC and ASC in Task A.**

	ACC	AUC	AUPR	MAP
SC-MLP	83.41 $\pm$ 0.02	91.09 $\pm$ 0.01	89.88 $\pm$ 0.01	93.15 $\pm$ 0.00
SC-CIP	84.51 $\pm$ 0.05	92.70 $\pm$ 0.04	91.46 $\pm$ 0.05	92.91 $\pm$ 0.11
ASC-MLP	85.44 $\pm$ 0.02	94.05 $\pm$ 0.00	92.91 $\pm$ 0.00	95.10 $\pm$ 0.00
ASC-CIP	86.80 $\pm$ 0.03	95.00 $\pm$ 0.00	93.70 $\pm$ 0.01	95.44 $\pm$ 0.00

**Table S5: Experimental results of DMCF-DDI with SC and ASC in Task B.**

	ACC	AUC	AUPR	MAP
SC-MLP	87.45 $\pm$ 0.01	94.59 $\pm$ 0.00	93.20 $\pm$ 0.00	92.21 $\pm$ 0.00
SC-CIP	88.40 $\pm$ 0.02	95.05 $\pm$ 0.00	93.75 $\pm$ 0.00	92.71 $\pm$ 0.00
ASC-MLP	89.96 $\pm$ 0.01	95.81 $\pm$ 0.00	94.53 $\pm$ 0.00	93.67 $\pm$ 0.00
ASC-CIP	89.93 $\pm$ 0.01	95.92 $\pm$ 0.00	94.63 $\pm$ 0.00	93.29 $\pm$ 0.00



### 3. Experimental results of DMCF-DDI and SOTA methods in regular and transductive situations.

**Table S6: Experimental results of DMCF-DDI and SOTA methods in the regular situation.**

	ACC	AUC	AUPR	MAP
STNN	81.92±0.00	89.54±0.00	87.35±0.00	85.73±0.00
Decagon	90.91±0.00	96.33±0.00	95.04±0.00	95.24±0.00
MDNN	89.98±0.00	95.65±0.00	93.94±0.00	93.18±0.00
GCNMK	90.71±0.00	96.11±0.00	94.69±0.00	94.60±0.00
MUFFIN	91.1±0.00	96.54±0.00	95.16±0.00	95.04±0.00
DMMHA-DDI	91.46±0.00	96.64±0.00	95.40±0.00	95.47±0.00
LR-GNN	91.13±0.00	96.40±0.00	95.05±0.00	95.16±0.00
Quate	91.42±0.00	96.84±0.00	95.53±0.00	95.37±0.00
DMCF-DDI	91.95±0.00	96.83±0.00	95.66±0.00	95.99±0.00

**Table S7A: Experimental results of DMCF-DDI and SOTA methods in Task A.**

	ACC	AUC	AUPR	MAP
STNN	63.12±0.14	70.63±0.13	68.60±0.09	55.10±0.34
Decagon	69.37±8.67	73.78±12.78	76.87±6.26	70.78±15.42
MDNN	81.58±0.09	92.86±0.00	91.37±0.00	94.48±0.00
GCNMK	81.37±0.06	90.65±0.01	88.89±0.01	92.53±0.01
MUFFIN	86.36±0.01	94.26±0.00	92.93±0.00	94.69±0.00
DMMHA-DDI	87.28±0.01	94.43±0.01	93.11±0.01	93.51±0.00
LR-GNN	86.49±0.01	94.39±0.00	92.94±0.00	94.48±0.00
Quate	85.93±0.01	94.33±0.00	92.85±0.00	94.99±0.00
DMCF-DDI	86.80±0.03	95.00±0.00	93.70±0.01	95.44±0.00

**Table S8A: Experimental results of DMCF-DDI and SOTA methods in Task B.**

	ACC	AUC	AUPR	MAP
STNN	76.81±0.04	85.07±0.02	82.60±0.02	78.59±0.04
Decagon	88.98±0.01	95.21±0.00	93.76±0.00	92.61±0.00
MDNN	86.74±0.01	94.52±0.00	93.12±0.00	92.31±0.01
GCNMK	83.39±0.03	91.87±0.00	90.23±0.00	88.86±0.02
MUFFIN	88.34±0.00	95.12±0.00	93.76±0.00	95.13±0.00
DMMHA-DDI	89.65±0.01	95.73±0.00	94.22±0.00	92.86±0.00
LR-GNN	89.35±0.01	95.32±0.00	93.81±0.00	92.52±0.00
Quate	89.78±0.00	96.05±0.00	94.69±0.00	93.60±0.00
DMCF-DDI	90.45±0.01	96.00±0.00	94.77±0.00	93.94±0.00

**Table S7B: Experimental results in Task A under different embedding dimensions.**

	ACC	AUC	AUPR	MAP
DMMHA-DDI-32	87.11±0.01	94.36±0.01	93.00±0.01	93.43±0.06
LR-GNN-32	85.76±0.01	93.42±0.00	91.80±0.00	92.28±0.00
DMCF-DDI-32	86.43±0.02	94.54±0.00	93.23±0.00	94.76±0.00
DMMHA-DDI-64	87.28±0.01	94.43±0.01	93.11±0.01	93.51±0.00
LR-GNN-64	86.49±0.01	94.39±0.00	92.94±0.00	94.48±0.00
Quate-32	86.19±0.01	94.21±0.00	92.76±0.00	94.98±0.00

DMCF-DDI-64	86.80 $\pm$ 0.03	95.00 $\pm$ 0.00	93.70 $\pm$ 0.01	95.44 $\pm$ 0.00
DMMHA-DDI-256	85.96 $\pm$ 0.02	94.67 $\pm$ 0.01	93.53 $\pm$ 0.01	95.23 $\pm$ 0.00
LR-GNN-256	86.19 $\pm$ 0.01	94.50 $\pm$ 0.00	93.14 $\pm$ 0.00	95.00 $\pm$ 0.00
Quate-64	85.93 $\pm$ 0.01	94.33 $\pm$ 0.00	92.85 $\pm$ 0.00	94.99 $\pm$ 0.00
DMCF-DDI-256	85.68 $\pm$ 0.03	94.67 $\pm$ 0.00	93.31 $\pm$ 0.00	94.99 $\pm$ 0.00
MUFFIN-32	85.46 $\pm$ 0.01	93.65 $\pm$ 0.00	92.38 $\pm$ 0.00	94.06 $\pm$ 0.01
DMCFPretrain-32	86.54 $\pm$ 0.00	93.61 $\pm$ 0.00	92.47 $\pm$ 0.00	93.15 $\pm$ 0.02

**Table S8B: Experimental results in Task B under different embedding dimensions.**

	ACC	AUC	AUPR	MAP
DMMHA-DDI-32	89.18 $\pm$ 0.01	95.26 $\pm$ 0.00	93.54 $\pm$ 0.01	91.58 $\pm$ 0.04
LR-GNN-32	89.42 $\pm$ 0.01	95.39 $\pm$ 0.00	93.94 $\pm$ 0.01	92.75 $\pm$ 0.02
DMCF-DDI-32	90.43 $\pm$ 0.01	95.98 $\pm$ 0.00	94.71 $\pm$ 0.00	93.84 $\pm$ 0.00
DMMHA-DDI-64	89.65 $\pm$ 0.01	95.73 $\pm$ 0.00	94.22 $\pm$ 0.00	92.86 $\pm$ 0.00
LR-GNN-64	89.35 $\pm$ 0.01	95.32 $\pm$ 0.00	93.81 $\pm$ 0.00	92.52 $\pm$ 0.00
Quate-32	88.86 $\pm$ 0.01	95.62 $\pm$ 0.00	94.07 $\pm$ 0.00	92.72 $\pm$ 0.00
DMCF-DDI-64	90.45 $\pm$ 0.01	96.00 $\pm$ 0.00	94.77 $\pm$ 0.00	93.94 $\pm$ 0.00
DMMHA-DDI-256	89.89 $\pm$ 0.00	95.90 $\pm$ 0.00	94.45 $\pm$ 0.00	93.10 $\pm$ 0.01
LR-GNN-256	89.69 $\pm$ 0.01	89.69 $\pm$ 0.01	95.78 $\pm$ 0.00	94.45 $\pm$ 0.00
Quate-64	89.78 $\pm$ 0.00	96.05 $\pm$ 0.00	94.69 $\pm$ 0.00	93.60 $\pm$ 0.00
DMCF-DDI-256	89.32 $\pm$ 0.01	89.32 $\pm$ 0.01	95.9 $\pm$ 0.00	94.69 $\pm$ 0.00
MUFFIN-32	88.33 $\pm$ 0.01	95.02 $\pm$ 0.00	93.69 $\pm$ 0.00	92.57 $\pm$ 0.00
DMCFPretrain-32	90.28 $\pm$ 0.01	95.99 $\pm$ 0.00	94.7 $\pm$ 0.00	93.79 $\pm$ 0.00

## 4. Experiment results of five variants and DMCF-DDI.

For DMCF-DDI-ASC\* variant, the DPRs of positive-reverse order  $H_{ij}, H_{ji}$  are constructed:

$$\begin{aligned} \text{Re}(H_{ij}) &= \text{LN}(H_{fp}[i] + S_{fp}[j]) = FP_{ij}^{asc*}, \\ \text{Im}(H_{ij}) &= \text{LN}(H_{kg}[i] + S_{kg}[j]) = KG_{ij}^{asc*}, \\ \text{Re}(H_{ji}) &= \text{LN}(H_{fp}[j] + S_{fp}[i]) = KG_{ij}^{asc*}, \\ \text{Im}(H_{ji}) &= \text{LN}(H_{kg}[j] + S_{kg}[i]) = FP_{ij}^{asc*} \end{aligned} \quad (9)$$

For DMCF-DDI-SSC variant, the DPRs of positive-reverse order  $H_{ij}, H_{ji}$  are constructed:

$$\begin{aligned} \text{Re}(H_{ij}) &= \text{LN}(H_{fp}[i] + H_{fp}[j]) = FP_{ij}^{ssc}, \\ \text{Im}(H_{ij}) &= \text{LN}(H_{kg}[i] + H_{kg}[j]) = KG_{ij}^{ssc}, \\ \text{Re}(H_{ji}) &= \text{LN}(S_{fp}[i] + S_{fp}[j]) = KG_{ij}^{ssc}, \\ \text{Im}(H_{ji}) &= \text{LN}(S_{kg}[i] + S_{kg}[j]) = FP_{ij}^{ssc} \end{aligned} \quad (10)$$

For DMCF-DDI-SSC\* variant, the DPRs of positive-reverse order  $H_{ij}, H_{ji}$  are constructed:

$$\begin{aligned} \text{Re}(H_{ij}) &= \text{LN}(H_{fp}[i] + H_{kg}[j]) = FP_{ij}^{ssc*}, \\ \text{Im}(H_{ij}) &= \text{LN}(S_{fp}[i] + S_{kg}[j]) = KG_{ij}^{ssc*}, \\ \text{Re}(H_{ji}) &= \text{LN}(S_{kg}[i] + S_{fp}[j]) = KG_{ij}^{ssc*}, \\ \text{Im}(H_{ji}) &= \text{LN}(H_{kg}[i] + H_{fp}[j]) = FP_{ij}^{ssc*} \end{aligned} \quad (11)$$

**Table S9: Experimental results of DMCF-DDI and SOTA methods in the regular situation.**

	ACC	AUC	AUPR	MAP
DMCF-DDI-SC*	82.55±0.00	89.99±0.00	88.22±0.00	88.24±0.00
DMCF-DDI-SC	91.95±0.00	96.77±0.00	95.57±0.00	95.74±0.00
DMCF-DDI-ASC*	86.72±0.00	94.99±0.00	93.69±0.00	95.35±0.00
DMCF-DDI-SSC	88.48±0.00	94.71±0.00	92.97±0.00	93.34±0.00
DMCF-DDI-SSC*	92.39±0.00	96.99±0.00	95.84±0.00	96.00±0.00
DMCF-DDI	91.95±0.00	96.83±0.00	95.66±0.00	95.99±0.00

**Table S10: Experimental results of DMCF-DDI and SOTA methods in Task A.**

	ACC	AUC	AUPR	MAP
DMCF-DDI-SC*	82.55±0.05	89.99±0.04	88.22±0.05	88.24±0.11
DMCF-DDI-SC	84.51±0.00	92.70±0.00	91.46±0.00	92.91±0.00
DMCF-DDI-ASC*	86.72±0.02	94.99±0.00	93.69±0.00	95.35±0.00
DMCF-DDI-SSC	81.79±0.03	89.87±0.04	88.12±0.04	89.23±0.18
DMCF-DDI-SSC*	84.52±0.05	93.74±0.01	92.48±0.01	94.71±0.01
DMCF-DDI	86.80±0.03	95.00±0.00	93.70±0.01	95.44±0.00

**Table S11: Experimental results of DMCF-DDI and SOTA methods in Task B.**

	ACC	AUC	AUPR	MAP
DMCF-DDI-SC*	86.04 $\pm$ 0.01	93.17 $\pm$ 0.00	91.40 $\pm$ 0.00	89.95 $\pm$ 0.00
DMCF-DDI-SC	88.40 $\pm$ 0.02	95.05 $\pm$ 0.00	93.75 $\pm$ 0.00	92.71 $\pm$ 0.00
DMCF-DDI-ASC*	89.78 $\pm$ 0.01	95.81 $\pm$ 0.00	94.46 $\pm$ 0.00	93.37 $\pm$ 0.00
DMCF-DDI-SSC	86.41 $\pm$ 0.01	93.42 $\pm$ 0.00	91.71 $\pm$ 0.00	90.46 $\pm$ 0.00
DMCF-DDI-SSC*	90.28 $\pm$ 0.01	96.01 $\pm$ 0.00	94.76 $\pm$ 0.00	93.93 $\pm$ 0.00
DMCF-DDI	90.45 $\pm$ 0.01	96.00 $\pm$ 0.00	94.77 $\pm$ 0.00	93.94 $\pm$ 0.00

## 5. Experiment results of case study for six diseases.

**Table S12: Experimental results on Task B under Alzheimer Disease scenario.**

Alzheimer Disease	ACC	AUC	AUPR	AP@50
MDNN	85.99	94.04	92.62	93.36
MUFFIN	87.23	94.38	92.86	92.64
DMMHA-DDI	89.82	95.69	94.24	94.53
LR-GNN	89.16	95.03	93.58	93.36
Quate	88.66	95.02	93.3	93.38
DMCF-DDI	90.31	95.99	94.69	95.01

**Table S13: Experimental results on Task B under Albuminuria scenario.**

Albuminuria	ACC	AUC	AUPR	AP@50
MDNN	84.46	92.52	90.94	91.36
MUFFIN	86.08	93.45	91.66	90.66
DMMHA-DDI	89.23	95.19	93.47	92.86
LR-GNN	88.42	94.48	92.83	91.84
Quate	86.88	93.76	91.71	90.92
DMCF-DDI	89.44	95.32	93.85	93.68

**Table S14: Experimental results on Task B under Astrocytoma scenario.**

Astrocytoma	ACC	AUC	AUPR	AP@50
MDNN	85.39	92.85	91.35	90.14
MUFFIN	85.16	92.74	91.12	89.39
DMMHA-DDI	89.93	95.63	94.14	94.04
LR-GNN	87.36	93.84	92.19	90.43
Quate	86.65	93.78	91.52	88.72
DMCF-DDI	89.48	95.1	93.58	92.48

**Table S15: Experimental results on Task B under Anaphylaxis scenario.**

Anaphylaxis	ACC	AUC	AUPR	AP@50
MDNN	83.20	90.88	89.13	86.00
MUFFIN	84.91	92.58	90.68	87.85
DMMHA-DDI	88.98	95	93.31	92.57
LR-GNN	87.10	93.67	91.91	89.21
Quate	85.76	92.71	90.36	86.74
DMCF-DDI	89.09	94.49	92.60	90.00

**Table S16: Experimental results on Task B under Prostate Cancer scenario.**

Prostate Cancer	ACC	AUC	AUPR	AP@50
MDNN	83.84	92.61	91.09	95.03
MUFFIN	87.27	94.38	92.97	93.81
DMMHA-DDI	88.03	94.58	92.84	91.24
LR-GNN	89.34	95.11	93.68	93.82
Quate	88.48	95.05	93.5	94.2
DMCF-DDI	89.84	95.61	94.3	95.23

**Table S17: Experimental results on Task B under Pneumonia scenario.**

Pneumonia	ACC	AUC	AUPR	AP@50
MDNN	85.61	93.49	91.94	92.32
MUFFIN	87.53	94.56	93.05	93.37
DMMHA-DDI	87.63	94.07	92.15	89.19

LR-GNN	88.54	94.53	92.95	92.41
Quate	87.92	94.54	92.77	92.31
DMCF-DDI	89.09	94.98	93.46	93.53

The std of All results are zero.

## 6. Experiment results of multi-class dataset.

### Materials and methods

#### Dataset

The multi-class dataset of DDIs is collected by Deng et al.[1] from DrugBank [2], which contains 572 drugs with 74 528 DDIs, which are associated with 65 types of side effects. For the sake of analysis, Deng et al.[1] removed the rare events and selected the events that had more than ten DDIs.

For this dataset, the first modality feature is a drug molecular graph that is constructed by rdkit according to its SMILES string. The second modality feature is heterogeneous similarity information which computes the similarity matrix between drugs based on various features (i.e. chemical substructures, targets, and enzymes), and uses PCA to extract key components of the similarity matrix.

#### Method

##### *GCN-encoder*

For drug molecular graph modality, each drug has a molecular graph  $g_i = (V, E)$ , that  $V$  represent atoms, and  $E$  represent chemical bonds. We use a GIN model pre-trained with supervised learning and attribute masking to generate the atoms representations of molecular graph  $g_i$ . Then, the graph-based drug representation can be acquired by the average pooling operation. The final drug molecular graph representation is a 300-dim embedding vector.

We first adopt a fully-connected layer to transfer the pre-trained drug molecular graph representation into a 256-dim embedding vector:  $H_{mg}^0 = X_{mg}W_{mg}^0$ , where  $W_{mg}^0$  is the weight matrix. Then, RGCN is used to update drug features with drug-drug adjacent relations:

$$H_{mg}^l = RGCN_{mg}(\{A^{dd}\}, H_{mg}^{l-1}, \{H_{mg}^{l-1}\}). \quad (12)$$

To simplify notation, we use  $H_{mg}$  to represent the output of the last layer of  $RGCN_{mg}$ .

For heterogeneous similarity information modality, we select three features of drugs: chemical substructures, targets, and enzymes. Each feature corresponds to a set of descriptors and is a binary feature vector, whose values (1 or 0) indicate the presence or absence of the descriptors. For instance, there are 1162 targets associated with drugs, and the target feature can encode a drug into an 1162-dimensional bit vector. We calculate the pairwise drug-drug similarity matrix from each feature using the Jaccard similarity measure and use PCA to extract  $n$  key components of the similarity matrix. Then, we concatenate the three extracted features as the initial heterogeneous similarity information (HSI) modality feature of drugs  $X_{hsi}$ . A multi-layer fully-connected network (MLP) is constructed and takes the initial HIS features as input to improve the expression ability of the HSI modality feature:

$$H_{hsi} = MLP_{hsi}(X_{hsi}). \quad (13)$$

### Complex-valued fusion

Asymmetric skip connection constructs the DPRs of positive-inverse order, i.e.  $H_{ij}, H_{ji}$ , thereby  $S_{mg}, S_{hsi}, H_{mg}$  and  $H_{hsi}$ :

$$\begin{aligned} Re(H_{ij}) &= LN(H_{mg}[i] + S_{mg}[j]) = MG_{i,j}^{asc}, \\ Im(H_{ij}) &= LN(H_{hsi}[i] + S_{hsi}[j]) = HSI_{i,j}^{asc}, \\ Re(H_{ji}) &= LN(H_{hsi}[j] + S_{hsi}[i]) = HSI_{j,i}^{asc}, \\ Im(H_{ji}) &= LN(H_{mg}[j] + S_{mg}[i]) = MG_{j,i}^{asc}, \end{aligned} \quad (14)$$

where  $S_{mg} = \sigma(\sigma(H_{mg}^0)W_{mg}^{cf})$ ,  $S_{hsi} = \sigma(\sigma(H_{hsi}^0)W_{hsi}^{cf})$ .  $W_{mg}^{cf}$  and  $W_{hsi}^{cf}$  are the trainable weight matrices,  $\sigma$  is the non-linear activation function and  $H_{hsi}^0$  is the output of the first layer of  $MLP_{hsi}$ . The element-wise multiplication is used to integrate complex-valued DPRs of positive-inverse order:

$$Z_{ij}^{asc} = H_{ij} \odot H_{ji}. \quad (15)$$

### Loss

We calculate the probability by Hermitian product in complex space:

$$P_{ij} = \text{softmax} \left( O \left( \begin{bmatrix} Re(WZ_{ij}^{asc}) \\ Im(WZ_{ij}^{asc}) \end{bmatrix} \right) \right), \quad (16)$$

and where  $W = (Re(W), Im(W))$  is the complex matrix,  $O()$  can be selected as the  $SUM()$  and  $Re()$  function.  $SUM()$  adds the real and imaginary parts of a complex number.  $Re()$  extracts the real part of a complex number.

## Experiment setting

According to DDIMDL, we apply 5-fold cross-validation (5-CV) to DDIs and split all DDIs into five folds. We select one of five folds as the test set in turn, the other four folds are used as the training set.

For the multi-class classification task, accuracy (ACC), micro AUPR, micro AUC, and macro F1 score are used as the evaluation metrics. The parameters of DMCF-DDI and compared methods are shown in Table S11.

**Table S18: The parameters of DMCF-DDI and compared methods.**

Methods	Parameters
DDIMDL	The number of principal components $n = 200$ . For the three-layer MLP, hidden dimension = {512, 256, 65}. The dropout rate = 0.3, the learning rate = 1e-2, the epoch = 100
MUFFIN	The feature dimension of the molecular graph = 300, and the number of principal components $n = 100$ . For the three-layer convolution neural network, the kernel size {64, 32, 16}. For the two-layer fully-connected network classifier, hidden dimension = {256, 128}. The learning rate = 1e-3, and the epoch = 300.
DMCF-DDI	The number of principal components $n = 200$ .



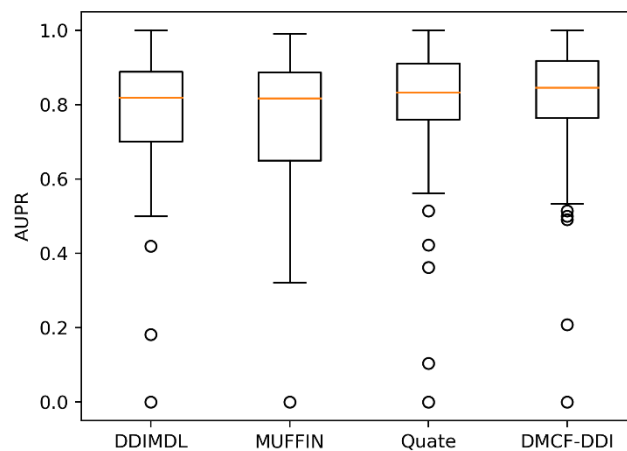
	For $RGCN_{mg}$ , the number of layers=2, the hidden dimension={256, 64}. For $MLP_{hsi}$ , the number of layers=3, the hidden dimension={256, 64, 64}. The dropout rate=0.3, the learning rate=5e-3, the epoch=500.
Quate	The encoders of Quate are the same as that of DMCF-DDI. For a fully connected network, hidden dimension=256. The learning rate=5e-3, and the epoch=500.

## Results

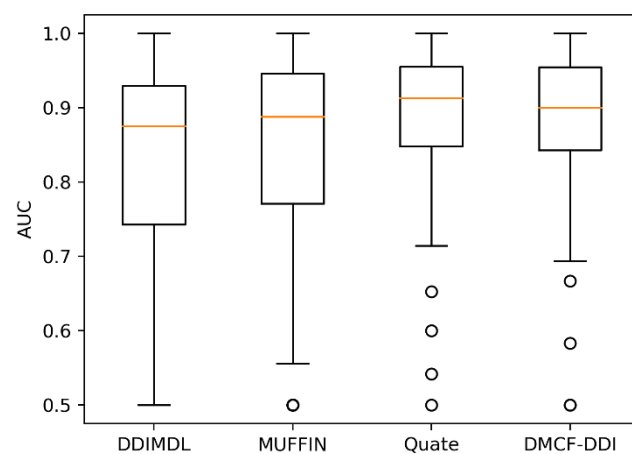
**Table S19: The 5-fold prediction performance of multi-class dataset.**

	ACC	AUPR- micro	AUC- micro	f1- macro
DDIMDL	86.50	92.01	99.71	72.08
MUFFIN	88.71	93.30	99.66	68.08
Quate	89.85	95.35	99.84	78.09
DMCF	<b>91.23</b>	<b>96.10</b>	99.83	<b>78.91</b>

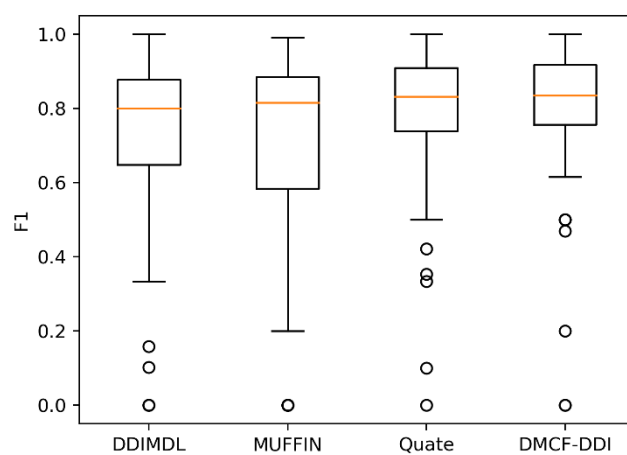
Table S19 shows that DMCF-DDI achieves the best performance on three metrics, ACC, AUPR, and F1. To further analyze the prediction performances of each class, Fig. S1-3 shows the AUPR, AUC, and F1 scores of different methods for 65 side effects.



**Fig. S 4 The boxplot of AUPR of different methods for 65 side effects.**



**Fig. S 5 The boxplot of AUC of different methods for 65 side effects.**



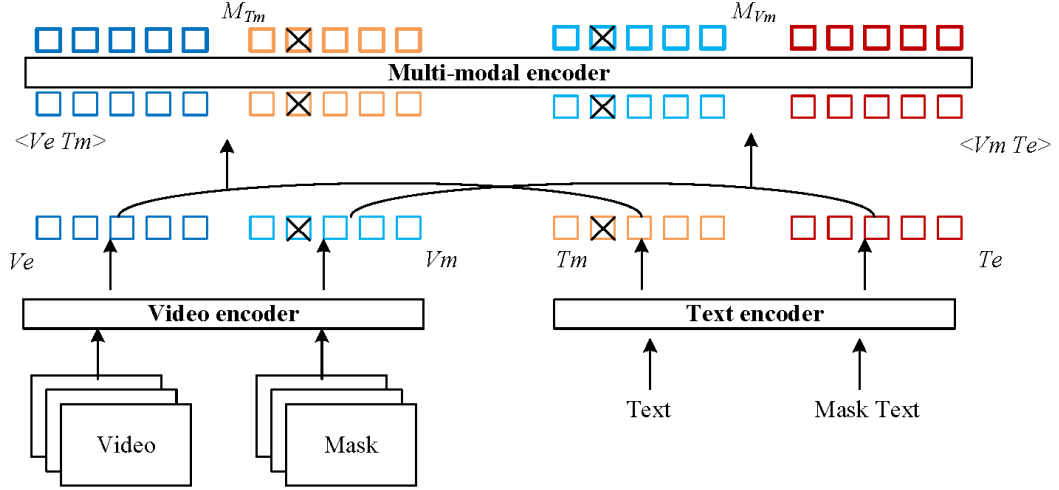
**Fig. S 6 The boxplot of the F1 score of different methods for 65 side effects.**

For each type of side effect, the AUPR and F1 score of DMCF-DDI are generally higher than other methods. The AUC of Quate is statistically better than DMCF-DDI, which shows the potential of quaternion embedding for predicting the side effects of DDIs.

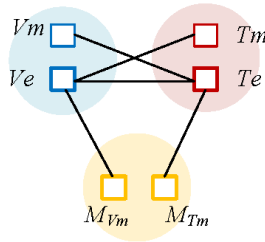
## 7. Experiment results of contrastive learning.

Cross-modal alignment, which is usually used for text-video (text-image) retrieval, aims to learn projection functions  $f()$  and  $g()$  that project videos and texts into a common embedding space where the similarity between the video and text with consistent semantic information is maximized and otherwise minimized.

Clover[3] leverages multi-modal representations as intermediate modalities between video and text, providing richer context information and serving as anchors for cross-modality alignment. The authors propose the Tri-Modal Alignment (TMA) approach to enhance Video-Language alignment and fusion by treating it as an auxiliary objective. TMA ensures that the fused representation maintains semantic consistency with uni-modal representations while minimizing inconsistencies. We briefly explain the architecture of Clover here, which is shown in Fig. S7.



(a) Model Architecture



(b) Pre-training Tasks

**Fig. S 7 Overview of Clover.**  $Ve$  is the video encoding feature;  $Vm$  is the masked video encoding feature;  $Te$  is the text encoding feature;  $Tm$  is the masked text encoding feature.  $\langle Ve Tm \rangle$  and  $\langle Vm Te \rangle$  are asymmetric combinations.

The analogy between the two designs of Clover and our method can verify the advancement of our method from two sides:

1. Compared with the previous cross-modal alignment method, Clover adds the fused multi-modal representation as the third modality. As an intermediate modality between video and text, the

multi-modal representations are good anchors for cross-modality alignment. Meanwhile, keeping the fused representation closer to the uni-modal representation containing consistent semantic information and away from the inconsistent one will enhance the learning of semantic information in the fused modality. We also achieve alignment on the fused multi-modal representation and improve the prediction results. It indicated that the exchange operation between real and imaginary parts is necessary. We borrow the idea of contrastive learning to make  $Re(H_{ij})Re(H_{ji})$  and  $Im(H_{ij})Im(H_{ji})$  consistent:

$$Z_{ij}^{asc} = H_{ij} \odot H_{ji}$$

$$p_{ij}^r = \text{sigmoid}(w_r^T Im(Z_{ij}^{asc}) + b_r) \quad (17)$$

$$loss_{CM}^+ = \frac{1}{|Pos|} \sum_{e_{ij} \in Pos} ||Re(Z_{ij}^{asc})||_2 + \log(p_{ij}^r)$$

$$loss_{CM}^- = \frac{1}{|Neg|} \sum_{e_{ij} \in Neg} ||Re(Z_{ij}^{asc})||_2 + \log(1 - p_{ij}^r) \quad (18)$$

$$loss = loss_{CM}^+ + loss_{CM}^-$$

The results are presented in Table S20. We observe that forcing the real part to be zero can't improve results. In other words, their difference is also important information to be used for classification without wasting real part features.

2. Clover uses the mask operation to construct two features of a video (text), and combines the text (video) mask feature with the video (text) original feature in an asymmetric manner, expanding the representation space of the sample combination. This is consistent with the underlying principle of asymmetric skip connection (ASC). We utilize ASC to construct two features of a drug pair in a single modality (i.e.  $H_{ij} = (Re(H_{ij}), Im(H_{ji}))$ ), which makes full use of the representation space.

**Table S20: Experimental results of DMCF-DDI with contrastive learning.**

Regular				
	ACC	AUC	AUPR	AP
w/o contrastive learning	87.14±0.00	94.05±0.00	92.12±0.00	92.13±0.00
with contrastive learning	88.23±0.00	94.56±0.00	92.80±0.00	92.97±0.00
DMCF-DDI	91.95±0.00	96.83±0.00	95.66±0.00	95.99±0.00
Task A				
	ACC	AUC	AUPR	AP
w/o contrastive learning	81.31±0.03	89.29±0.01	87.34±0.01	86.20±0.01
with contrastive learning	82.09±0.02	88.38±0.01	86.57±0.01	84.03±0.01
DMCF-DDI	86.80±0.03	95.00±0.00	93.70±0.01	95.44±0.00
Task B				
	ACC	AUC	AUPR	AP
w/o contrastive learning	85.36±0.02	93.55±0.00	91.78±0.00	90.48±0.01
with contrastive learning	85.89±0.02	93.18±0.00	91.32±0.00	89.71±0.01
DMCF-DDI	90.45±0.01	96.00±0.00	94.77±0.00	93.94±0.00

## 8. Experiment results of PPI dataset.

To prove the generalizability of DMCF-DDI, we apply the complex-valued fusion operator to the PPI dataset that is extracted from the drug knowledge graph. The dataset contains 708188 PPIs with 18138 protein nodes. The first modality of proteins is the pre-trained sequence embedding from ProtBert. We use PCA to reduce the sequence embedding dimension from 1024 to 300. The second modality of proteins is the protein knowledge graph which contains the protein-drug, and drug-drug associations.

We evaluate the performance of models through regular and inductive situations. In the regular situations, we divide the PPIs into training, validation, and test sets with a ratio of 80%: 5%: 15%. The results are shown in the Table S21. To evaluate the performance of DMCF-PPI under a few training samples, we conduct experiments with a splitting ratio of 45%: 5%: 50%. The results under a low training ratio are shown in the Table S22. Additionally, to evaluate the performance of predicting interactions of unknown proteins, we randomly split proteins into five subsets, and used four of them as training protein sets and the remaining as test protein sets. Thus, two interactions are obtained: Train-Test and Test-Test. The difference between transductive and inductive learning is whether the training dataset contains the interactions of test proteins. The inductive experiment results are shown in the Table S23.

We use AUC, AP, and Hit@500 to evaluate the performance of models. Specifically, AUC is the area under the curve. Letting TP (FP) be the number of true (false) positive links,  $AP = TP / (TP + FP)$ . Hit@500 ranks each positive edge in the validation/test set against randomly sampled negative edges and counts the ratio of positive edges that are ranked at 500th place or above.

**Table S21: Experimental results of PPI under ratio 80%: 5%: 15%.**

Method	AUC	AP	Hit@500
MUFFIN	0.9095 $\pm$ 0.0010	0.9068 $\pm$ 0.0008	0.1917 $\pm$ 0.0057
LR-GNN	<b>0.9476 <math>\pm</math> 0.0004</b>	0.9510 $\pm$ 0.0006	0.4246 $\pm$ 0.0054
Quate	0.9441 $\pm$ 0.0008	0.9490 $\pm$ 0.0005	<u>0.4360 <math>\pm</math> 0.0129</u>
DMCF-PPI	<u>0.9457 <math>\pm</math> 0.0003</u>	<b>0.9511 <math>\pm</math> 0.0003</b>	<b>0.4508 <math>\pm</math> 0.0063</b>

**Table S22: Experimental results of PPI under ratio 45%: 5%: 50%.**

Method	AUC	AP	Hit@500
MUFFIN	0.8891 $\pm$ 0.0018	0.8849 $\pm$ 0.0020	0.0689 $\pm$ 0.0072
LR-GNN	<b>0.9358 <math>\pm</math> 0.0027</b>	<u>0.9377 <math>\pm</math> 0.0033</u>	0.1773 $\pm$ 0.0156
Quate	0.9318 $\pm$ 0.0010	0.9366 $\pm$ 0.0011	<u>0.1997 <math>\pm</math> 0.0088</u>
DMCF-PPI	<u>0.9323 <math>\pm</math> 0.0004</u>	<b>0.9381 <math>\pm</math> 0.0005</b>	<b>0.2179 <math>\pm</math> 0.0049</b>

**Table S23: Experimental results of inductive experiments.**

Train-Test			
Method	AUC	AP	Hit@500
MUFFIN	<u>0.8062 <math>\pm</math> 0.0206</u>	<u>0.8040 <math>\pm</math> 0.0169</u>	<u>0.2165 <math>\pm</math> 0.0197</u>
LR-GNN	0.7607 $\pm$ 0.0118	0.7504 $\pm$ 0.0221	0.1386 $\pm$ 0.0333
Quate	0.7847 $\pm$ 0.0305	0.7898 $\pm$ 0.0318	0.2123 $\pm$ 0.0409
DMCF-PPI	<b>0.8246 <math>\pm</math> 0.0158</b>	<b>0.8306 <math>\pm</math> 0.0147</b>	<b>0.2785 <math>\pm</math> 0.0283</b>
Test-Test			
Method	AUC	AP	Hit@500
MUFFIN	0.6465 $\pm$ 0.0192	0.6322 $\pm$ 0.0140	0.3242 $\pm$ 0.0395

LR-GNN	$0.6526 \pm 0.0076$	$0.6413 \pm 0.0173$	$0.3367 \pm 0.0207$
Quate	$0.6455 \pm 0.0148$	$0.6459 \pm 0.0178$	<b><math>0.3402 \pm 0.0107</math></b>
DMCF-PPI	<b><math>0.6537 \pm 0.0130</math></b>	<b><math>0.6516 \pm 0.0112</math></b>	$0.3319 \pm 0.0261$

The results show that DMCF-PPI has the best Hit@500 under different splitting ratios. The Hit@500 measures the significance of positive sample prediction results. In the inductive experiments, DMCF-PPI has the best performance under all metrics on Train-Test interactions, which indicates that the complex-valued fusion operator is not limited to specific biological entities. When positive pairs have similar neighbors, the complex-valued fusion operator makes the real part be gathered in an imaginary axis to achieve feature alignment. Knowledge graphs can make the representations of negative pairs separate in the complex space, even if they have similar pre-trained sequence embedding.

## 9. Experiment results of multi-modality comparison.

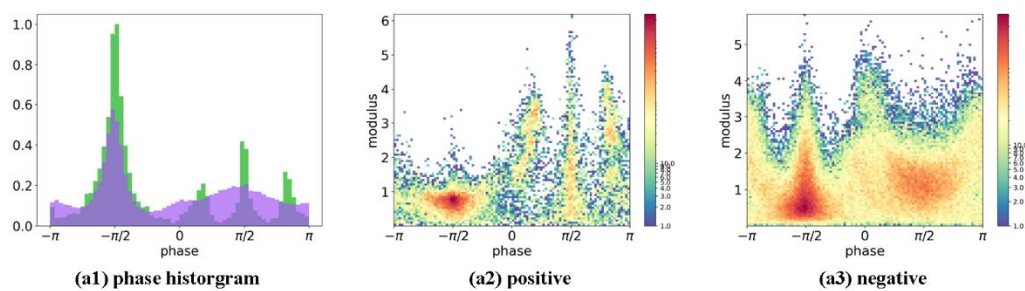
We collect the molecular properties of each drug from the Pubchem website which contains 17 items. Then, three modalities of information are obtained: fingerprint (FP), drug knowledge graph (DKG), and molecular property (MP). Any combination of two modalities gives us three combinations: FP-DKG, MP-DKG, and FP-MP. The results of the three combinations are shown in Table S24.

**Table S24: Experimental results of inductive experiments.**

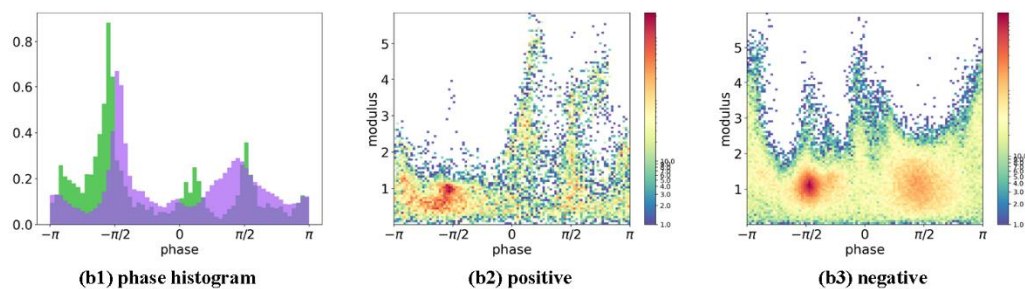
Regular				
Method	ACC	AUC	AUPR	AP@50
MP-KG	92.03±0.00	96.84±0.00	95.59±0.00	95.48±0.00
MP-FP	90.62±0.00	96.1±0.00	94.86±0.00	95.24±0.00
FP-KG	91.95±0.00	96.83±0.00	95.66±0.00	95.99±0.00
Task A				
Method	ACC	AUC	AUPR	AP@50
MP-KG	85.84±0.02	94.81±0.00	93.52±0.00	95.45±0.00
MP-FP	84.22±0.10	93.67±0.01	92.42±0.00	94.75±0.00
FP-KG	86.80±0.03	95.00±0.00	93.70±0.01	95.44±0.00
Task B				
Method	ACC	AUC	AUPR	AP@50
MP-KG	0.9010±0.01	0.9606±0.00	0.9483±0.00	0.9407±0.00
MP-FP	0.8816±0.00	0.9471±0.00	0.9338±0.00	0.9240±0.00
FP-KG	90.45±0.01	96.00±0.00	94.77±0.00	93.94±0.00

The results show that MP-KG and FP-KG have similar performance, but the performance of FP-MP is poor. According to the distribution of the DPR phase of DMCF-DDI, negative drug pairs also achieve alignment even with dissimilar neighbors. We suspect that FP and MP are highly similar between drugs. Thus, we calculate the ratio of common compound structures in each drug and plot the histogram distribution of ratios. We also calculate the MP similarity between 645 drugs and plot the histogram distribution of similarities. We collect all distribution graphs in Fig S8. Fig S8 shows that more than half of the structures of most drugs are common compound structures. The similarity of the MP between drugs is mainly distributed between 0.6 and 1.0. These findings suggest that while MP and FP of drugs exhibit high similarity, the KG modality plays a crucial role in prediction tasks. Consequently, the poor performance of the FP-MP combination could be attributed to this high degree of similarity between MP and FP, emphasizing the importance of the KG modality in our predictions.

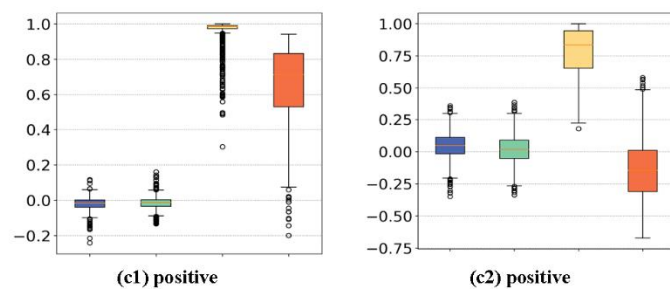
(a) The histogram distributions of MP-KG modality.



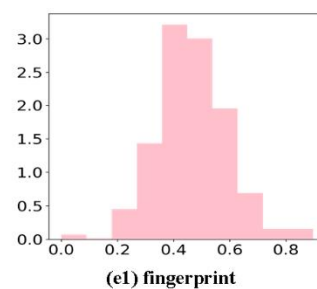
(b) The histogram distributions of FP-MP modality.



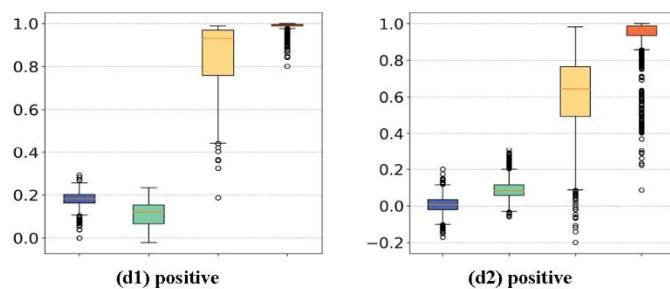
(c) Similarity between parts of complex features for MP-KG modality.



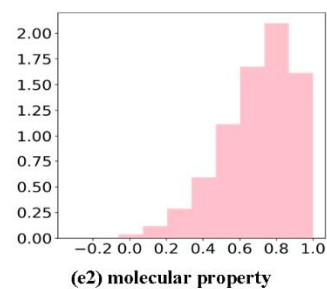
(e) Similarity in fingerprint and molecular property



(d) Similarity between parts of complex features for FP-MP modality.



(e2) molecular property



**Fig. S 8 The histogram distributions of positive and negative pairs representation for MP-KG and FP-MP modality combinations.**



## 10. Discussion about databases.

**STITCH (Drug-Protein):** STITCH is a comprehensive database that integrates information on chemical-protein interactions and chemical-chemical interactions. It provides a valuable resource for exploring the interactions between small molecules (such as drugs or metabolites) and proteins, as well as the relationships between different chemicals. STITCH offers curated data on binding affinities, functional associations, and predicted interactions, facilitating the investigation of drug-target interactions and the identification of potential drug candidates.

**STRING (Protein-Protein):** STRING is a widely used database that provides information on protein-protein interactions (PPIs) and functional associations. It integrates data from various sources, including experimental studies, computational predictions, and curated databases, to generate comprehensive interaction networks. STRING offers valuable insights into the molecular interactions underlying cellular processes, disease mechanisms, and drug action. It provides users with access to protein interaction networks, functional enrichment analysis, and visualization tools, enabling the exploration of complex biological systems and the identification of novel therapeutic targets.

**PubChem (FingerPrint):** PubChem is a large-scale repository of chemical information maintained by the National Center for Biotechnology Information (NCBI). It provides access to a vast collection of chemical compounds, their properties, and biological activities. PubChem offers various datasets, including chemical structures, bioassay results, and toxicity information, making it a valuable resource for drug discovery and development. Researchers can utilize PubChem to explore chemical structures, identify potential drug candidates, and investigate the pharmacological properties of compounds.

**TwoSIDES (Side effects of DDIs):** TwoSIDES is a comprehensive database that contains information on drug-drug interactions (DDIs) and their associated side effects. It integrates data from multiple sources, including pharmacovigilance databases, clinical trials, and literature reports, to provide comprehensive coverage of known DDIs. TwoSIDES offers valuable insights into the adverse effects resulting from the concomitant use of different medications, facilitating the identification and management of potentially harmful drug combinations. Researchers can leverage TwoSIDES to explore DDI networks, predict potential interactions, and assess the clinical significance of drug combinations.

These databases provided valuable resources for investigating drug action mechanisms, identifying potential drug targets, and predicting drug interactions, thereby advancing our understanding of complex biological systems and facilitating drug discovery and development efforts.

## 11. Supplementary discussion about using of ComplexE and RotatE for link prediction in knowledge graph.

ComplexE and RotatE are two successful models for link prediction in knowledge graphs due to their ability to capture intricate relationships between entities by leveraging complex numbers. Here's a detailed explanation of why these models are effective:

### ComplexE

ComplexE (Complex Embeddings) is a knowledge graph embedding model that represents entities and relations using complex vectors.

#### 1. Complex Number Representation:

Entities and Relations as Complex Vectors: Entities and relations are represented in a complex vector space, where each element of the vector is a complex number. This allows for richer representations as complex numbers can encode both magnitude and phase information.

#### 2. Modeling Symmetric and Asymmetric Relations:

Symmetric Relations: For symmetric relations (e.g., "sibling of"), ComplexE can effectively model them because the complex conjugate operation naturally captures symmetry.

Asymmetric Relations: For asymmetric relations (e.g., "parent of"), the phase information (angle) in complex numbers helps in distinguishing directionality. The model can capture the asymmetry by learning appropriate phase shifts.

#### 3. Efficient and Scalable Training:

ComplexE can be trained efficiently using gradient-based optimization techniques. The use of complex arithmetic does not significantly increase computational complexity compared to real-valued models.

### RotatE

RotatE (Rotation Embeddings) is another knowledge graph embedding model that represents relations as rotations in the complex plane.

#### 1. Relation as Rotations:

Rotational Transformation: In RotatE, relations are modeled as rotations in the complex plane. If  $e_1$  and  $e_2$  are entity vectors and  $r$  is a relation vector, the model learns that applying the rotation  $r$  to  $e_1$  should result in  $e_2$ . Mathematically, this is represented as  $e_2 \approx e_1 \circ r$ , where  $\circ$  denotes element-wise multiplication.

Phase Shifts: The rotation  $r$  is a unit complex number, representing a phase shift. This makes

it easy to model cyclic and transitive relations, as rotations inherently capture these properties.

## 2. Handling Complex Relations:

**Complex Relationships:** RotatE effectively handles various types of relationships including one-to-one, one-to-many, many-to-one, and many-to-many. The phase shift introduced by rotations can naturally encode such diverse relational patterns.

**Inverse Relations:** RotatE can model inverse relations efficiently. If  $r$  is the relation for "parent of," the inverse relation "child of" can be modeled by the inverse rotation  $r^{-1}$ .

## 3. Geometric Interpretability:

**Geometric Intuition:** The geometric nature of rotations in the complex plane provides an intuitive understanding of how relations are modeled. Rotations are a simple yet powerful transformation that makes the learning process more interpretable.

## Common Strengths of ComplexE and RotatE

### 1. Expressive Power:

Both models leverage the expressive power of complex numbers to capture rich relational patterns. Complex numbers can encode more information per dimension compared to real numbers, allowing for more compact and informative embeddings.

### 2. Handling of Symmetry and Asymmetry:

The ability to naturally handle both symmetric and asymmetric relations is a key strength. Symmetric relations benefit from the conjugate properties of complex numbers (ComplexE), while asymmetric relations benefit from phase shifts (RotatE).

### 3. Scalability:

Both models are designed to be scalable, allowing them to be applied to large-scale knowledge graphs. They use efficient training algorithms and can leverage hardware acceleration.

## 12. Visualization of Quate.

Unit quaternions provide a convenient mathematical notation for representing spatial orientations and rotations of elements in three-dimensional space.

Given a quaternion,  $q = q_r + q_i i + q_j j + q_k k$ , the vector  $a$  in 3d space and  $\theta$  can be found from the following equations:

$$(a_x, a_y, a_z) = \frac{(q_i, q_j, q_k)}{\left(\sqrt{(q_i^2 + q_j^2 + q_k^2)}\right)}$$

$$rotation = 2arctan2\left(\sqrt{(q_i^2 + q_j^2 + q_k^2)}, q_r\right)$$

Vector  $a$  is transformed from cartesian to spherical coordinates:

$$r = \sqrt{(a_x^2 + a_y^2 + a_z^2)}$$

$$\theta = arctan(a_y/a_x)$$

$$\phi = arccos(z/\sqrt{(a_x^2 + a_y^2 + a_z^2)})$$

We can visualize the positive and negative samples in 3d space with rotation,  $\theta$ , and  $\phi$ , since  $r = 1$ .

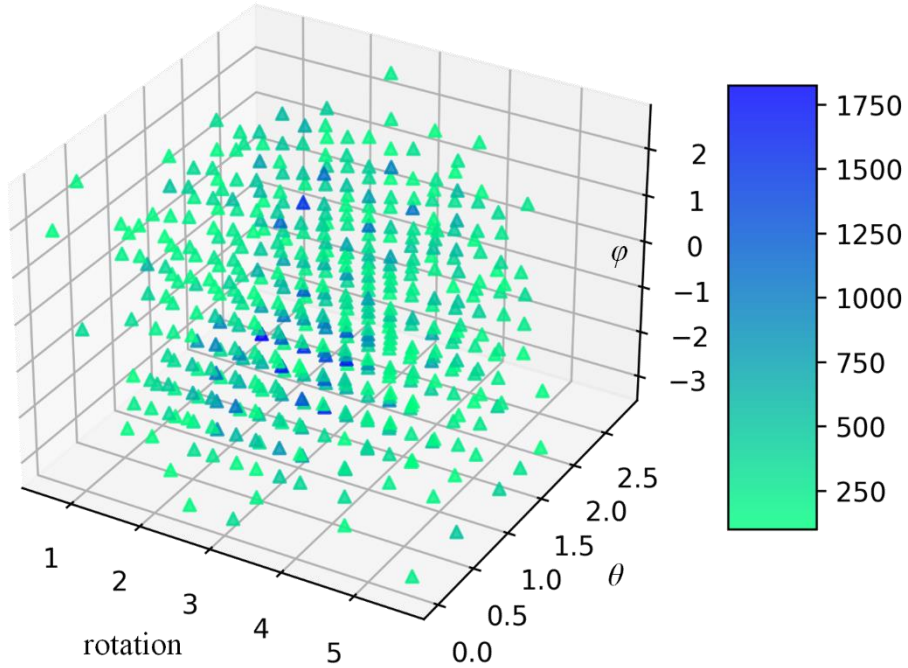


Fig. S 9 The distribution of positive samples in 3d space.

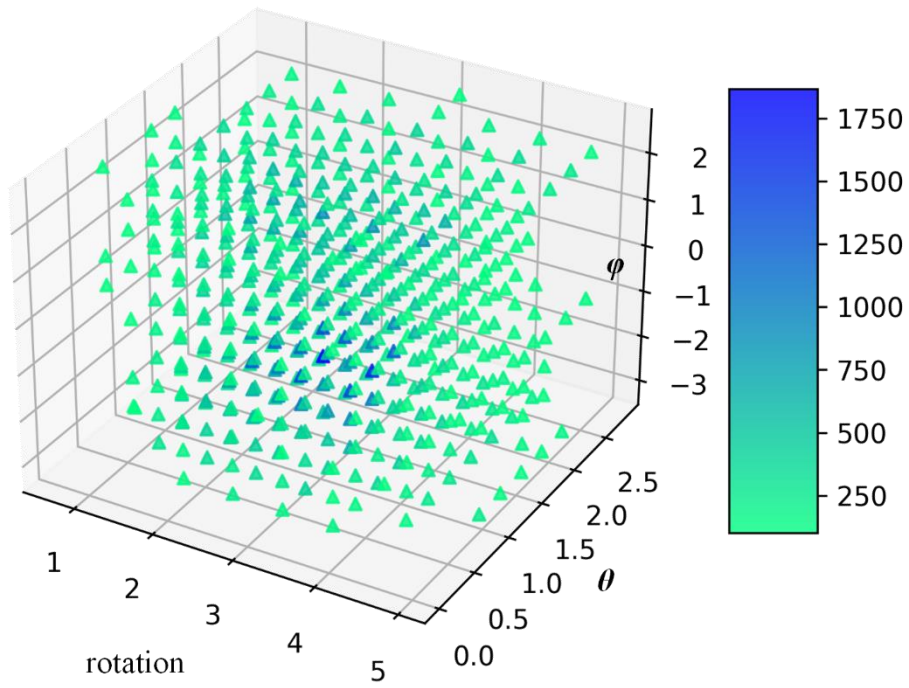


Fig. S 10 The distribution of negative samples in 3d space.

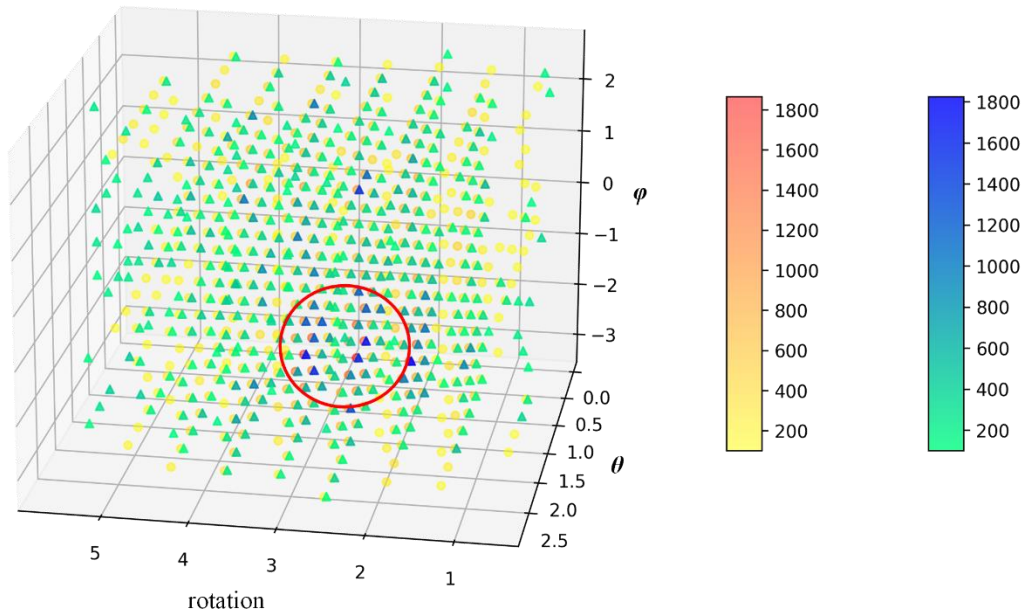


Fig. S 11 The distribution of positive and negative samples in 3d space.

Fig. shows that it is difficult to distinguish the distributions of positive and negative samples. The classification pattern learned by Quate is not clear in visualization as well. More theoretical derivation is required to improve the way of visualization.

### 13. New Interactions predicted by DMCF-DDI.

In this section, we confirm the new interactions predicted by DMCF-DDI on Pubmed website. In case study section, the specific side effects associated three diseases are: {Prostate Cancer: dizziness, asthenia, decreased libido, nasopharyngitis, nasal congestion}, {Pneumonia: aortic aneurysm, acute kidney failure, diarrhea, nausea, headache, tendinitis}, {Alzheimer Disease: nausea, diarrhea, fatigue, dizziness, itch, confusion, edema}. We rank the negative samples by prediction probabilities under involved side effects. Top 10 predictions are selected and confirmed on Pubmed website. We only consider the relations <drug A, drug B, side effect>, because constraints about disease aren't introduced into training of DMCF-DDI.

**Tabel S25: Confirmed DDIs in top 10 predictions under side effects during treatment for Prostate Cancer.**

Drug A	Drug B	Side Effect	Rank	Evidence
Amoxicillin	Omeprazole	dizziness	1	PMID: 29119724
Amoxicillin	Omeprazole	asthenia	3	PMID: 29119724
Ofloxacin	Rabeprazole	dizziness	5	PMID: 23429422
Ofloxacin	Amoxicillin	asthenia	7	PMID: 22372560
Betamethasone	Amoxicillin	dizziness	10	PMID: 17523719

**Tabel S26: Confirmed DDIs in top 10 predictions under side effects during treatment for Pneumonia.**

Drug A	Drug B	Side Effect	Rank	Evidence
Ofloxacin	Rabeprazole	nausea	1	PMID: 23429422
Betamethasone	Methylprednisolone	nausea	3	PMID: 29859017 PMID: 35184049
Betamethasone	Methylprednisolone	diarrhea	4	PMID: 9884551
Ofloxacin	Rabeprazole	diarrhea	5	PMID: 23429422
Metformin	Ofloxacin	acute kidney failure	7	PMID: 24940713
Betamethasone	Methylprednisolone	tendinitis	10	PMID: 12431045

**Tabel S27: Confirmed DDIs in top 10 predictions under side effects during treatment for Alzheimer Disease.**

Drug A	Drug B	Side Effect	Rank	Evidence
DL-Thyroxine	Metformin	diarrhea	1	PMID: 34907435, PMID: 27987248
DL-Thyroxine	Metformin	fatigue	2	PMID: 35265033, PMID: 31485975
Enoxaparin	Omeprazole	edema	3	PMID:

				33942541
Baclofen	Omeprazole	fatigue	5	PMID: 23885151
Baclofen	Omeprazole	diarrhea	7	PMID: 23885151
Baclofen	Omeprazole	confusion	8	PMID: 28018156

## 14. How dual-modality improve the single-modal FP performance.

We evaluate the performance of the complex-valued single-modal models in Task B across Prostate Cancer, Pneumonia, and Alzheimer Disease. FP-DDI and KG-DDI are constructed to use a single modality to predict the side effects of DDI. The way FP-DDI constructing drug complex-valued representation is as follows:

$$\begin{aligned} Re(z_i) &= LN(S_{fp}[i]) \\ Im(z_i) &= LN(H_{fp}[i]) \\ Re(z_j) &= LN(H_{fp}[j]) \\ Im(z_j) &= LN(S_{fp}[j]) \\ Z_{ij}^{FP} &= z_i \odot z_j \end{aligned}$$

The way KG-DDI constructing drug complex-valued representation is as follows:

$$\begin{aligned} Re(z_i) &= LN(S_{kg}[i]) \\ Im(z_i) &= LN(H_{kg}[i]) \\ Re(z_j) &= LN(H_{kg}[j]) \\ Im(z_j) &= LN(S_{kg}[j]) \\ Z_{ij}^{KG} &= z_i \odot z_j \end{aligned}$$

Results are shown in Tabel S28:

**Tabel S28: Experimental results for FP-DDI, KG-DDI, and DMCF-DDI in Task B across Prostate Cancer, Pneumonia, and Alzheimer Disease.**

Prostate Cancer				
	ACC	AUC	AUPR	MAP@50
FP-DDI	88.67	94.77	93.44	94.60
KG-DDI	89.72	95.40	94.01	94.81
DMCF-DDI	90.30	95.96	95.96	95.96
Pneumonia				
	ACC	AUC	AUPR	MAP@50
FP-DDI	86.74	93.44	92.02	92.25
KG-DDI	88.80	94.84	93.20	92.96
DMCF-DDI	89.63	95.46	94.00	94.07
Alzheimer Disease				
	ACC	AUC	AUPR	MAP@50
FP-DDI	88.14	94.69	93.31	93.78
KG-DDI	89.48	95.22	93.71	93.28
DMCF-DDI	90.22	95.83	94.55	95.00

Results show that the performance of the dual-modal model is better than that of the single-modal model. The FP requires the KG feature to improve its prediction results. Therefore, we explore how dual-modality improves the performance of FP-DDI. For the prediction results of the FP-DDI



and KG-DDI, we collect the true and false predicted positive samples. First, we sort the prediction probability from large to small and select the top 2,000 true positive samples. Secondly, the top 2000 false positive samples are selected. Most of the predictive probability of these samples is around 0.499, which is near the classification hyperplane. Finally, we select the bottom 2,000 false positive samples. Most of the predictive probability of these samples is between 0.1-0.2, that is, they are completely misclassified. We verify whether the DMCF-DDI can correct the false positive samples predicted by FP-DDI and KG-DDI. The results are as follows:

**Tabel S29: Experimental results for FP-DDI, KG-DDI, and DMCF-DDI in Task B across Prostate Cancer, Pneumonia, and Alzheimer Disease.**

	FP: Prob. $\rightarrow$ 0.5		FP: Prob. $\rightarrow$ 0.0	
	p-value	Correct Ratio	p-value	Correct Ratio
Prostate Cancer	1.85e-80	0.6535	4.08e-265	0.2365
Pneumonia	1.12e-218	0.7555	0	0.3995
Alzheimer Disease	3.44e-264	0.7770	1.66e-320	0.2920
	KG: Prob. $\rightarrow$ 0.5		KG: Prob. $\rightarrow$ 0.0	
	p-value	Correct Ratio	p-value	Correct Ratio
Prostate Cancer	1.19e-9	0.5455	1.02e-58	0.1800
Pneumonia	0.999	0.4290	1.13e-23	0.1545
Alzheimer Disease	9.63e-11	0.5435	1.82e-83	0.2355

In the Tabel S29, the p-value represents the significance that the predictive probability of DMCF-DDI is greater than that of FP-DDI/KG-DDI. The statistical results show that DMCF-DDI has a significant correction ability for false positive samples near the FP modal classification hyperplane. There is a limited correction performance for false positive samples that are completely mistaken by the single-modal model, but the overall prediction probability of DMCF-DD is significantly greater than that of single-modal models.

Taking Pneumonia as an example, we visualize the feature similarity of the corrected positive sample in the Top 2000(Prob.  $\rightarrow$  0.5):

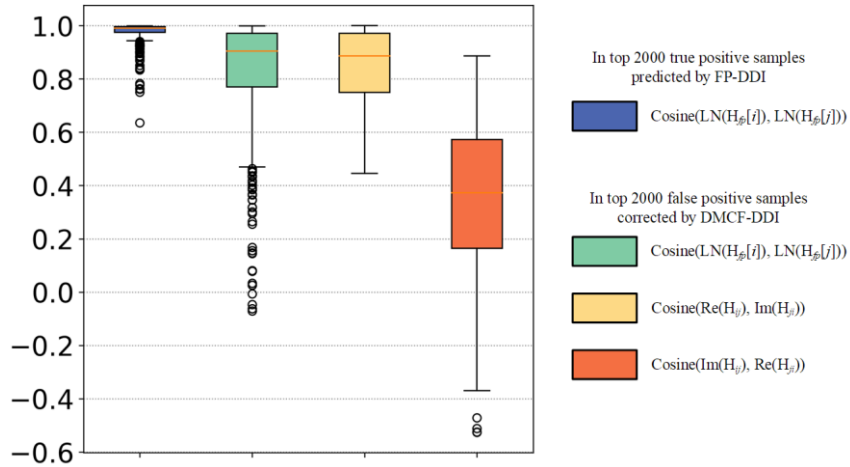


Fig. S 12 Distributions of cosine similarity between the real and imaginary parts of  $H_{fp}$  and  $H_{ij}$ .

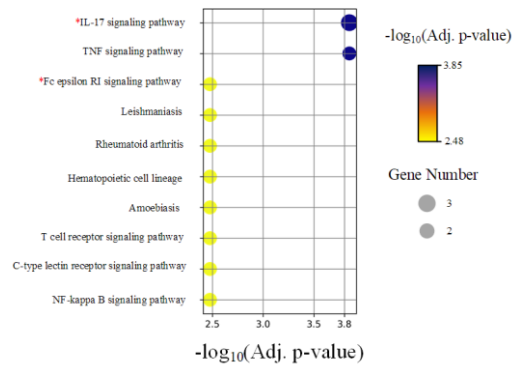


Fig. S 13 KEGG enrichment analysis of cardiac disease. IL-17 signaling pathway-PMID: 31630418; Fc epsilon RI signaling pathway-PMID: 34158869

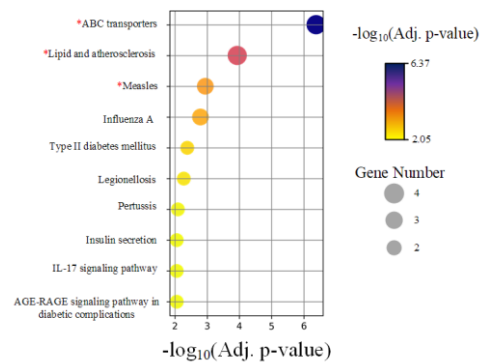


Fig. S 14 KEGG enrichment analysis of prostate cancer. ABC transporters-PMID: 25147980; Lipid and atherosclerosis-PMID: 18349064, PMID: 34738072; Measles-PMID: 19773744

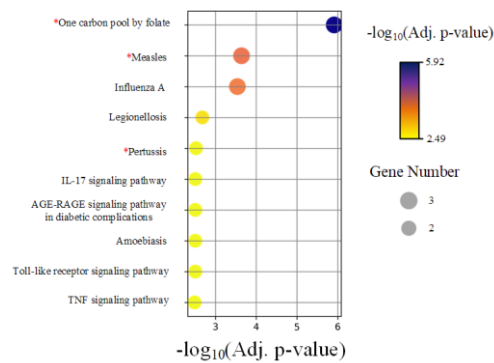


Fig. S 15 KEGG enrichment analysis of adenocarcinoma. One carbon pool by folate-PMID: 33968204; Measles-PMID: 27657109; Pertussis-PMID: 27071941

## Reference

- [1] Deng YF. et al. (2020) A multimodal deep learning framework for predicting drug–drug interaction events. *Bioinformatics*, 36(15): 4316–4322.
- [2] Knox C. et al. (2011) Drugbank 3.0: a comprehensive resource for ‘omics’ research on drugs. *Nucleic Acids Res.*, 39, D1035–41.
- [3] Huang, Jingjia, et al. "Clover: Towards a unified video-language alignment and fusion model." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.
- [4] Yu, Z. J., et al. (2020). Efficacy and side effects of drugs commonly used for the treatment of lower urinary tract symptoms associated with benign prostatic hyperplasia. *Frontiers in Pharmacology*, 11, 658.
- [5] Tang, J., et al. (2022). Nanomaterials for delivering antibiotics in the therapy of pneumonia. *International Journal of Molecular Sciences*, 23(24), 15738.
- [6] Michailidis, M., et al. (2022). Antidiabetic drugs in the treatment of Alzheimer’s disease. *International journal of molecular sciences*, 23(9), 4641.
- [7] Briggs, R., et al. (2016). Drug treatments in Alzheimer’s disease. *Clinical medicine*, 16(3), 247-253.
- [8] Hoffman, L., & Bloemer, J. (2021). Side effects of drugs used in the treatment of Alzheimer's disease. In *Side effects of drugs annual* (Vol. 43, pp. 71-77). Elsevier.