# 基于多任务学习和超图神经网络的微生物-药物关联预测

王波,王钧祺,杜晓昕,孙明,王彤轩,黎景威

(齐齐哈尔大学 计算机与控制工程学院;黑龙江省大数据网络安全检测分析重点实验室,黑龙江 齐齐哈尔 161006)

**摘 要**:传统的生物实验方法寻找微生物与药物关系不仅耗时费力,而且成本极高.因此,为了降低实验成本并提高效率,计算方法被用于预测微生物-药物关联.然而,现有方法忽视了疾病作为中介的关键作用,导致数据稀疏性问题.为此,提出了基于多任务学习的模型 (MTLTPMDA),用于同时预测微生物-药物和疾病-药物关联.模型通过共享药物节点的特征来增强任务间的联系,并利用超图神经网络(HGNN)探索微生物、药物和疾病之间的复杂交互.通过构建微生物-药物和疾病-药物超图,HGNN 有效捕捉了多节点间的高阶关系.在五重交叉验证下,MTLTPMDA 实现了 AUC 为 0.903 3 和 AUPR 为 0.893 0,优于多种现有方法,展示了模型在预测潜在关联上的有效性.

**关键词**:微生物与药物关联;疾病与药物关联;多任务学习技术;数据稀疏性;超图神经网络

**中图分类号**:TP391 **文献标志码**:A **文章编号**:1000-2367(2026)01-0068-09

微生物与人类之间存在着密切的关系.人类与微生物共同生活在同一个生态系统中.一方面,人体微生物与人体共生共存,在维持人体健康和免疫系统功能等方面起着重要作用[1].例如,在皮肤上,大约有 500~1 000 个共生细菌参与培养免疫系统和维持皮肤炎症的稳态控制[2].另一方面,已有研究表明,微生物的异常生长或下降会影响人体健康.例如细菌、病毒和真菌可以引发传染病,沙门氏菌[3]、大肠杆菌[4]等可以污染食品,导致食物中毒等[5].

微生物具有高度的适应性和变异能力,在抗生素选择压力下易产生耐药性[6].近年来,微生物多样性在抗生素和抗癌药物开发中发挥了重要作用,常见的微生物来源抗肿瘤药物包括蒽环霉素、博来霉素等[7].传统上,微生物与药物关联的识别依赖生物实验,成本高、周期长,且难以精准筛选.因而,构建高效准确的计算模型预测微生物与药物的潜在关联,有助于指导实验、加速药物发现.

随着微生物-药物关联数据库的建立,相关的计算预测模型不断发展.LONG 等[8]提出了基于图卷积网络的 GCNMDA 模型,在隐藏层引入条件随机场与注意机制以提升邻域特征聚合效果,随后又开发了具分层注意力的图注意网络框架 EGATMDA,用于更精细地学习节点嵌入[9].尽管这些方法取得良好效果,但从相似性中提取的特征仍可能包含噪声,影响预测准确性.为此,研究者提出了更多新方法,如 LONG 等[10]结合 metapath2vec 和二分网络推荐构建异构网络嵌入模型 HNERMDA;DENG 等[11]基于多模态特征构建属性图,训练 VGAE 获取可解释表示后用深度神经网络进行预测;MA 等[12]整合多组数据并利用超图结构引入正则化以优化矩阵分解模型.

尽管现有模型在微生物-药物关联预测方面取得进展,但仍存在数据稀疏和模型稳健性不足的问题.为此,提出了一种基于多任务学习的新方法,构建微生物-药物与疾病-药物两个关联网络,并通过共享药物节点

**收稿日期**:2024-10-08;**修回日期**:2024-11-18.

**基金项目**:黑龙江省省属高等学校基本科研业务费国自然培育一般项目(145409324).

**作者简介(通信作者)**:王波(1980—),男,黑龙江齐齐哈尔人,齐齐哈尔大学教授,博士,研究方向为生物医学大数据分析与处理,E-mail:bowangdr@qqhru.edu.cn.

**引用本文**:王波,王钧祺,杜晓昕,等.基于多任务学习和超图神经网络的微生物-药物关联预测[J].河南师范大学学报(自然科学版),2026,54(1):68-76.(Wang Bo,Wang Junqi,Du Xiaoxin,et al.Predicting microbe-drug associations based on multi-task learning and hypergraph neural network[J].Journal of Henan Normal University(Natural Science Edition),2026,54(1):68-76.DOI:10.16366/j.cnki.1000-2367.2024.10.08.0001.)

实现任务间的信息交互.利用超图神经网络(HGNN)挖掘微生物、药物和疾病之间的复杂关系,在共享潜在特征空间的基础上提升模型泛化能力,有效缓解了数据稀疏问题[13-14].

# 1　原理与方法

## 1.1　数据集整理

近年来,多个公开数据库支持微生物-药物关联研究.2018 年,SUN 等[15]构建了首个微生物-药物关联数据库 MDAD,包含 142 种微生物与 627 种药物的 1 152 条关联;同年,RAJPUT 等[16]开发了 ABiofilm,收录了 1 720 种药物与 140 种微生物的 2 884 条关联.2020 年,ANDERSEN 等[17]建立了 DrugVirus 数据集,涵盖 95 种病毒和 175 种药物,共 933 条关联.WANG 等[18]整合 17 个数据库并生成 3 个新数据集,其中使用的 MDgAs_(DS20)和 DgDsAs_(DS19)分别包含微生物-药物和疾病-药物的实验验证关联.MDgAs_(DS20)包含 446 种药物与 226 种微生物的 1 634 条关联,DgDsAs_(DS19)包含 446 种药物与 2 667 种疾病的 17 123 条关联.因其数据全面且唯一涵盖两类关联,本文实验均基于该数据库进行.

## 1.2　子网的构建

在这一部分中,首先描述了“微生物与药物”和“疾病与药物”两个子网络的形成过程,然后描述了构建子网络相似度的方法.

### 1.2.1　人类微生物与药物的关联

在研究中,微生物与药物的关联来源于 WANG 等[18]收集整合的微生物与药物关联数据集,包含了经过实验验证的 $n_d$(446)种药物和 $n_m$(226)种微生物的关联.在实验中,鉴定出的微生物与药物之间的关系用 $n_m$ 列 $n_d$ 行矩阵 $\boldsymbol{A}$ 表示.如果相应的微生物与相应的药物相关,则 $\boldsymbol{A}(i,j)=1$,否则为 0(表示关系未知).矩阵 $\boldsymbol{A}$ 表示为:

$$\boldsymbol{A}(i,j)=\begin{cases}1,\text{如果 } m_i \text{ 和 } d_j \text{ 存在关系},\\0,\text{否则},\end{cases}$$

其中,关联矩阵 $\boldsymbol{A}\in\mathbf{R}^{n_d\times n_m}$,$m_i$ 表示第 $i$ 个微生物,$d_j$ 表示第 $j$ 个药物.

### 1.2.2　人类疾病与药物的关联

类似于微生物与药物子网络,创建了一个有 $n_D$ 行和 $n_d$ 列的矩阵 $\boldsymbol{B}$.如果疾病与药物相关,则 $\boldsymbol{B}(i,j)=1$,否则为 0.矩阵 $\boldsymbol{B}$ 表示为:

$$\boldsymbol{B}(i,j)=\begin{cases}1,\text{如果 } D_i \text{ 和 } d_j \text{ 存在关系},\\0,\text{否则},\end{cases}$$

其中,$D_i$ 表示第 $i$ 个疾病,$d_j$ 表示第 $j$ 个药物.

### 1.2.3　高斯相互作用谱核相似性

高斯相互作用谱核相似性用于评估分子间的结构相似性,通过对分子中原子特性进行编码,并利用高斯分布函数模拟原子间相互作用,计算其核函数值以衡量相似度.因此,本研究采用该方法来描述微生物与药物之间的相似性.微生物 $m_i$ 与 $m_j$ 的高斯相互作用谱核相似度计算如下:$K_{\text{gip},m}(m_i,m_j)=\exp(-r_m\|\boldsymbol{A}_{mi}-\boldsymbol{A}_{mj}\|^2)$,式中,$r_m$ 表示内核带宽,计算公式为:$r_m=r'_m/(\frac{1}{n}\sum_{i=1}^{n_m}\|\boldsymbol{A}_{mi}\|^2)$,其中 $n_m$ 为微生物总数,$r'_m$ 为归一化常数,根据以往研究[19],将其设为 1.同样,可以根据下式得到药物的高斯相互作用谱核相似度:$K_{\text{gip},d}(d_i,d_j)=\exp(-r_d\|\boldsymbol{A}_{di}-\boldsymbol{A}_{dj}\|^2)$,$r_d=r'_d/(\frac{1}{n}\sum_{i=1}^{n_d}\|\boldsymbol{A}_{di}\|^2)$,其中,$n_d$ 表示药物总数,归一化常数 $r'_d$ 设为 1.

### 1.2.4　余弦相似性

微生物余弦相似度原理是基于假设微生物 $i$ 和微生物 $j$ 彼此相似,这是一种常用的相似度计算方法.那么二进制向量 $\boldsymbol{A}_{MD}(i,:)$ 和 $\boldsymbol{A}_{MD}(j,:)$ 也应该彼此相似.同样下面以微生物以及药物余弦相似度为例.根据已知的微生物-药物关系数据,计算微生物的余弦相似度为:

$$C(i,j)=\frac{\boldsymbol{A}_{MD}(i,:)\cdot\boldsymbol{A}_{MD}(j,:)}{\|\boldsymbol{A}_{MD}(i,:)\|\ \|\boldsymbol{A}_{MD}(j,:)\|},$$

其中,$\boldsymbol{A}_{MD}(i,:)$为微生物与药物邻接矩阵中的第$i$行向量,表示微生物$i$的关系特征;$\boldsymbol{A}_{MD}(j,:)$为微生物与药物邻接矩阵$\boldsymbol{A}(i,j)$中的第$j$行向量,表示微生物$j$的关系特征;$\boldsymbol{A}_{MD}(i,:)\cdot\boldsymbol{A}_{MD}(j,:)$表示两个行向量的点积.

与微生物余弦相似度计算方法类似,药物的余弦相似度计算如下:

$$D(i,j)=\frac{\boldsymbol{A}_{MD}(:,i)\cdot\boldsymbol{A}_{MD}(:,j)}{\|\boldsymbol{A}_{MD}(:,i)\|\ \|\boldsymbol{A}_{MD}(:,j)\|},$$

其中,$\boldsymbol{A}_{MD}(:,i)$为微生物与药物邻接矩阵中的第$i$列向量,表示药物$i$的关系特征;$\boldsymbol{A}_{MD}(:,j)$为微生物与药物邻接矩阵$\boldsymbol{A}(i,j)$中的第$j$列向量,表示微生物$j$的关系特征;$\boldsymbol{A}_{MD}(:,i)\cdot\boldsymbol{A}_{(MD)}(:,j)$表示两个列向量的点积.

**1.2.5** 多源特征融合

多源相似度的有效融合也是本文应用深度学习方法的一个重要任务.据估计,特征融合可以产生更重要的特征,全面捕捉微生物和药物的特征.

对于微生物的相似度,将微生物高斯相互作用谱核相似度和微生物余弦相似度结合起来,形成微生物相似度$MM$.故微生物融合相似度计算公式如下:

$$M(m_i,m_j)=\begin{cases}\dfrac{K_{\text{gip},m}(m_i,m_j)+C(i,j)}{2}, C(i,j)\neq 0,\\ K_{\text{gip},m},\text{否则},\end{cases}$$

同理,对于药物的融合相似度,计算如下:

$$D(d_i,d_j)=\begin{cases}\dfrac{K_{\text{gip},d}(d_i,d_j)+D(i,j)}{2}, D(i,j)\neq 0,\\ K_{\text{gip},D},\text{否则}.\end{cases}$$

**1.3 方法概述**

该模型采用多任务学习框架,旨在同时预测微生物-药物和疾病-药物关联,以挖掘三者间的复杂关系.通过构建微生物-药物和疾病-药物超图,利用超图神经网络(HGNN)提取高阶特征.模型引入交叉压缩机制,实现任务间特征共享与融合,特别是通过药物节点的信息共享增强任务互补性.解码阶段使用双线性解码器,捕捉更复杂的二阶交互关系,并生成关联预测分数矩阵.整体通过联合二元交叉熵损失进行优化.

**1.3.1** 网络一致性投影

在两个子网络中,节点拥有不同维数的特征向量.为了在后续步骤中简化计算过程,使用网络一致性投影,将不同的节点特征统一到一个公共向量空间中.例如,在微生物与药物子网中,利用过渡矩阵将微生物和药物节点特征映射到统一的维度空间,具体为:$H_m(i)=M(i)\cdot\boldsymbol{W}_m$,$H_{d1}(i)=D_1(i)\cdot\boldsymbol{W}_d$,$H_D(i)=D_D\cdot\boldsymbol{W}_D$,$H_{d2}(i)=D_2(i)\cdot\boldsymbol{W}_d$,其中,$H_m(i)\in\mathbf{R}^{1\,024}$,$H_{d1}(i)\in\mathbf{R}^{1\,024}$为微生物-药物网络中微生物节点$m_i$和药物节点$d_i$的投影特征.同样,$H_D(i)\in\mathbf{R}^{1\,024}$和$H_{d2}(i)\in\mathbf{R}^{1\,024}$是疾病$D_i$和药物$d_i$在疾病-药物网络中的投影特征.将特征投影到1 024维空间,主要是为了提升模型的表达和学习能力.高维空间有助于捕捉微生物与药物之间复杂、非线性的关系,从而增强模型的泛化能力和预测准确性,特别适用于处理生物医学中的复杂关联数据.根据所设计的空间向量的大小要求,将学习权矩阵设置为$\boldsymbol{W}_m\in\mathbf{R}^{226\times 1\,024}$,$\boldsymbol{W}_D\in\mathbf{R}^{2\,667\times 1\,024}$和$\boldsymbol{W}_d\in\mathbf{R}^{446\times 1\,024}$.为了减少实验中的冗余参数和学习时间,这里使用权值矩阵来共享,完成两个网络中将药物节点映射到向量空间的任务.

**1.3.2** 多任务学习

多任务学习(multi-task learning,MTL)是一种机器学习方法,旨在通过同时学习多个相关任务来改善整体学习性能.在多任务学习中,不同任务之间可以是相关联的,通过共享信息和特征,可以提高模型的泛化能力、减少过拟合风险、提升数据效率,并且可以通过任务之间的相互促进来提高模型的性能.

为增强多任务间的信息流动与共享,采用交叉压缩方法.该方法通过融合不同任务的特征,实现细粒度

的信息共享，提升协同学习能力并缓解数据稀疏问题.同时，它能挖掘微生物、药物和疾病间的高阶关联，在共享信息的同时保留各任务的特异性，提升整体模型性能.具体的，通过交叉压缩单元模块连接两个子网，并通过分析 $\boldsymbol{MD}$ 和 $\boldsymbol{DD}$ 矩阵同时从两个子网中提取辅助信息.$\boldsymbol{H}_{aux-m} \in \mathbf{R}^{226\times 1\,024}$ 和 $\boldsymbol{H}_{aux-d1} \in \mathbf{R}^{446\times 1\,024}$ 分别代表微生物-药物网络中的微生物和药物节点，它们从自身网络和疾病-药物网络中获得辅助信息：$\boldsymbol{H}_{aux-m}=\boldsymbol{MD}\cdot\boldsymbol{W}_{aux-m}$，$\boldsymbol{H}_{aux-d1}=\boldsymbol{DD}^{\mathrm{T}}\cdot\boldsymbol{W}_{aux-d1}$.其中，$\boldsymbol{H}_{aux-m} \in \mathbf{R}^{226\times 1\,024}$ 和 $\boldsymbol{H}_{aux-d1} \in \mathbf{R}^{2\,667\times 1\,024}$ 为两个权重矩阵.类似的过程在疾病 - 药物网络中发生如下：$\boldsymbol{H}_{aux-D}=\boldsymbol{DD}\cdot\boldsymbol{W}_{aux-D}$，$\boldsymbol{H}_{aux-d2}=\boldsymbol{MD}^{\mathrm{T}}\cdot\boldsymbol{W}_{aux-d2}$.

最后，将节点的初始特征与辅助特征连接起来，形成节点的新特征.可以总结如下：$\boldsymbol{H}_M=\mathrm{cat}(\boldsymbol{H}_m,\boldsymbol{H}_{aux-m})$，$\boldsymbol{H}_{D1}=\mathrm{cat}(\boldsymbol{H}_{d1},\boldsymbol{H}_{aux-d1})$，$\boldsymbol{H}_D=\mathrm{cat}(\boldsymbol{H}_D,\boldsymbol{H}_{aux-D})$，$\boldsymbol{H}_{D2}=\mathrm{cat}(\boldsymbol{H}_{d2},\boldsymbol{H}_{aux-d2})$，其中 $\boldsymbol{H}_{D1} \in \mathbf{R}^{446\times 2\,048}$，$\boldsymbol{H}_{D1} \in \mathbf{R}^{446\times 2\,048}$ 为微生物-药物网络中节点的综合特征表示，$\boldsymbol{H}_D \in \mathbf{R}^{2\,667\times 2\,048}$，$\boldsymbol{H}_{D2} \in \mathbf{R}^{446\times 2\,048}$ 分别为疾病 - 药物子网络中药物和疾病节点的综合特征表示.

### 1.3.3　超图神经网络

超图神经网络(hypergraph neural networks，HGNN)是一种特殊类型的图神经网络(GNN)，它专门设计来处理超图数据[20].而超图是一种比传统图更一般化的数据结构，在超图中，一条超边可以连接两个以上的顶点，这使得超图能够自然地表示更复杂的多对多关系.HGNN 的核心思想在于利用超图结构来表示和学习数据中的复杂多对多关系，通过特定的超图卷积和特征聚合机制，为顶点学习富有表现力的特征表示，以支持各种下游任务.具体框架如附录图 S1 所示.在 MTLTPMDA 中，进一步利用两个子网络节点在各自网络中的强邻居信息，基于 HGNN 编码器获得两个子网络节点的表示.

具体而言，上文已经给出微生物-药物关系矩阵，$\boldsymbol{A}(i,j)=1$ 若 $\boldsymbol{A}(i,j)=1$，则表示药物 $d_j$ 对微生物 $m_i$ 具有作用；同理药物 - 疾病关系矩阵 $\boldsymbol{B}(i,j)=1$，若 $\boldsymbol{B}(i,j)=1$，则表示药物 $d_j$ 可用于治疗疾病 $\boldsymbol{D}_i$.在超图中，微生物、药物和疾病均作为节点存在，而每个药物节点 $d_j$ 会生成一个超边 $e_i$，连接到所有与其相关的微生物和疾病节点.通过这种方式，超边可以捕捉到药物与多个微生物及疾病间的多对多关联关系.例如，若药物 $d_1$ 关联微生物 $m_1$ 和疾病 $\boldsymbol{D}_1$，则超边 $e_1$ 链接 $d_1$ 与 $m_1$，$\boldsymbol{D}_1$.

为超边分配权重以表征超边中关联强度，在这设为 1 表示等权重.

接下来，构建节点度矩阵和超边度矩阵.在超图中，节点度矩阵 $\boldsymbol{D}_v$ 是一个对角矩阵，用于表示每个节点在超图中连接的超边数量，即每个节点的度数.具体来说，超图包含 3 339 个节点和 18 757 条超边，节点度矩 $\boldsymbol{D}_v \in \mathbf{R}^{3\,339\times 3\,339}$ 的第 $i$ 个对角元素 $\boldsymbol{D}_v(i,i)$ 表示第 $i$ 个节点的度，即节点 $v_i$ 所连接的超边的总数.可以通过节点-超边关联矩阵 $\boldsymbol{H}$ 计算得到 $\boldsymbol{D}_v=\mathrm{diag}(\boldsymbol{HWH}^{\mathrm{T}}\mathbf{1})$.其中，$\boldsymbol{H} \in \mathbf{R}^{3\,339\times 18\,757}$ 是节点-超边关联矩阵，如果节点 $v_i$ 属于超边 $e_j$，则 $\boldsymbol{H}_{ij}=1$，否则 $\boldsymbol{H}_{ij}=0$；$\boldsymbol{D} \in \mathbf{R}^{18\,757\times 18\,757}$ 是超边的权重矩阵，$\mathbf{1}$ 是一个列向量，长度为 18 757，用于将 $\boldsymbol{HWH}^{\mathrm{T}}$ 矩阵每一列进行求和.超边度矩阵 $\boldsymbol{D}_e$ 同样是一个对角矩阵，用于表示每条超边中连接的节点数量，即每条超边的度数.超图中存在 18 757 条超边，则超边度矩阵 $\boldsymbol{D}_e \in \mathbf{R}^{18\,757\times 18\,757}$ 的 $j$ 个对角元素 $\boldsymbol{D}_e(j,j)$ 表示超边 $e_j$ 的度，即该超边所连接的节点总数.可以通过节点-超边关联矩阵 $\boldsymbol{H}$ 计算得到：$\boldsymbol{D}_e=\mathrm{diag}(\boldsymbol{H}^{\mathrm{T}}\mathbf{1})$.其中 $H^{\mathrm{T}}$ 是 $H$ 的转置，$\mathbf{1}$ 是一个列向量，长度为 3 339.$\boldsymbol{H}^{\mathrm{T}}\mathbf{1}$ 会计算出每条超边连接的节点总数，从而得到超边的度数.

在超图神经网络模型中，通过超图卷积来实现特征的聚合更新.给定初始节点特征矩阵 $\boldsymbol{X} \in \mathbf{R}^{3\,339\times 2\,048}$，其中 2 048 为特征维度，超图卷积操作可以表述如下：$\boldsymbol{X}^{(l+1)}=\sigma(\boldsymbol{D}_v^{-1}\boldsymbol{HWD}_e^{-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{X}^{(l)}\Theta^{(l)})$，其中，$\boldsymbol{X}^{(l)}$ 是第 $l$ 层的节点特征矩阵；$\sigma$ 为激活函数；$\Theta^{(l)}$ 为第 $l$ 层的可学习权重参数；$\boldsymbol{D}_v^{-1}$ 和 $\boldsymbol{D}_e^{-1}$ 节点和超边的度进行归一化，确保不同节点间特征聚合不受其度的影响.

通过超图卷积，微生物、药物和疾病节点间的信息得以传播和聚合，从而捕获数据中蕴含的高阶结构关系.

### 1.3.4　双线性解码器

双线性解码通常用于学习节点之间的关系或边的表示.它通过将两个节点的特征进行双线性变换，从而捕捉节点之间的复杂关系.相比一般的线性解码器，双线性解码器采用双线性变换操作，在处理图数据中的关系建模任务时具有更强的表达能力，能够更好地捕捉节点之间的复杂关系.

在 MTLTPMDA 中获得两个子网络的节点表示后,使用线性解码器重构两个子网络中异构图的链路,具体为:$\hat{y}_{md}=\mathrm{Sigmoid}((\boldsymbol{F}_{m(i)})^{\mathrm{T}}\boldsymbol{Q}_1\boldsymbol{F}_{d1(j)})$,$\hat{y}_{dd}=\mathrm{Sigmoid}((\boldsymbol{F}_{g(i)})^{\mathrm{T}}\boldsymbol{Q}_2\boldsymbol{F}_{d2(j)})$.其中,$\hat{y}_{md}$ 表示微生物-药物子网络中微生物节点 $m(i)$ 和药物节点 $d(j)$ 的预测关联概率,$\boldsymbol{F}_m$ 和 $\boldsymbol{F}_{d1}$ 分别表示通过 MTLTPMDA 编码器在微生物-药物子网络中获得的最终微生物-药物节点嵌入表示,$\boldsymbol{Q}_1$ 表示可训练参数矩阵,该矩阵为 64 * 64 维:$\mathrm{Sigmoid}(x)=\dfrac{1}{1+\mathrm{e}^{-x}}$.同理,$\hat{y}_{dd}$ 表示疾病-药物子网中疾病节点与药物节点的预测关联概率.

### 1.3.5 模型训练

MTLTPMDA 模型的损失函数是两个子网中所有训练样本的重构误差之和.在这里,选择交叉熵损失函数来度量子网络中每个关联的真实值 $y$ 与预测概率值 $\hat{y}$ 之间的误差.具体如下:

$$L_{m-d}=-\sum y_{ij}\ln\hat{y}_{ij}+(1-y_{ij})\ln(1-\hat{y}_{ij}),$$

$$L_{d-d}=-\sum y_{ij}\ln\hat{y}_{ij}(1-y_{ij})\ln(1-\hat{y}_{ij}).$$

其中 $L_{m-d}$ 表示微生物-药物子网络中的功能损失,$\hat{y}_{ij}$ 表示微生物-药物节点之间的预测链接概率,由双线性解码器部分可知当 $m=i,d=j$ 时 $\hat{y}_{md}=\hat{y}_{ij}$,而 $y_{ij}$ 表示该连接的真实标签,在已知的微生物-药物关联矩阵 $\boldsymbol{A}$ 中,其值为 1 或 0.相应的,$L_{d-d}$ 表示疾病-药物子网络中的功能损失.

将两个子网的损耗之和作为 MTLTPMDA 的损耗,其形式为:$L=L_{d-d}+L_{m-d}$,然后,使用上述 Loss 函数,以端到端方式通过反向传播算法对整个模型进行训练.

## 2 实验结果与分析

### 2.1 评价指标

为了综合评价提出的 MTLTPMDA 的性能,选择 Precision($P$)、Accuracy($A$)、Recall($R$)、$F1$ 分数、AUC 和 Precision-Recall(P-R)曲线作为评价标准.相应的数学计算表示如下:

$$P=\frac{TP}{TP+FP},A=\frac{TP+TN}{TP+TN+FP+FN},R=\frac{TP}{TP+FN},F1=\frac{2TP}{2TP+FN+FP}.$$

其中 $TP$、$FP$、$TN$、$FN$ 分别为真阳性、假阳性、真阴性、假阴性.AUC 是指受试者工作特征(ROC)曲线下的面积,可以定量反映基于 ROC 曲线所测得的模型性能.ROC 曲线横坐标为 $FPR$,纵坐标为 $TPR$,其中 $TPR$ 和 $TPR$ 计算公式如下:

$$TPR=\frac{TP}{TP+FN},FPR=\frac{FP}{FP+TN},$$

P-R 曲线的横坐标表示模型的召回率,纵坐标表示精度.P-R 曲线面积越大,模型性能越好.附录表 S1 详细描述了本文的模型使用 5 倍交叉验证的各种评价指标的值.从表 S1 中方差普遍低于 0.01 可以看出模型的表现非常接近,实验结果的一致性和稳定性很高.

为提升实验稳健性,采用 2 折、5 折和 10 折交叉验证,以减少数据划分带来的影响.由图 1 显示,MTLTPMDA 模型在不同折数下的 ROC 和 P-R 曲线 AUC 均在 88%以上,变化幅度较小,表明模型具有良好的预测能力和鲁棒性.综合表现来看,5 折交叉验证效果最佳.附录表 S2 进一步验证了模型在 10 折交叉验证中的稳定性,各项指标方差均低于 0.001,说明模型在不同数据划分下表现一致.

为验证 MTLTPMDA 中引入疾病-药物信息的价值,设计了一项实验,将疾病-药物网络随机打乱,同时保持微生物-药物网络不变.5 折和 10 折交叉验证结果见附录表 S3、表 S4 所示.尽管打乱了原始关联,各项指标与未打乱时差距较小,说明即便在非特异性任务初期,两个网络在疾病节点的向量空间中仍可能表现出结构相似性[15],打乱后的疾病-药物网络依然能为微生物-药物关联预测提供辅助信息.

### 2.2 与其他最新方法的比较

为了评估模型的性能,将 MTLTPMDA 与其他关联预测模型进行比较.选取了该领域最新、最具代表性的模型,分别是 AGAEMD[21]、GATECDA[22]、KATZHMDA[23]、LAGCN[24].为了公平实验,5 种比较算法都在相同数据集上进行了 5 折交叉验证实验.图 2 显示了 MTLTPMDA 与其他 4 种对比算法的 ROC 曲线对

比以及每个模型的 AUC 值.与其他模型相比，MTLTPMDA 的 AUC 值分别高出 2%、4%、3%以及 6%左右，由此可以看出本文的模型在预测微生物与药物新关联方面有较好的效果.

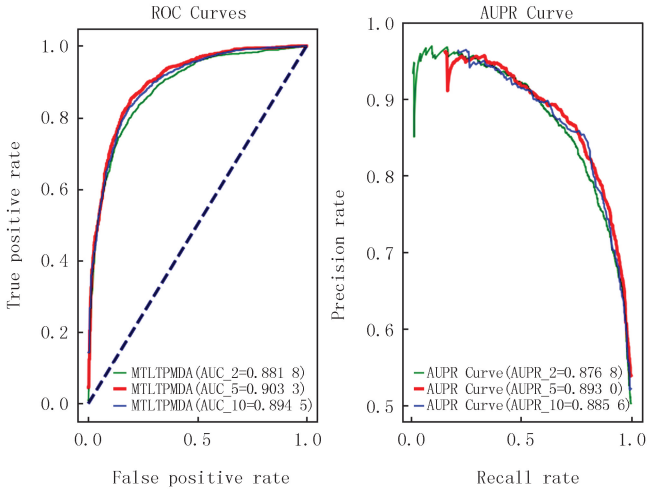



图1　MTLTPMDA的2折、5折及10折交叉验证ROC曲线以及P-R曲线

Fig. 1　The ROC curve and P-R curve of MTLTPMDA are verified by 2-fold cross-validation, 5-fold cross-validation and 10-fold cross-validation

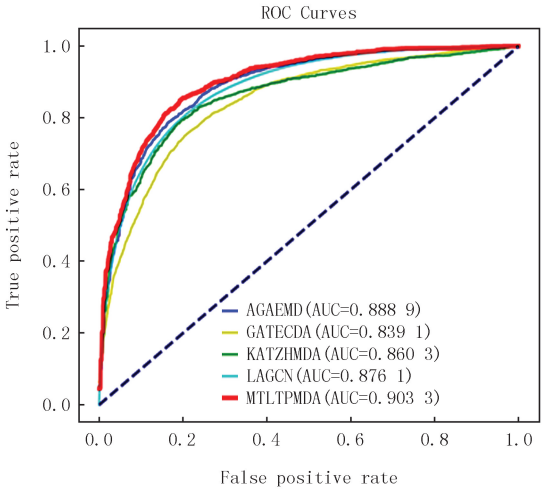



图2　不同模型的ROC曲线比较

Fig. 2　Comparison of ROC curves for different models

## 2.3　消融实验

为验证模型有效性，设计了 3 组消融实验：同时使用高斯相互作用谱核相似性和余弦相似性(Cosine+GIP)、仅使用高斯相互作用谱核相似性(GIP)、仅使用余弦相似性(Cosine).这些实验全面评估了不同相似性度量对模型性能的影响，揭示各组件的贡献与互补性，为后续优化提供依据.表 1 展示了各组合下的模型表现.

表 1 显示，单独使用高斯相互作用谱核相似性或余弦相似性性能均不及二者结合，且余弦相似性主要起辅助作用.表 2 对比了不使用交叉压缩的简单拼接法与使用交叉压缩法的模型性能，后者各项指标均提升超过 1.5 百分点，说明交叉压缩通过更深层次的特征交互，有效增强了任务间信息融合，提升了模型表现，验证了其在多任务学习中的优势.

**表 1　比较不同相似性特征下的结果**

**Tab. 1　Comparison of results under different similarity features**

| 方法 | AUC | ACC | Pre | F1 | Recall |
|---|---|---|---|---|---|
| GIP+Cosine | 90.33 | 81.78 | 82.60 | 82.80 | 83.96 |
| GIP | 90.24 | 81.23 | 82.46 | 82.22 | 82.07 |
| Cosine | 86.64 | 79.64 | 79.46 | 79.13 | 79.32 |

**表 2　比较交叉压缩和简单拼接下的结果**

**Tab. 2　Compare the results of cross-compression and simple concatenation**

| 方法 | AUC | ACC | Pre | F1 | Recall |
|---|---|---|---|---|---|
| 交叉压缩法 | 90.33 | 81.78 | 82.60 | 82.80 | 83.96 |
| 简单拼接 | 88.05 | 80.58 | 80.41 | 80.63 | 81.13 |

## 2.4　参数分析

学习率的大小是影响模型性能的重要因素.学习率设置过大容易导致参数更新过快，错过最优解；而学习率设置过小会导致参数更新过慢，训练过程变得冗长.因此，合适的学习率是很关键的.在实验中，学习率 $lr\in\{5\times10^{-5},10^{-4},5\times10^{-4},10^{-3},5\times10^{-3}\}$.实验结果表明，当学习率达到 $5\times10^{-4}$ 时，模型的性能最好.因此，选择 $5\times10^{-4}$ 作为 MTLTPMDA 编码器的默认学习率.

权重衰减通过在损失函数中加入正则项，抑制权重大小，减少过拟合，提高模型泛化能力.实验中，测试了不同权重衰减对模型性能的影响.取权重衰减值 $w_d\in\{3\times10^{-3},3\times10^{-4},10^{-4},10^{-5}\}$.实验结果表明，当权重衰减数为 $10^{-5}$ 时，该模型的整体性能最优.因此，在后续的实验中，选择 $10^{-5}$ 作为权重衰减的默认值.

HGNN 卷积层用于通过聚合邻居节点特征来学习图中节点表示.实验中，测试了 1 到 5 层 HGNN 的性能，结果显示 3 层时模型表现最佳，因此后续实验默认使用 3 层.

## 3 案例分析

为验证 MTLTPMDA 预测潜在微生物-药物关联的能力,选取 HIV 和姜黄素进行案例研究.对每个微生物,按预测得分排序药物并验证前 10 名;对每个药物,同样排序微生物并验证前 10 名.

HIV(human immunodeficiency virus)学名人类免疫缺陷病毒,是一种感染人类免疫系统的病毒,它会损害人体的免疫系统,使其无法有效应对感染[25].吐根生物碱是药用植物吐根草中产生的次生代谢产物.吐根碱是吐根的主要生物碱,也是吐根糖浆中具有催吐作用的活性成分之一.文献[26]分析了 HIV-1 RT 上的 Emetine 分子对接,其研究结果表明吐根碱能够穿透完整的 HIV 颗粒,并结合和阻断逆转录反应,表明其可以用作抗 HIV 杀微生物剂.对 HIV 的预测结果如表 3 所示,前 10 名候选药物中有 9 个得到确认,高比例的正确预测表明模型在处理类似任务时表现出稳定性和鲁棒性,预测结果能够被较大程度信赖.

Curcumin 学名姜黄素,源自姜黄素(姜黄香料)的根,是姜科植物和热带植物的一员.这种植物化学物质的功效已被证明可以对抗各种人类疾病[27],IZUI 等[28]主要探讨姜黄素对牙周病细菌,特别是牙龈卟啉单胞菌的抑菌作用,最终他们得出结论:姜黄素对牙周病细菌具有抗菌作用,可能是预防牙周病的有效药物.虽然姜黄素因其广泛的治疗和预防应用而获得了巨大的重要性,但是,MARATHE 等[29]研究发现,它可以调节肠沙门氏菌血清型鼠伤寒沙门氏菌的防御途径,增强其致病性.这一结果促使我们需要重新考虑姜黄素的滥用,特别是在沙门氏菌爆发期间.姜黄素的预测结果如表 4 所示,前 10 名候选药物中有 8 个得到确认,高比例的正确预测表明模型在处理类似任务时表现出稳定性和鲁棒性,预测结果能够被较大程度信赖.

**表 3 MTLTPMDA 鉴定 HIV 前 10 名**

**Tab. 3 Top 10 HIV identifications by MTLTPMDA**

| 排名 | 药物名 | PMID |
|---|---|---|
| 1 | chloroquine | 11166661 |
| 2 | ivermectin | 9300635 |
| 3 | gemcitabine | 23994876 |
| 4 | memantine | 1620355 |
| 5 | emetine | 36139404 |
| 6 | saquinavir | 35883499 |
| 7 | atazanavir | 15482137 |
| 8 | zidovudine | 2500164 |
| 9 | thymalfasin | Unconfirmed |
| 10 | rilpivirine | 10848067 |

**表 4 MTLTPMDA 鉴定 Curcumin 前 10 名**

**Tab. 4 Top 10 Curcuminidentifications by MTLTPMDA**

| 排名 | 微生物名 | PMID |
|---|---|---|
| 1 | Candida albicans | 33592455 |
| 2 | Enterococcus faecalis | 34320428 |
| 3 | Vibrio harveyi | 34371200 |
| 4 | Bacillus subtilis | 35456852 |
| 5 | Proteus mirabilis | 21808656 |
| 6 | Stenotrophomonas maltophilia | 33143940 |
| 7 | Eikenella corrodens | unconfirmed |
| 8 | Aeromonas hydrophila | 32947883 |
| 9 | Vibrio anguillarum | unconfirmed |
| 10 | Vibrio vulnificus | 22092562 |

## 4 结 论

人类的各种恶性疾病都与人体微生物有着很大的关联,寻找到针对人体恶性疾病的有效药物,就可以找出与药物相关的微生物,进而从根源上抑制疾病的发生.因此,准确预测微生物与药物的关系可以促进人类健康的进步.在本文中,提出了多任务学习模型(即 MTLTPMDA)来预测潜在的微生物-药物关联.该方法的新颖之处在于,根据微生物-药物网络中的药物,构建相应的疾病-药物子网络,辅助微生物-药物关联预测任务,同时运用超图神经网络挖掘两个子网的深层特征表示,挖掘微生物与药物之间的复杂相互作用.与 4 种最新的经典基准模型相比,提出的 MTLTPMDA 模型获得了更优的 AUC.

此外,通过 HIV、SARS-COV、咖啡因和姜黄素这 4 个常见微生物和药物的研究,证实了 MTLTPMDA 模型在预测过程中的准确性和可靠性.

然而,MTLTPMD 仍然具有一些局限性,HGNN 方法由于其复杂性,特别是在处理大规模超图时,可能需要大量的计算资源,包括内存和处理时间.

**附录见电子版(DOI:10.16366/j.cnki.1000-2367.2024.10.08.0001).**

## 参 考 文 献


[1] MARCOS-ZAMBRANO L J,KARADUZOVIC-HADZIABDIC K,LONCAR TURUKALO T,et al.Applications of machine learning in human microbiome studies:a review on feature selection,biomarker identification,disease prediction and treatment[J].Frontiers in Microbiology,2021,12:634511.

[2] HUANG Y,YOU Z H,CHEN X,et al.Prediction of microbe-disease association from the integration of neighbor and graph with collaborative recommendation model[J].Journal of Translational Medicine,2017,15(1):209.

[3] 邵孟，郭苗苗，邱文娜，等.2021—2023 年苏州市某儿童医院临床沙门氏菌血清型和耐药性分析[J].中华医院感染学杂志，2025，35(11)：1679-1683.
SHAO M,GUO M M,QIU W N,et al.Analysis of serotypes and antibiotic resistance of clinically isolated Salmonella in a children's hospital in Suzhou,2021-2023[J].Chinese Journal of Nosocomiology,2025,35(11):1679-1683.

[4] 刘晓雯，牟小琴，程创，等.Ebselen 抑制巨噬细胞过度炎性反应抗大肠杆菌急性感染[J].中国药理学通报，2025，41(7)：1346-1353.
LIU X W,MOU X Q,CHENG C,et al.Inhibition of excessive inflammatory response of macrophages by Ebselen against acute Escherichia coli infection[J].Chinese Pharmacological Bulletin,2025,41(7):1346-1353.

[5] KUMAR A,CHORDIA N.Role of microbes in human health[J].Applied Microbiology:Open Access,2017,3(2): 1-3..

[6] QADRI H.Novel strategies to combat the emerging drug resistance in human pathogenic microbes[J].Current Drug Targets,2021,22(12):1424-1436.

[7] KAPOOR R,SAINI A,SHARMA D.Indispensable role of microbes in anticancer drugs and discovery trends[J].Applied Microbiology and Biotechnology,2022,106(13):4885-4906.

[8] LONG Y H,WU M,KWOH C K,et al.Predicting human microbe-drug associations via graph convolutional network with conditional random field[J].Bioinformatics,2020,36(19):4918-4927.

[9] LONG Y H,WU M,LIU Y,et al.Ensembling graph attention networks for human microbe-drug association prediction[J].Bioinformatics,2020,36(Supplement_2):i779-i786.

[10] LONG Y H,LUO J W.Association mining to identify microbe drug interactions based on heterogeneous network embedding representation[J].IEEE Journal of Biomedical and Health Informatics,2020,25(1):266-275.

[11] DENG L,HUANG Y B,LIU X J,et al.Graph2MDA:a multi-modal variational graph embedding model for predicting microbe-drug associations[J].Bioinformatics, 2022,38(4): 1118-1125.

[12] MA Y J,LIU Q Q.Generalized matrix factorization based on weighted hypergraph learning for microbe-drug association prediction[J].Computers in Biology and Medicine,2022, 145:105503.

[13] LONG M, CAO Z, WANG J, et al. Learning multiple tasks with multilinear relationship networks [EB/OL].[2024-09-17].https://doi.org/10.48550/arXiv.1506.02117.

[14] ZHANG Y,YANG Q.A survey on multi-task learning[J].IEEE Transactions on Knowledge and Data Engineering,2022,34(12):5586-5609.

[15] SUN Y Z,ZHANG D H,CAI S B,et al.MDAD:a special resource for microbe-drug associations[J].Frontiers in Cellular and Infection Microbiology,2018,8:424.

[16] RAJPUT A,THAKUR A,SHARMA S,et al.aBiofilm:a resource of anti-biofilm agents and their potential implications in targeting antibiotic drug resistance[J].Nucleic Acids Research,2018,46(D1):D894-D900.

[17] ANDERSEN P I,IANEVSKI A,LYSVAND H,et al.Discovery and development of safe-in-man broad-spectrum antiviral agents[J].International Journal of Infectious Diseases,2020,93:268-276.

[18] WANG L,TAN Y Q,YANG X Y,et al.Review on predicting pairwise relationships between human microbes,drugs and diseases:from biological data to computational models[J].Briefings in Bioinformatics,2022,23(3):bbac080.

[19] VAN LAARHOVEN T,NABUURS S B,MARCHIORI E.Gaussian interaction profile kernels for predicting drug-target interaction[J].Bioinformatics,2011,27(21):3036-3043.

[20] FENG Y F,YOU H X,ZHANG Z Z,et al.Hypergraph neural networks[J].Proceedings of the AAAI Conference on Artificial Intelligence,2019,33(1):3558-3565.

[21] ZHANG H Z,FANG J T,SUN Y P,et al.Predicting miRNA-disease associations via node-level attention graph auto-encoder[J].IEEE/ACM Transactions on Computational Biology and Bioinformatics,2022,20(2):1308-1318.

[22] DENG L,LIU Z X,QIAN Y R,et al.Predicting circRNA-drug sensitivity associations via graph attention auto-encoder[J].BMC Bioinformatics,2022,23(1):160.

[23] CHEN X,HUANG Y,YOU Z H,et al.A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases[J].Bioinformatics,2017,33(5):733-739.

[24] YU Z X,HUANG F,ZHAO X H,et al.Predicting drug-disease associations through layer attention graph convolutional network[J].Briefings in Bioinformatics,2021,22(4):bbaa243.

[25] LI G D,DE CLERCQ E.HIV genome-wide protein associations:a review of 30 years of research[J].Microbiology and Molecular Biology Reviews,2016,80(3):679-731.

[26] VALAD? O A,ABREU C,DIAS J,et al.Natural plant alkaloid (emetine) inhibits HIV-1 replication by interfering with reverse transcriptase activity[J].Molecules,2015,20(6):11474-11489.

[27] OSKOUIE M N,AGHILI MOGHADDAM N S,BUTLER A E,et al.Therapeutic use of curcumin-encapsulated and curcumin-primed exosomes[J].Journal of Cellular Physiology,2019,234(6):8182-8191.

[28] IZUI S,SEKINE S,MAEDA K,et al.Antibacterial activity of curcumin against periodontopathic bacteria[J].Journal of Periodontology,2016,87(1):83-90.

[29] MARATHE S A,RAY S,CHAKRAVORTTY D.Curcumin increases the pathogenicity of Salmonella enterica serovar typhimurium in murine model[J].PLoS One,2010,5(7):e11511.

# Predicting microbe-drug associations based on multi-task learning and hypergraph neural network

Wang Bo, Wang Junqi, Du Xiaoxin, Sun Ming, Wang Tongxuan, Li Jingwei

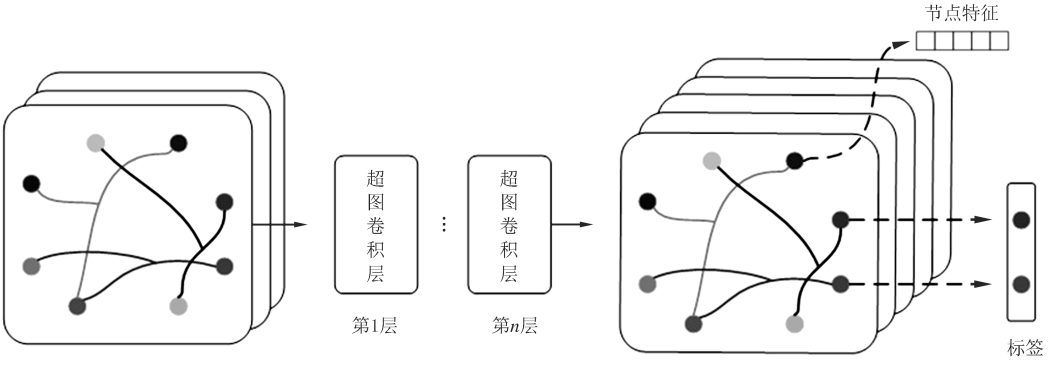
(College of Computer and Control Engineering;Heilongjiang Key Laboratory of Big Data Network Security Detection and Analysis, Qiqihar University, Qiqihar 161006, China)

**Abstract**: Traditional biological experiments to discover microbe-drug relationships are not only time-consuming and labor-intensive but also highly expensive. Therefore, to reduce experimental costs and improve efficiency, computational methods have been employed to predict microbe-drug associations. However, the existing methods neglect the crucial role of diseases as intermediaries, which leads to the problem of data sparsity. To address this, we propose a multi-task learning model(MTLTPMDA) that simultaneously predicts microbe-drug and disease-drug associations. The model enhances the connections between tasks by sharing drug node features and utilizes a hypergraph neural network(HGNN) to explore the complex interactions between microbes, drugs, and diseases. By constructing microbe-drug and disease-drug hypergraphs, the HGNN effectively captures higher-order relationships among multiple nodes. In a five-fold cross-validation framework, MTLTPMDA achieved an AUC of 0.903 3 and an AUPR of 0.893 0, outperforming several existing methods, demonstrating the model's effectiveness in predicting potential associations.

**Keywords**: microbe-drug associations; disease-drug associations; multi-task learning technology; data sparsity; hypergraph neural network

[责任编校 陈留院 杨浦]

## 附　录




图S1　HGNN结构图

Fig.S1　HGNN architecture diagram

**表 S1　5 折交叉验证**

**Tab. S1　5-fold cross-validation**

| 测试集 | 精确度 | 准确度 | 召回率 | F1 得分 | AUC 值 | AUPR 值 |
|---|---|---|---|---|---|---|
| 1 | 0.82 | 0.83 | 0.85 | 0.84 | 0.90 | 0.90 |
| 2 | 0.75 | 0.80 | 0.90 | 0.81 | 0.89 | 0.87 |
| 3 | 0.83 | 0.86 | 0.90 | 0.86 | 0.92 | 0.91 |
| 4 | 0.84 | 0.84 | 0.84 | 0.84 | 0.92 | 0.90 |
| 5 | 0.82 | 0.80 | 0.79 | 0.80 | 0.89 | 0.88 |
| 平均值 | 0.81 | 0.83 | 0.85 | 0.83 | 0.90 | 0.89 |
| 方差 | 0.001 27 | 0.000 68 | 0.002 13 | 0.000 6 | 0.000 23 | 0.000 27 |

**表 S2　10 折交叉验证**

**Tab. S2　10-fold cross-validation**

| 测试集 | 精密度 | 准确度 | 召回率 | F1 得分 | AUC 值 | AUPR 值 |
|---|---|---|---|---|---|---|
| 1 | 0.80 | 0.81 | 0.77 | 0.79 | 0.89 | 0.88 |
| 2 | 0.77 | 0.85 | 0.70 | 0.77 | 0.90 | 0.92 |
| 3 | 0.81 | 0.85 | 0.76 | 0.82 | 0.87 | 0.84 |
| 4 | 0.80 | 0.75 | 0.84 | 0.80 | 0.89 | 0.87 |
| 5 | 0.81 | 0.76 | 0.92 | 0.83 | 0.90 | 0.90 |
| 6 | 0.83 | 0.86 | 0.79 | 0.82 | 0.90 | 0.89 |
| 7 | 0.82 | 0.81 | 0.81 | 0.81 | 0.92 | 0.90 |
| 8 | 0.84 | 0.83 | 0.87 | 0.85 | 0.91 | 0.90 |
| 9 | 0.81 | 0.85 | 0.76 | 0.80 | 0.89 | 0.91 |
| 10 | 0.77 | 0.81 | 0.75 | 0.78 | 0.85 | 0.85 |
| 平均值 | 0.81 | 0.82 | 0.80 | 0.80 | 0.89 | 0.89 |
| 方差 | 0.000 52 | 0.001 46 | 0.004 18 | 0.000 58 | 0.000 40 | 0.000 67 |

**表 S3　5 折交叉验证(将疾病-药物关联随机打乱)**

**Tab. S3　5-fold cross-validation(with randomized disease-drug associations)**

| 测试集 | 精密度 | 准确度 | 召回率 | F1 得分 | AUC 值 |
|---|---|---|---|---|---|
| 1 | 0.82 | 0.82 | 0.83 | 0.82 | 0.89 |
| 2 | 0.79 | 0.81 | 0.82 | 0.81 | 0.89 |
| 3 | 0.81 | 0.85 | 0.91 | 0.86 | 0.92 |
| 4 | 0.85 | 0.84 | 0.83 | 0.84 | 0.92 |
| 5 | 0.82 | 0.80 | 0.80 | 0.81 | 0.89 |
| 平均得分 | 0.82 | 0.83 | 0.84 | 0.83 | 0.90 |
| 方差 | 0.000 47 | 0.000 43 | 0.001 77 | 0.000 47 | 0.000 27 |

**表 S4　10 折交叉验证(将疾病-药物关联随机打乱)**

**Tab. S4　10-fold cross-validation(with randomized disease-drug associations)**

| 测试集 | 精密度 | 准确度 | 召回率 | F1 得分 | AUC 值 |
|---|---|---|---|---|---|
| 1 | 0.81 | 0.83 | 0.76 | 0.79 | 0.89 |
| 2 | 0.80 | 0.76 | 0.90 | 0.83 | 0.90 |
| 3 | 0.81 | 0.83 | 0.78 | 0.80 | 0.87 |
| 4 | 0.80 | 0.80 | 0.74 | 0.77 | 0.89 |
| 5 | 0.81 | 0.76 | 0.91 | 0.83 | 0.91 |
| 6 | 0.83 | 0.84 | 0.80 | 0.82 | 0.91 |
| 7 | 0.83 | 0.84 | 0.79 | 0.81 | 0.92 |
| 8 | 0.83 | 0.83 | 0.88 | 0.84 | 0.92 |
| 9 | 0.80 | 0.85 | 0.78 | 0.80 | 0.89 |
| 10 | 0.76 | 0.84 | 0.68 | 0.75 | 0.86 |
| 平均值 | 0.81 | 0.82 | 0.80 | 0.80 | 0.90 |
| 方差 | 0.000 44 | 0.001 11 | 0.005 44 | 0.000 80 | 0.000 40 |