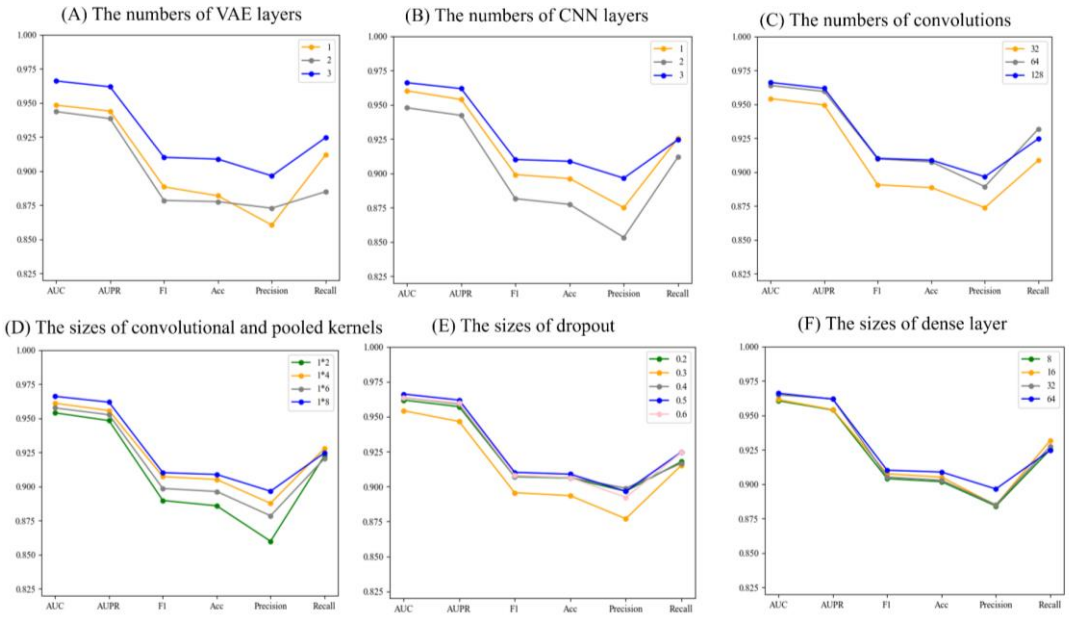
Supplemental document

Supplemental instructions for 5CV:

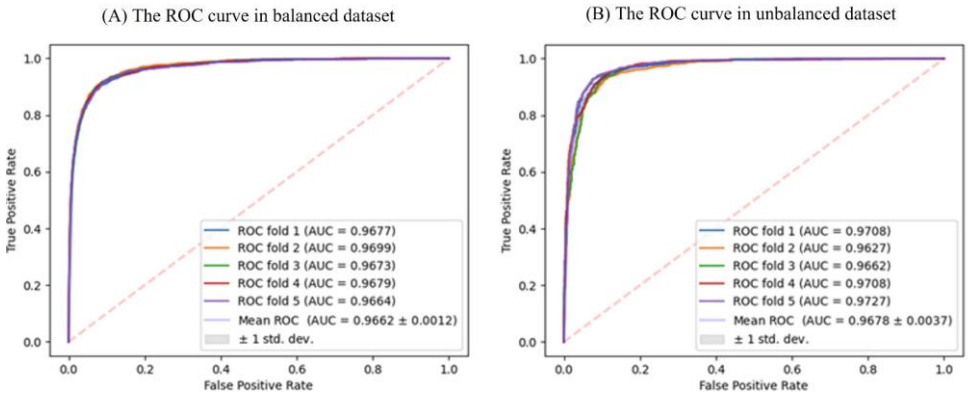
In the experiment, we use four datasets, which are the balanced dataset and the unbalanced dataset on the HMDD v2.0 dataset and the balanced and the unbalanced dataset on the HMDD v3.2 dataset. For the balanced dataset, the ratio of positive to negative samples is 1:1, where the negative samples are randomly selected from all unknown samples in the same number as the positive samples. For the unbalanced dataset, all unknown samples are considered as negative samples. To verify the ability of DMFVAE to infer the association between unknown miRNAs and diseases, we conduct 10 times of 5CV in the experiment. For example, for the HMDD v2.0 balanced dataset, there are a total of 10860 samples, and the steps for its 10 times 5CV are as follows:

First, all samples are randomly divided into five equal parts, the training set is four of them, with 8688 samples, and the test set is one of them, with 2174 samples. Then, the test set is taken in turn, and the average of the 5 results can be used to obtain a result of 5CV once. At last, after the above process is cycled 10 times, the average result of 10 times is calculated as shown in this paper.

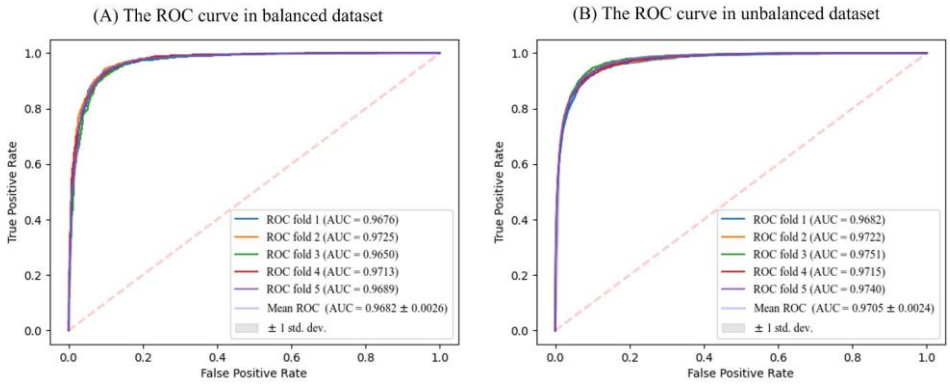
Supplemental Figures:



Supplemental Fig S1 The detailed parameters tuning experiments on HMDD v2.0 balanced dataset. The F1 and Acc represent F1-score and accuracy, respectively



Supplementary Fig S2 The ROC curves of DMFVAE in five-fold cross-validation on HMDD v2.0 balanced and unbalanced datasets



Supplementary Fig S3. The ROC curves of DMFVAE in five-fold cross-validation on HMDD v3.2 balanced and unbalanced datasets

Supplemental Tables:

Supplement Table S1-S2: The individual values of the evaluation indicators on HMDD v2.0 dataset. For balanced dataset, the average value is calculated from AUC, AUPR, F1 and Acc. For unbalanced dataset, the average value is calculated from AUPR and F1:

Supplement Table S2. DMFVAE compared with other models on balanced dataset

models	AUC	AUPR	F1	Acc	Precision	Recall	Avg
SMALF	0.9503	0.9472	0.8868	0.8860	0.8808	0.8931	0.9176
ERMDA	0.9013	0.9043	0.8356	0.8300	0.8096	0.8639	0.8678
VAEMDA	0.9206	0.9215	0.8466	0.8382	0.8045	0.8938	0.8817
GRPAMDA	0.9346	0.9331	0.8606	0.8600	0.8579	0.8637	0.8971
DMFVAE	0.9662	0.9619	0.9102	0.9089	0.8966	0.9247	0.9368

Supplement Table S3. DMFVAE compared with other models on unbalanced dataset

models	AUC	AUPR	F1	ACC	Precision	Recall	Avg
MDA-GCNFTG	0.9448	0.6137	0.5080	0.9718	0.6324	0.5628	0.5609
GBDT-LR	0.9397	0.4816	0.6688	0.9757	0.3029	0.4168	0.5752
NIMGSA	0.9354	0.4567	0.4346	0.9721	0.5229	0.3518	0.4457
GAEMDA	0.9321	0.4432	0.6836	0.9748	0.2262	0.3382	0.5634

DMFVAE	0.9678	0.6556	0.6055	0.9802	0.7053	0.5313	0.6306

Supplement Table S4-S4: The individual values of the evaluation indicators on HMDD v3.2 dataset. For balanced dataset, the average value is calculated from AUC, AUPR, F1 and Acc. For unbalanced dataset, the average value is calculated from AUPR and F1:

Supplement Table S5. DMFVAE compared with other models on unbalanced dataset

	AUC	AUPR	F1	Acc	Precision	Recall	Avg
ABMDA	0.9152	0.9069	0.8402	0.8439	0.8398	0.8479	0.8766
VGAMF	0.9443	0.9397	0.8764	0.8763	0.8754	0.8775	0.9092
MLRDFM	0.9545	0.9550	0.8833	0.8833	0.8833	0.8834	0.9190
DMFVAE	0.9682	0.9639	0.9140	0.9123	0.8967	0.9322	0.9396

Supplement Table S6. DMFVAE compared with other models on unbalanced dataset

	AUC	AUPR	F1	Acc	Precision	Recall	Avg
VGAMF	0.9260	0.5217	0.40437	0.9748	0.7338	0.28796	0.4630
MLRDFM	0.9311	0.5387	0.4228	0.9757	0.7630	0.2928	0.4808
DMFVAE	0.9705	0.6853	0.6191	0.9795	0.7142	0.5514	0.6522

Supplement Table S5. Top 20 candidate miRNAs associated with CN, where H3, DEMC and miR represent HMDD v3.2, DEMCDEMC and miR2Disease respectively

rank	miRNA	evidence	rank	miRNA	evidence
1	hsa-mir-483	DEMC	11	hsa-mir-19a	HM, DEMC, miR
2	hsa-mir-135a	DEMC, miR	12	hsa-mir-125b	DEMC
3	hsa-mir-141	DEMC, miR	13	hsa-mir-302b	DEMC, HM
4	hsa-mir-20b	DEMC, miR	14	hsa-mir-125a	DEMC, miR
5	hsa-mir-146a	HM, DEMC, miR	15	hsa-mir-155	DEMC, miR
6	hsa-mir-196a	HM, DEMC, miR	16	hsa-mir-199b	DEMC
7	hsa-mir-31	DEMC	17	hsa-mir-20a	DEMC, miR
8	hsa-mir-34a	DEMC, miR	18	hsa-mir-21	HM, DEMC, miR
9	hsa-let-7a	DEMC, miR	19	hsa-mir-182	DEMC, miR
10	hsa-mir-148a	DEMC	20	hsa-mir-143	HM, DEMC, miR

Supplement Table S6. Top 20 candidate miRNAs associated with EN, where H3, DEMC and miR represent HMDD v3.2, DEMCDEMC and miR2Disease respectively

rank	miRNA	evidence	rank	miRNA	evidence

1	hsa-mir-23a	DEMC	11	hsa-mir-124	DEMC
2	hsa-mir-125a	DEMC	12	hsa-mir-10b	DEMC
3	hsa-mir-7	DEMC, miR	13	hsa-mir-182	DEMC
4	hsa-mir-20b	DEMC, miR	14	hsa-mir-224	DEMC
5	hsa-mir-200b	DEMC	15	hsa-mir-142	DEMC
6	hsa-mir-708	DEMC, miR	16	hsa-mir-199b	DEMC
7	hsa-mir-122	PMID: 22751839	17	hsa-mir-10a	DEMC
8	hsa-mir-16	DEMC	18	hsa-mir-378a	DEMC
9	hsa-mir-135a	DEMC	19	hsa-mir-133b	DEMC, HM
10	hsa-mir-19b	DEMC	20	hsa-mir-125b	DEMC

Supplemental Formulas:

Supplemental Formula 1: The specific formula of each performance evaluation metric to measure the prediction performance of DMFVAE:

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}$$

$$precision = \frac{TP}{TP + FP} \tag{2}$$

$$recall = \frac{TP}{TP + FN} \tag{3}$$

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \tag{4}$$

$$TPR = \frac{TP}{TP + FN} \tag{5}$$

$$FPR = \frac{FP}{TN + FP} \tag{6}$$

where TP represents true positive, TN represents true negative, FP represents false positive, FN represents false negative. TPR and FPR represent true positive rate and false positive rate, according to them, the ROC curve is obtained, and the corresponding area is AUC. Similarly, AUPR is obtained according to precision and recall.

Supplemental Formula 2: The specific formulas on how to obtain semantic similarity of diseases:

Firstly, we calculate the contribution $DC1_p$ of disease d in DAG_p to disease p as follow:

$$\begin{cases} DC1_p(d) = 1, & if\ d = p \\ DC1_p(d) = \max\left\{\Delta * DC1_p(d') \middle| d' \in children\ of\ d\right\}, & if\ d \neq p \end{cases} \tag{7}$$

where Δ represents the semantic contribution decay factor and we set it as 0.5. Then, we can calculate the semantic value $SV1$ of disease p, which is defined as follow:

$$SV1(p) = \sum_{d \in T_p} DC1_p(d) \tag{8}$$

Then, we can get the semantic similarity $SD1(d_i, d_j)$ between disease d_i and d_j as below:

$$SD1(d_i, d_j) = \frac{\sum_{d \in T_{d_i} \cap T_{d_j}} \left(DC1_{d_i}(d) + DC1_{d_j}(d)\right)}{SV1(d_i) + SV1(d_j)} \tag{9}$$

where $SD1$ is the first kind of disease semantic similarity and it is a 383×383 or 374×374 symmetry matrix.

Additionally, we also introduce another method to calculate the disease semantic similarity. The semantic contribution $DC2_p$ of disease d to disease p can be described as below:

$$DC2_p(d) = -\log\left(\frac{the\ number\ of\ DAGs\ including\ \mathrm{d}}{the\ number\ of\ diseases}\right). \tag{10}$$

Then, we can obtain the semantic value $SV2$ of disease p from equation (5), and the disease semantic similarity $SD2(d_i, d_j)$ between disease d_i and d_j from equation (6).

$$SV2(p) = \sum_{d \in T(p)} DC2_p(d) \tag{11}$$

$$SD2(d_i, d_j) = \frac{\sum_{d \in T_{d_i} \cap T_{d_j}} \left(DC2_{d_i}(d) + DC2_{d_j}(d)\right)}{SV2(d_i) + SV2(d_j)} \tag{12}$$

Finally, we can obtain the disease semantic similarity SD by taking the average of be $SD1$ and $SD2$.

Supplemental Formula 3: The detailed process to obtain enhanced association matrix A' by using EASNN:

Firstly, using SNN to obtain suitable location: sort the similarity interaction profile SD_i (SM_i) in descending order and search backwards from the first point until a position k that satisfies the following conditions is found:

$$SD_i(k) - SD_i(k+l) < SD_i(k+l) - SD_i(k+3l) \tag{13}$$

where l is the step length, which is generally set at one-fifth of the length of SD_i.

Secondly, for each similarity matrix, a new matrix M and D are obtained:

$$M\left(m_i,m_j\right)=\begin{cases}SM\left(m_i,m_j\right) & SM\left(m_i,m_j\right)\in SNN\left(m_i\right)\\ 0 & SM\left(m_i,m_j\right)\notin SNN\left(m_i\right)\end{cases} \tag{14}$$

$$D\left(d_i,d_j\right)=\begin{cases}SD\left(d_i,d_j\right) & SD\left(d_i,d_j\right)\in SNN\left(d_i\right)\\ 0 & SD\left(d_i,d_j\right)\notin SNN\left(d_i\right)\end{cases} \tag{15}$$

Thirdly, M and D combine with the association matrix A to obtain a new interaction profile A_{md} :

$$A_m\left(m_i\right)=\frac{1}{N_{m_i}}\sum_{j=1}^{K}w_jA\left(m_j\right) \tag{16}$$

$$A_d\left(d_i\right)=\frac{1}{N_{d_i}}\sum_{j=1}^{K}w_jA\left(d_j\right) \tag{17}$$

$$A_{md}=\frac{\mu A_m+\nu A_d}{\mu+\nu} \tag{18}$$

where m_1 (d_i) to m_K (d_j) are sorted in descending order by their similarity to m_i (d_i), $N_{m_i}=\sum_{j=1}^{K}M\left(m_i,m_j\right)$, $N_{d_i}=\sum_{j=1}^{K}D\left(d_i,d_j\right)$, $\mu=\nu=0.5$

Finally, using the new interaction profile A_{md} to update the original miRNA-disease associations matrix A :

$$A'=\max\left(A,A_{md}\right) \tag{19}$$