SUPPLEMENT MATERIALS

This document provides the details of cross checking for top-50 predicted associations of GTGenie, necessity of extracted BBS and DDS, hyperparameters tuning, the impact of different ratios between positive and negative samples, the performance comparison of different fusion modules and strategies, and sampling negative in the begin of the study. **Table S1** The global top-50 predicted miRNA-disease associations trained on HMDD v2.0. **Table S2** The global top-50 predicted microbe-disease associations trained on HMDAD. **Table S3** The global top-50 predicted lncRNA-disease associations trained on LncRNADisease v2017. **Table S4** The average probabilities of the diseases regarding of all the known related associations. **Table S5** Default hyperparamters settings. **Table S6** Performance comparison of different fusion modules and strategies. **Fig. S1** AUC of GTGenie using different similarities. **Fig. S2** AUC of different learning rates from 1e-1 to 1e-4. **Fig. S3** AUC of different dropout rates. **Fig. S4** AUC of different learning rates from 5e-4 to 5e-3. **Fig. S5** Curves of training loss, test loss, and test AUC. **Fig. S6** Performance comparison of different ratio between positive and negative samples. **Fig. S7** Concat-based/Early fusion variant of GTGenie. **Fig. S8** Add-based variant of GTGenie. **Fig. S9** Late fusion variant of GTGenie.

1. Details of cross checking for top-50 predicted associations of GTGenie

The global top-50 predicted biomarker-disease associations trained on HMDD v2.0, HMDAD, and lncRNADisease v2017 are provided in Tables S1, S2, & S3, respectively.

Table S1. The global top-50 predicted miRNA-disease associations trained on HMDD v2.0

Top 1-25			Top 26-50		
Disease	**miRNA**	**Evidence**	**Disease**	**miRNA**	**Evidence**
Adenocarcinoma	*hsa-mir-1*	PMID:33305905	Carcinoma, hepatocellular	*hsa-mir-9*	dbDEMC
Adenocarcinoma	*hsa-mir-146a*	PMID:32382320	Carcinoma, hepatocellular	*hsa-mir-149*	dbDEMC
Adenocarcinoma	*hsa-mir-18a*	PMID:33942935	Carcinoma, hepatocellular	*hsa-mir-135b*	dbDEMC
Adenocarcinoma	*hsa-mir-34a*	PMID:30700696	Carcinoma, non-small-cell lung	*hsa-mir-29a*	PMID:29495918
Adenoviridae infections	*hsa-mir-29a*	PMID:30405317	Carcinoma, non-small-cell lung	*hsa-mir-200a*	dbDEMC
Adrenocortical carcinoma	*hsa-mir-21*	dbDEMC	Carcinoma, non-small-cell lung	*hsa-mir-203*	PMID:28921827
Breast neoplasms	*hsa-mir-150*	dbDEMC	Carcinoma, non-small-cell lung	*hsa-mir-20a*	dbDEMC
Breast neoplasms	*hsa-mir-106a*	dbDEMC	Carcinoma, non-small-cell lung	*hsa-mir-31*	dbDEMC
Breast neoplasms	*hsa-mir-192*	dbDEMC	Carcinoma, non-small-cell lung	*hsa-mir-92a*	dbDEMC
Breast neoplasms	*hsa-mir-99a*	dbDEMC	Carcinoma, non-small-cell lung	*hsa-mir-18a*	dbDEMC
Breast neoplasms	*hsa-mir-130a*	dbDEMC	Carcinoma, non-small-cell lung	*hsa-mir-19b*	PMID:31564891
Breast neoplasms	*hsa-mir-15b*	dbDEMC	Carcinoma, non-small-cell lung	*hsa-mir-19a*	PMID:23609137
Breast neoplasms	*hsa-mir-144*	dbDEMC	Carcinoma, non-small-cell lung	*hsa-mir-22*	PMID:34729252
Breast neoplasms	*hsa-mir-138*	dbDEMC	Carcinoma, renal cell	*hsa-mir-29a*	PMID:29156819
Breast neoplasms	*hsa-mir-142*	dbDEMC	Carcinoma, renal cell	*hsa-mir-29b*	PMID:30153702
Breast neoplasms	*hsa-mir-212*	dbDEMC	Carcinoma, renal cell	*hsa-mir-1*	PMID: 21745735
Breast neoplasms	*hsa-mir-130b*	dbDEMC	Carcinoma, renal cell	*hsa-mir-133a*	dbDEMC
Breast neoplasms	*hsa-mir-32*	dbDEMC	Carcinoma, renal cell	*hsa-mir-17*	dbDEMC
Carcinoma	*hsa-mir-9*	PMID:30795814	Carcinoma, renal cell	*hsa-mir-181b*	PMID:33560588
Carcinoma, hepatocellular	*hsa-mir-143*	dbDEMC	Carcinoma, renal cell	*hsa-mir-182*	PMID:29424922
Carcinoma, hepatocellular	*hsa-mir-133a*	dbDEMC	Carcinoma, renal cell	*hsa-mir-183*	PMID:30689558
Carcinoma, hepatocellular	*hsa-mir-215*	dbDEMC	Carcinoma, renal cell	*hsa-mir-20a*	dbDEMC
Carcinoma, hepatocellular	*hsa-mir-26b*	dbDEMC	Carcinoma, renal cell	*hsa-mir-221*	PMID:24379138
Carcinoma, hepatocellular	*hsa-mir-34b*	dbDEMC	Carcinoma, renal cell	*hsa-mir-25*	PMID:31202813
Carcinoma, hepatocellular	*hsa-mir-23b*	dbDEMC	Carcinoma, renal cell	*hsa-mir-31*	dbDEMC

Table S2. The global top-50 predicted microbe-disease associations trained on HMDAD

Top 1-25			Top 26-50		
Disease	**microbe**	**Evidence**	**Disease**	**microbe**	**Evidence**
Bacterial Vaginosis	*Proteobacteria*^{P}	PMID:32296412	Ileal Crohn's disease	*Proteobacteria*^{P}	PMID:31530835
Type 2 diabetes	*Prevotella*^{G}	PMID:34040023	Necrotizing Enterocolitis	*Clostridia*^{C}	PMID:30980889
Type 2 diabetes	*Actinobacteria*^{P}	PMID:28177125	Bacterial Vaginosis	*Staphylococcus aureus*^{S}	PMID:25926958
Colorectal carcinoma	*Proteobacteria*^{P}	PMID:34650531	Colorectal carcinoma	*Haemophilus*^{G}	PMID:28988196
Crohn's disease	*Actinobacteria*^{P}	PMID:31911822	Necrotizing Enterocolitis	*Lactobacillus*^{G}	PMID:10575148
Crohn's disease	*Proteobacteria*^{P}	PMID:31530835	Bacterial Vaginosis	*Faecalibacterium prausnitzii*^{S}	unconfirmed
Crohn's disease	*Prevotella*^{G}	PMID: 28542929	Irritable bowel syndrome	*Faecalibacterium prausnitzii*^{S}	PMID:24713205
Crohn's disease	*Bacteroidetes*^{P}	PMID:33669168	Clostridium difficile infection	*Actinobacteria*^{P}	PMID:31876614
Liver cirrhosis	*Actinobacteria*^{P}	PMID:31726747	Irritable bowel syndrome	*Firmicutes*^{P}	PMID:30829919
Asthma	*Firmicutes*^{P}	PMID:32072252	Colorectal carcinoma	*Clostridium coccoides*^{S}	PMID: 26541655
Asthma	*Actinobacteria*^{P}	PMID:29318023	Clostridium difficile infection	*Fusobacterium*^{G}	PMID:33227279
Colorectal carcinoma	*Actinobacteria*^{P}	PMID:35049922	Irritable bowel syndrome	*Clostridium*^{G}	PMID:18026576
Bacterial Vaginosis	*Bacteroidetes*^{P}	PMID:20819230	Liver cirrhosis	*Firmicutes*^{P}	PMID:33708204
Clostridium difficile infection	*Prevotella*^{G}	PMID:33854066	Irritable bowel syndrome	*Lactobacillus*^{G}	PMID:21860817
Clostridium difficile infection	*Bacteroides ovatus*^{S}	PMID:29076071	Ileal Crohn's disease	*Firmicutes*^{P}	PMID:25844959
Crohn's disease	*Lactobacillus*^{G}	PMID:15113451	Bacterial Vaginosis	*Clostridium leptum*^{S}	unconfirmed
Psoriasis	*Prevotella*^{G}	PMID:32545459	Type 2 diabetes	*Streptococcus*^{G}	PMID:32246318
Clostridium difficile infection	*Bacteroides*^{G}	PMID:32660525	Crohn's disease	*Clostridium coccoides*^{S}	unconfirmed
Clostridium difficile infection	*Bacteroides vulgatus*^{S}	PMID:32660525	Cystic fibrosis	*Clostridium difficile*^{S}	PMID:28078087
Psoriasis	*Bacteroidetes*^{P}	PMID:33384669	COPD	*Actinobacteria*^{P}	PMID:23071781
Colorectal carcinoma	*Lactobacillus*^{G}	PMID:32419125	Type 2 diabetes	*Haemophilus*^{G}	PMID:32246318
Irritable bowel syndrome	*Actinobacteria*^{P}	PMID:31244784	Colorectal carcinoma	*Lachnospiraceae*^{F}	PMID:28988196
Clostridium difficile infection	*Clostridia*^{C}	PMID:22555464	Type 2 diabetes	*Fusobacterium*^{G}	PMID:31901868
Clostridium difficile infection	*Lactobacillus*^{G}	PMID:24856984	Ileal Crohn's disease	*Klebsiella*^{G}	unconfirmed
Clostridium difficile infection	*Haemophilus*^{G}	unconfirmed	Crohn's disease	*Clostridia*^{C}	PMID:31911822

The superscripts *D, P, C, O, F, G,* and *S* represent the domain, phylum, class, order, family, genus, and species, respectively.

Table S3. The global top-50 predicted lncRNA-disease associations trained on LncRNADisease v2017

Top 1-25			Top 26-50		
Disease	**lncRNA**	**Evidence**	**Disease**	**lncRNA**	**Evidence**
Alzheimer's disease	*H19*	PMID:30107531	Osteosarcoma	*PVT1*	Lnc2Cancer;MNDR
Alzheimer's disease	*MALAT1*	MNDR	Breast cancer	*AFAP1-AS1*	Lnc2Cancer;MNDR
Cancer	*HOTTIP*	MNDR	Breast cancer	*BANCR*	Lnc2Cancer;MNDR
Cancer	*NEAT1*	MNDR	Breast cancer	*HOTTIP*	Lnc2Cancer;MNDR
Cancer	*TUG1*	MNDR	Breast cancer	*HULC*	Lnc2Cancer;MNDR
Gastric cancer	*AFAP1-AS1*	Lnc2Cancer;MNDR	Breast cancer	*PCAT1*	Lnc2Cancer;MNDR
Gastric cancer	*BCYRN1*	Lnc2Cancer;MNDR	Cervical cancer	*UCA1*	Lnc2Cancer;MNDR
Gastric cancer	*PCAT1*	Lnc2Cancer;MNDR	Lung cancer	*NEAT1*	Lnc2Cancer;MNDR
Gastric cancer	*SOX2-OT*	MNDR	Lung cancer	*PVT1*	Lnc2Cancer;MNDR
Lung adenocarcinoma	*CDKN2B-AS1*	MNDR	Lung cancer	*TUG1*	PMID:28069000
Lung adenocarcinoma	*H19*	Lnc2Cancer;MNDR	Esophageal cancer	*HOTAIR*	Lnc2Cancer;MNDR
Lung adenocarcinoma	*UCA1*	Lnc2Cancer;MNDR	Ovarian cancer	*MEG3*	Lnc2Cancer;MNDR
Glioma	*PVT1*	Lnc2Cancer;MNDR	Schizophrenia	*H19*	PMID:33093650
Glioma	*UCA1*	Lnc2Cancer;MNDR	Endometrial cancer	*CDKN2B-AS1*	PMID:34712660
Nasopharyngeal carcinoma	*GAS5*	Lnc2Cancer;MNDR	Endometrial cancer	*H19*	Lnc2Cancer;MNDR
Nasopharyngeal carcinoma	*PVT1*	Lnc2Cancer;MNDR	Esophageal squamous cell carcinoma	*GAS5*	Lnc2Cancer;MNDR
Nasopharyngeal carcinoma	*UCA1*	Lnc2Cancer;MNDR	Esophageal squamous cell carcinoma	*HOTTIP*	Lnc2Cancer;MNDR
Hepatocellular carcinoma	*BCYRN1*	PMID:31339046	Esophageal squamous cell carcinoma	*PVT1*	Lnc2Cancer
Hepatocellular carcinoma	*CRNDE*	PMID:30230527	Esophageal squamous cell carcinoma	*LINC-ROR*	Lnc2Cancer;MNDR
Papillary thyroid carcinoma	*H19*	PMID:31403942	Colorectal cancer	*CCAT2*	Lnc2Cancer;MNDR
Papillary thyroid carcinoma	*MALAT1*	PMID:29987950	Colorectal cancer	*XIST*	Lnc2Cancer;MNDR
Prostate cancer	*BANCR*	unconfirmed	Colorectal cancer	*ATB*	MNDR
Melanoma	*MEG3*	Lnc2Cancer;MNDR	Lymphoma	*CDKN2B-AS1*	unconfirmed
Melanoma	*NEAT1*	Lnc2Cancer;MNDR	Lymphoma	*H19*	PMID:30610809
Osteosarcoma	*GAS5*	Lnc2Cancer;MNDR	Kidney cancer	*CDKN2B-AS1*	PMID:32814766

Table S4. The average probabilities of the diseases regarding of all the known related associations

Dataset	Position	Disease	Average Probabilities	Rank
HMDD	1-4	Adenocarcinoma	0.4512	32
	5	Adenoviridae infections	0.4858	24
	6	Adrenocortical carcinoma	0.5019	19
	7-18	Breast neoplasms	0.7892	2
	19	Carcinoma	0.4459	34
	20	Carcinoma, hepatocellular	0.8608	1
HMDAD	1,13	Bacterial Vaginosis	0.4155	4
	2,3	Type 2 diabetes	0.2557	9
	4,12	Colorectal carcinoma	0.1928	16
	5,6,7,8,16	Crohn's disease	0.2935	5
	9	Liver cirrhosis	0.5848	2
	10-11	Asthma	0.2314	11
	14-15,18-19	Clostridium difficile infection	0.2647	7
	17,20	Psoriasis	0.2042	15
LncRNADisease	1-2	Alzheimer's disease	0.1788	34
	3-5	Cancer	0.2990	13
	6-9	Gastric cancer	0.4969	3
	10-12	Lung adenocarcinoma	0.3139	9
	13,14	Glioma	0.3068	12
	15-17	Nasopharyngeal carcinoma	0.2818	15
	18,19	Hepatocellular carcinoma	0.5250	1
	20	Papillary thyroid carcinoma	0.2432	22

2. AUC of GTGenie with or without BBS and DDS

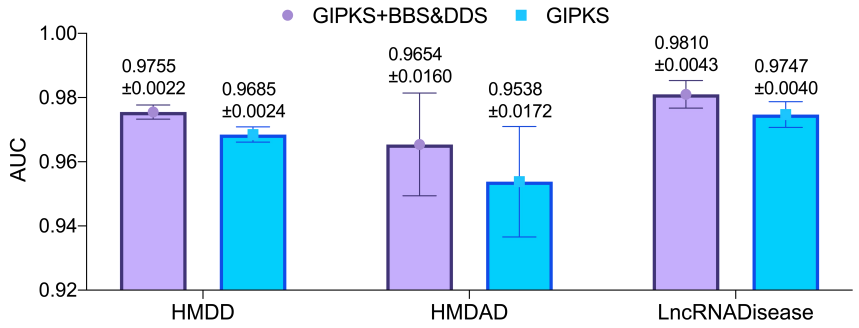

Fig. S1. AUC of GTGenie using different similarities

3. Hyperparameters tuning

We conducted the experiments of tuning the hyperparameters including learning rate, dropout rate, and training epoch. We first set up the learning rate and dropout rate in the common setting ranges of [1e-1, 1e-2, 1e-3, 1e-4] and [0.0, 0.1, 0.2, 0.3], respectively. The results are provided as Fig. S2 and Fig. S3:

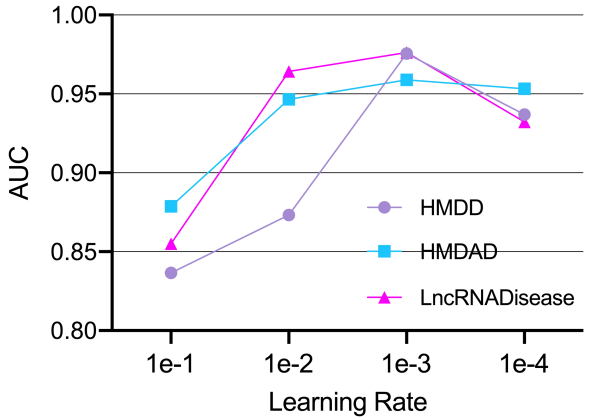

Fig. S2. AUC of different learning rates from 1e-1 to 1e-4

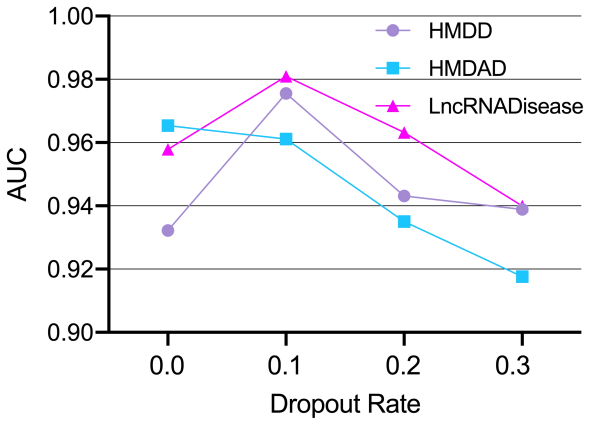

Fig. S3. AUC of different dropout rates

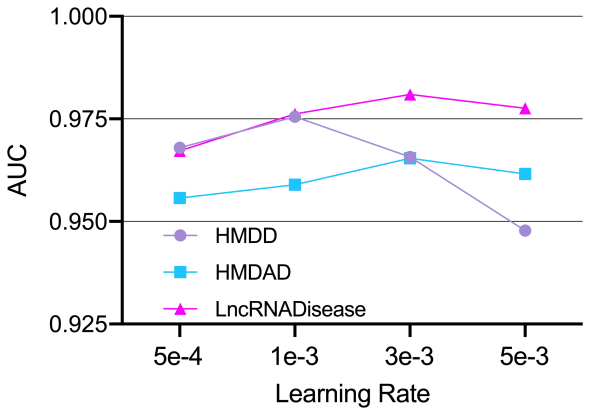

Fig. S4. AUC of different learning rates from 5e-4 to 5e-3

As we can see, the optimal learning rates and dropout rates appear on the combinations of 1e-3/0.1, 1e-3/0.0, and 1e-3/0.0 for the HMDD, HMDAD, and LncRNADisease datasets, respectively. We also note that significant differences can be observed by tuning the learning rate around 1e-3. Therefore, we further performed a fine-grained parameter tuning of the learning rates ranging from 5e-4 to 5e-3 as shown in Fig. S6. Accordingly, the optimal learning rates for HMDD, HMDAD, and LncRNADisease were ultimately set to 1e-3, 3e-3, and 3e-3, respectively.

As for the training epoch, we set the maximum epoch to 700 and the curves of training loss, test loss, and test AUC are plotted as follows:

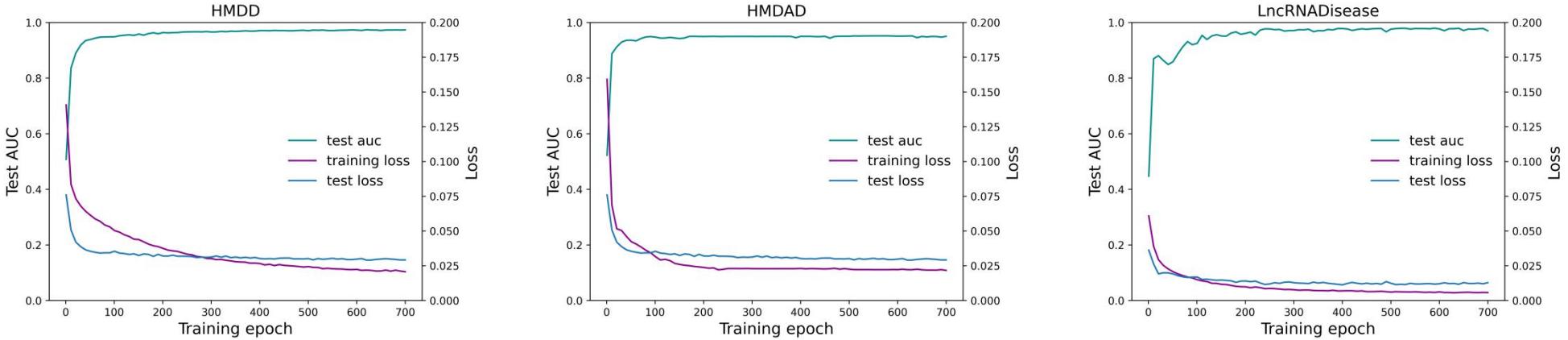

Fig. S5. Curves of training loss, test loss, and test AUC

The results demonstrate the convergence of GTGenie on both training and testing sets. The loss decreases rapidly in the first 100 epoch and then reaches a relative stable stage. We therefore set up the training epochs according to the different convergence situations on different datasets. Finally, the default hyperparamters settings of different datasets are provided as follows:

Table S5. Default hyperparamters settings

Dataset	Learning Rate	Dropout Rate	Training Epoch
HMDD	1e-3	0.1	500
HMDAD	3e-3	0.0	300
LncRNADisease	3e-3	0.1	300

4. Different ratios between positive and negative samples

Since the number of potential negative samples is much larger than the known positive sample, a simulation experiment has been conducted for the evaluation of using different ratios between positive and negative samples on HMDD, ranging from 1:1 to 1:9. As we can see in Fig. S6, as the proportion of negative samples increases, the corresponding AUC and AUPR values present the decline of different degrees. Clearly, the resultant class imbalance issue overwhelms the model training and thereby degenerating the classification performance. Therefore, it is not necessary to use larger amount of negative samples in model training compared with that of positive ones.

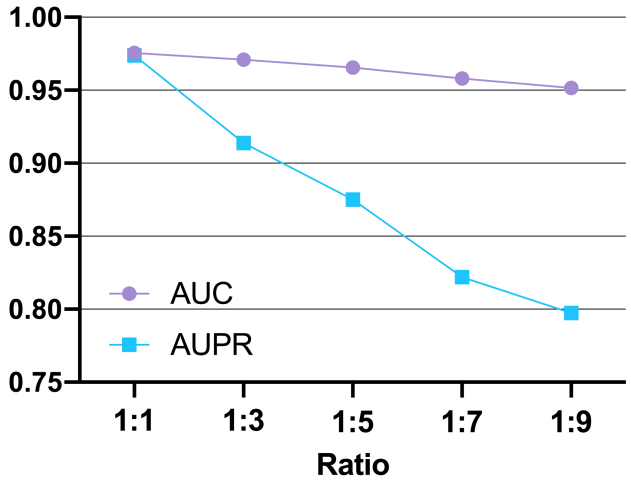

Fig. S6. Performance comparison of different ratio between positive and negative samples

5. Different fusion modules and strategies

Different fusion model variants including concatenation and addition operation were implemented for the comparison of the proposed BFN. In addition, based on which stage fusion occurs, the existing multimodal fusion strategies can be roughly divided into three categories, i.e., early fusion (at feature level), late fusion (at decision level), intermediate fusion (mixing early and late fusion) [1]. BFN as the key fusion block in GTGenie should belong to intermediate fusion strategy. We conducted an additional experiment to evaluate the performance of early fusion and late fusion strategies in our model separately. As for early fusion, however, we cannot directly fuse the graph features and text features at the earliest stage because of dimension mismatch issue. Therefore, the graph features and text features were first extracted via GAT and BioBERT, respectively, and then projected into the same dimensional embedding space for early fusion by concatenation. In this way, the model of early fusion is the same as the concatenation operation. The architectures diagrams of concat-based/early fusion, add-based, and late fusion are designed as Fig S7, S8, & S9, respectively.

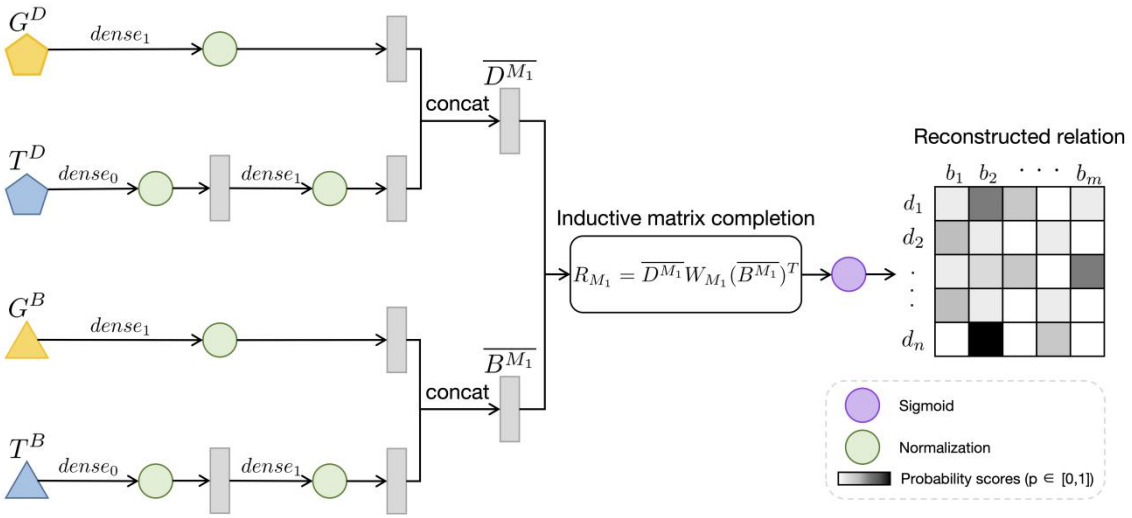

Fig. S7. Concat-based/Early fusion variant of GTGenie

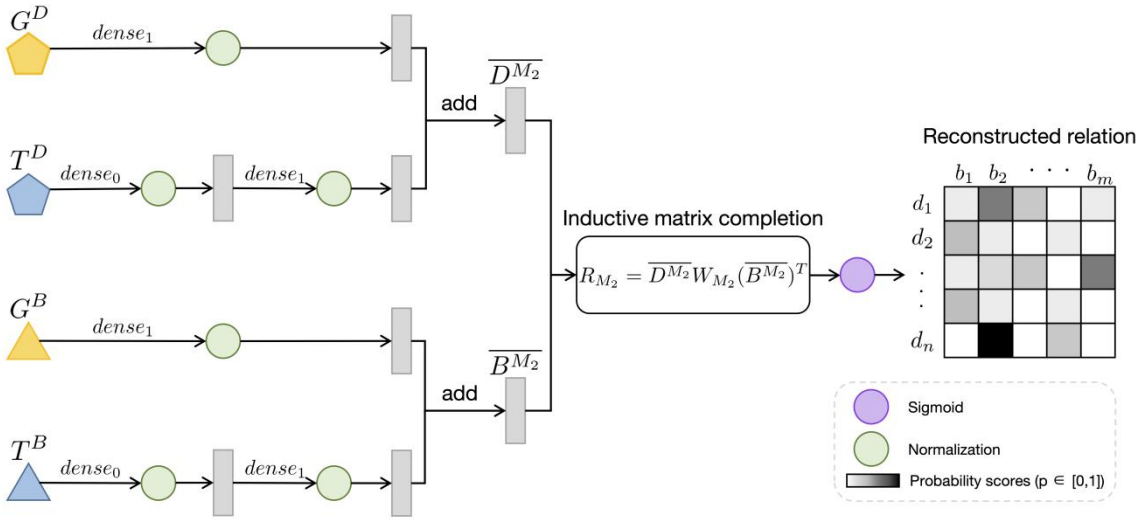

Fig. S8. Add-based variant of GTGenie

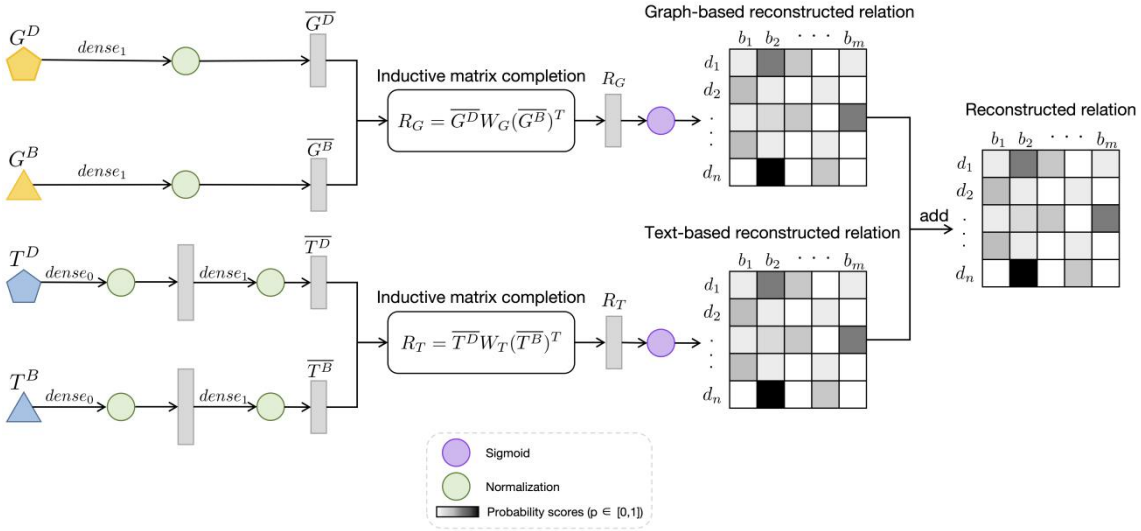

Fig. S9. Late fusion variant of GTGenie

where $\overline{D^{M_1}} \in R^{n\times(r+t)}$, $\overline{B^{M_1}} \in R^{m\times(r+t)}$, $\overline{D^{M_2}} \in R^{n\times(r+t)}$, and $\overline{B^{M_2}} \in R^{m\times(r+t)}$, and t represents the output dimension of TRR. The results on HMDD, HMDAD, and LncRNADisease are tabulated as Table S76.

Table S6. Performance comparison of different fusion modules and strategies

Dataset	Modules/Strategies	AUC	AUPR
HMDD	Concat-based/Early fusion	0.9702±0.0025	0.9667±0.0031
	Add-based	0.9693±0.0025	0.9658±0.0027
	Late fusion	0.9657±0.0041	0.9605±0.0049
	BFN	**0.9755±0.0022**	**0.9739±0.0028**
HMDAD	Concat-based/Early fusion	0.9571±0.0184	0.9517±0.0240
	Add-based	0.9543±0.0158	0.9532±0.0170
	Late fusion	0.9535±0.0331	0.9518±0.0448
	BFN	**0.9654±0.0160**	**0.9635±0.0174**
LncRNADisease	Concat-based/Early fusion	0.9743±0.0045	0.9685±0.0066
	Add-based	0.9713±0.0057	0.9656±0.0072
	Late fusion	0.9680±0.0064	0.9573±0.0100
	BFN	**0.9810±0.0043**	**0.9788±0.0066**

The results demonstrate that the fusion effectiveness of BFN is superior to all these variants on the three datasets. The main difference between them is that BFN-based module separately reconstructs the associations matrices (i.e., R_G and R_T) within the same single modality by inductive matrix completion whereas both concat-based/early fusion and add-

based modules directly integrate features of different modalities into the same entity representation (i.e., $\overline{D^{M_1}}$, $\overline{B^{M_1}}$, $\overline{D^{M_2}}$ and $\overline{B^{M_2}}$). The late fusion strategy obtains the final reconstructed matrix based on the results of different classifiers. The main reasons of the BFN-based module achieves the best results are as follows: i) inductive matrix completion as the key component in BFN is effective in fusing the same modality representation of different entities, rather than the same entity representation of different modalities learned by concat-based or add-based modules; ii) Compare with early fusion and late fusion, the intermediate fusion adopted by BFN enables a better joint multimodal representation by effectively modeling intra- and inter-modality interactions. We also note that the result of add-based module is slightly inferior to that of concat-based module. Perhaps this is because add-based module could suffer from the risk of information loss by messing up the learning for important, related features with redundant, irrelevant ones. In addition, the early fusion variant performs better than late fusion generally. This could be attributed to the fact that such a decision fusion-based method cannot mitigate the contradiction between the individual decisions made from different modal inference.