



天津科技大学学报

Journal of Tianjin University of Science & Technology

ISSN 1672-6510, CN 12-1355/N

## 《天津科技大学学报》网络首发论文

题目：基于精简可信超图的中药处方筛选方法  
作者：赵婷婷，张碧霞，李金羽，史浩天，陈亚瑞，庄朋伟  
DOI：10.13364/j.issn.1672-6510.20250162  
收稿日期：2025-10-27  
网络首发日期：2026-09-03  
引用格式：赵婷婷，张碧霞，李金羽，史浩天，陈亚瑞，庄朋伟. 基于精简可信超图的中药处方筛选方法[J/OL]. 天津科技大学学报.  
<https://doi.org/10.13364/j.issn.1672-6510.20250162>



**网络首发：**在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

**出版确认：**纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。



## 基于精简可信超图的中药处方筛选方法

赵婷婷<sup>1</sup>, 张碧霞<sup>1</sup>, 李金羽<sup>1</sup>, 史浩天<sup>1</sup>, 陈亚瑞<sup>1</sup>, 庄朋伟<sup>2,3</sup>

(1. 天津科技大学人工智能学院, 天津 300457; 2. 天津中医药大学第一附属医院, 天津 300193;  
3. 国家中医针灸临床科学研究中心, 天津 300193)

**摘要:** 在中药配伍研究中, 由于中药种类繁多, 现有基于临床经验的处方策略未能充分挖掘潜在的高效药物组合。研究者引入基于蛋白质相互作用网络距离的药效评价方法, 然而该方法因计算复杂度高和可解释性差限制了大规模筛选的能力。本文提出一种高效且具有可解释性的基于精简可信超图的中药处方筛选方法, 即靶点特异性中药(target-specificity traditional Chinese medicine, TS-TCM) 框架。通过功能富集获取在特定疾病中显著表达的靶点, 采用图卷积网络生成靶点社区; 为充分利用超图对高阶关系的表达能力, 设计扩展密度聚类算法生成精简超图, 并基于表示差异的超图卷积网络获取可信超图; 通过捕获处方疗效差异与靶点微观结构之间的模式, 设计处方评价网络, 实现对中药处方疗效的准确预测。本文方法在帕金森病上的准确率为 88.9%, 在哮喘、中风疾病上的准确率均达到 100%。在执行效率上, 该方法在 0.5 s 内即可完成处方筛选任务, 显著提高了大规模中药处方筛选速度。

**关键词:** 中药; 蛋白质相互作用; 超图; 图卷积网络; 处方评价

**中图分类号:** TP391; R289 **文献标志码:** A **文章编号:** 1672-6510 (0000)00-0000-00

## Screening Method of Traditional Chinese Medicine Formulas Based on Simplified and Reliable Hypergraph

ZHAO Tingting<sup>1</sup>, ZHANG Bixia<sup>1</sup>, LI Jinyu<sup>1</sup>, SHI Haotian<sup>1</sup>, CHEN Yarui<sup>1</sup>, ZHUANG Pengwei<sup>2,3</sup>

(1. College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin 300457, China;

2. First Teaching Hospital of Tianjin University of Traditional Chinese Medicine, Tianjin, 300193, China;

3. China National Clinical Research Center for Chinese Medicine Acupuncture and Moxibustion, Tianjin 300193, China)

**Abstract:** In the research on the compatibility of traditional Chinese medicine (TCM), due to the wide variety of medicines, the existing prescription strategies based on clinical experience have not fully explored the potential efficient drug combinations. Therefor researchers introduced a drug efficacy evaluation method based on protein-protein interaction (PPI) network distance. However, it is limited in its ability to perform large-scale screening due to high computational complexity and poor interpretability. We propose an efficient and interpretable method, which is target-specificity traditional Chinese medicine(TS-TCM), for screening TCM prescriptions based on a streamlined and trusted hypergraph. Firstly, targets significantly expressed in specific diseases were obtained through functional enrichment, and a graph convolutional network was used to generate target communities. To fully utilize the expressive ability of hypergraphs for high-order relations, an extended density clustering algorithm was designed to generate a streamlined hypergraph, and a trusted hypergraph was obtained based on a hypergraph neural network representing differences. Furthermore, by capturing the patterns between the

收稿日期: 2025-10-27; 修回日期: 2026-04-08

基金项目: 天津市科技计划项目 (25JCYBJC00960)

作者简介: 赵婷婷 (1986—), 女, 内蒙古赤峰人, 教授, tingting@tust.edu.cn

differences in therapeutic efficacy of the formula and the microstructure of the target, a formula evaluation network was designed to achieve accurate prediction of the therapeutic efficacy of TCM prescriptions. The proposed method achieved an accuracy rate of 88.9% for Parkinson's disease and 100% accuracy for both asthma and apoplexy. In terms of execution efficiency, this method completes prescription screening tasks within 0.5 seconds, significantly accelerating the large-scale screening process for TCM prescriptions.

**Key words:** traditional Chinese medicine; protein-protein interaction; hypergraph; graph convolutional network; formula evaluation

传统的中药处方开具过程是由经验丰富的中医根据患者的症状和不同中药的特性进行选择 and 配伍, 其药效通常优于随机组合处方或安慰剂处方<sup>[1-2]</sup>。然而, 由于中药种类繁多且药理作用机制复杂, 仍可能存在部分尚未发现的具有良好治疗效果的中药组合<sup>[3]</sup>。因此, 迫切需要从大量药物组合中快速识别出针对特定病症具有潜在疗效的处方<sup>[4]</sup>。基于蛋白质相互作用 (protein-protein interaction, PPI) 网络距离的中药处方评价方法能够有效评估现有中药处方的有效性<sup>[5-7]</sup>, 但难以准确区分有效处方与疾病中常用中药的随机组合。由于药物与疾病靶点之间存在多靶点协同作用的复杂机制, 其作用关系远比 PPI 网络描述的简单二元相互作用更为复杂, 所以该方法在机制解释层面存在一定的局限性。此外, 该方法受到运行效率的限制, 通常检验一个处方需要少则 1~2 h, 多则 1~2 d 的时间。因此, 在处理大规模中药处方筛选时会遭遇计算瓶颈。

针对上述问题, 本文提出融合微观与宏观双重视角的靶点特异性中药 (target-specificity traditional Chinese medicine, TS-TCM) 框架。通过 PPI 网络构建的可信超图结构从微观层面解析靶点分布特征, 同时结合中药处方的实际疗效从宏观层面评估药效, 由此解决中药配伍过程中存在的可解释性难题以及计算效率瓶颈问题, 以期为中药处方的高效筛选提供理论依据与技术支持。

## 1 研究背景

### 1.1 中药配伍研究

中医根据患者的病历, 结合望、闻、问、切等诊疗手段, 对患者病情进行全面评估。依据病情的需要, 将中药按照七情和合、升降浮沉、君臣佐使、四气五味等原则组合形成中药处方<sup>[8-9]</sup>。然而, 由于中药成分和机制的复杂性, 对中药处方的研究长期局限于疾病症状层面。近年来, 为了更深入地理解中药处方的

原理, 提高中药处方在靶点层面的可解释性, 研究人员引入了靶点层面基于 PPI 网络距离的处方评价。

### 1.2 基于 PPI 网络距离的处方评价方法

PPI 网络包含蛋白质靶点之间的相互作用关系<sup>[10]</sup>, 尽管目前 PPI 网络尚未完善, 但基于 PPI 网络距离的方法已成功应用于西药评价领域。Guney 等<sup>[6]</sup>提出了一种基于 PPI 网络的无监督药效评价方法, 其中 PPI 网络上药物和疾病之间的距离越近, 药物就越有效。Cheng 等<sup>[7]</sup>提出了一种基于 PPI 网络距离的两种药物组合评价方法, 利用两种药物的网络拓扑关系, 预测药物组合的效果。然而, 这类方法在处理多种中药组合时依赖于药物和疾病之间的拓扑特征, 从而限制其推广到更多种药物组合的能力<sup>[5,7]</sup>。此外, 中药的靶点数量远超西药, 带来了严重的计算瓶颈, 在扩大中药处方筛选规模时面临重大挑战。因此, 难以将现有的基于 PPI 网络距离的方法应用于大规模中药处方筛选。

### 1.3 基于聚类的生物分子社区划分

在网络药理学研究领域中, 常采用聚类方法对蛋白质靶点或基因基于相似性特征进行分类分析<sup>[11]</sup>。此类方法能够根据靶点或基因的特征, 识别出具有相似功能的生物分子集合, 并将复杂生物分子网络进行模块化划分, 从而提高运算效率。Liu 等<sup>[12]</sup>通过谱聚类构建了整合组织与细胞系特异性网络框架, 使胰腺导管腺癌中具有协同功能的蛋白质靶点在 PPI 网络中呈现模块化聚集, 进而可实现潜在药物靶点的优先排序。Yuan 等<sup>[13]</sup>采用加权基因共表达网络分析 (WGCNA) 结合 k 均值聚类 (k-means) 算法, 根据表达模式相似性对差异表达基因进行划分, 石斛中与活性成分合成相关的基因集群及其调控路径被有效识别。然而, 传统的聚类方法将生物分子划归特定簇, 其自身的特异性会丢失并且无法恢复。

### 1.4 基于图卷积网络的生物分子社区划分

图卷积网络 (graph convolution network, GCN) 是一种专门用于处理图结构数据的深度学习模型<sup>[14]</sup>。

在靶点识别任务中, 图的拓扑结构依据 GCN 可以将靶点划分为多个允许重叠的功能性社区, 这更符合实际靶点分布特征。Xiao 等<sup>[15]</sup>提出了一种新颖的蛋白质嵌入方法, 该方法将 GCN 与 PageRank 算法相结合, 对 PPI 网络的缺失部分进行补全。Qiu 等<sup>[16]</sup>提出了一种基于 GCN 的癌症特异性子网络识别方法, 通过分析蛋白拓扑特征, 构建癌症特异性动态子网络, 揭示了关键致癌模块及其网络扰动特征。然而, 在使用 GCN 划分靶点社区时, 无法准确定义社区的数量。社区规模过小会导致靶点间复杂关系表达不完整, 实践中较大尺度社区的优先设置成为常见选择, 但社区尺度过大引发冗余社区增多和大量靶点同质化的负面效应同样不容忽视。

本文提出一种高效且具有可解释性的基于精简可信超图的中药处方筛选方法: 针对基于 PPI 网络距离方法的计算瓶颈问题和 GCN 社区划分方法的冗余信息问题, 提出一种精简可靠的超图生成方法; 针对靶点特异性缺失问题, 设计联系靶点微观层面结构信息和药物-疾病宏观药效信息的 TS-TCM 框架; 通过性能评价, 验证本文方法在中医药医学实践中的有效性与适用性。

## 2 TS-TCM 框架

TS-TCM 框架主要包括超图生成部分和处方评价网络部分, 整体框架结构如图 1 所示。超图生成部分包含数据预处理、GCN 社区划分和超图构建 3 个主要过程, 用于获得可靠的靶点超图结构。处方评价网络部分利用专家开具处方和随机处方之间的药物疗效差异, 训练全连接神经网络, 用于从现有的中药处方和超图结构中学习模式, 最终生成处方评分。

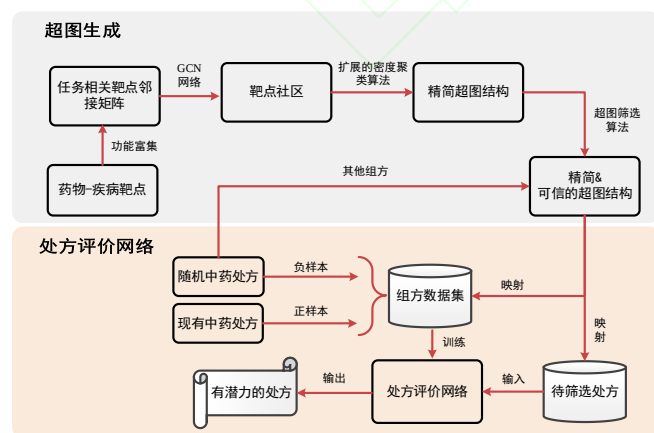


图 1 TS-TCM 框架整体结构

Fig. 1 Overall structure of the TS-TCM framework

针对某一特定疾病, TS-TCM 框架首先获取疾病

靶点和该疾病常用中药的全部靶点, 对这些靶点进行功能富集, 以获得对应药物-疾病过程中显著表达的功能。通过这些富集到的功能可以获得任务相关靶点, 并重构其邻接矩阵。然后, 利用 GCN 网络训练这些靶点的社区关系, 借助扩展的密度聚类算法, 获得初步超图结构, 通过基于动态表示和静态表示差异的超图卷积网络, 筛选具有较高置信度的超边, 加入最终超图结构。最后, 通过靶点超图结构约束网络防止过拟合, 利用专家针对特定疾病开具的若干有效处方的正样本与该疾病常用中药随机组合生成的负样本之间的药物疗效差异, 训练全连接神经网络。训练好的处方评价网络可用于在该疾病上对大规模中药处方进行高效、准确筛选。

### 2.1 数据预处理方法

本文针对帕金森病、哮喘和中风 3 种疾病进行具体研究。帕金森病是一种神经退行性疾病, 病理机制极为复杂, 目前中医在处方治疗方面仍面临巨大挑战。哮喘和中风的病理机制较为明确, 研究更为充分。3 种疾病的靶点来自 GEO (gene expression omnibus)<sup>[17]</sup>和 TTD (therapeutic target database)<sup>[18]</sup>数据集。帕金森病的处方数据由中医专家提供, 药物靶点来自 TCMSP (traditional Chinese medicine systems pharmacology database)<sup>[19]</sup>数据集; 哮喘、中风的处方数据和药物靶点来自 TCM-ID (traditional Chinese medicine information database)<sup>[3]</sup>数据集。

基因本体论 (gene ontology, GO) 数据集<sup>[20]</sup>是生物信息学领域用于描述基因及其产物特性的标准化本体, 涵盖生物过程、细胞组分和分子功能 3 个生物学方面。GO 富集分析用于确定基因集在生物过程、细胞组分和分子功能中的富集情况, 它是一种基于超几何分布的统计方法<sup>[21]</sup>。当特定基因集内某一功能出现的频率远高于其数学期望时, 该功能被认为在该基因集中显著表达。GO 富集的显著性通常由  $P$  表示,  $P$  值越小表示表达越显著, 计算公式为

$$P(k, M, n, N) = 1 - \sum_{i=0}^{k-1} \frac{C_M^i C_{N-M}^{n-i}}{C_N^n} \quad (1)$$

其中:  $C_N^n$  为从  $N$  个不同元素中取出  $n$  个元素的所有组合个数,  $N$  为 PPI 网络靶点的总数,  $n$  为输入靶点的数量,  $M$  为某个功能的靶点数量,  $k$  为输入靶点在某个功能上的靶点数量。

为了控制误报概率, 在实践中通常使用  $Q$  替换  $P$ <sup>[22]</sup>,  $Q$  的计算公式为

$$Q = P \times L / R \quad (2)$$

其中:  $L$  为统计检验的次数,  $R$  为功能  $P$  的重要性排

名。

图 2 展示了任务相关靶点获取流程。首先,通过查询 GO 数据集,对药物-疾病相关的所有靶点进行 GO 富集分析,获取与靶点相关的所有功能,并筛选出显著表达的功能。随后,从 GO 数据集中通过基因-功能映射获取每个候选靶点的功能集合,并与显著表达的功能进行交集分析。保留具有重叠功能的靶点,作为与特定任务相关的关键靶点。

PPI 网络由目前已知的全部蛋白质靶点组成。TS-TCM 中使用的靶点仅为 PPI 网络的一部分,其相互作用关系可能与 PPI 网络的子网络不同。为了解决这个问题,框架中使用显著表达的功能重建这些靶点的邻接矩阵,包含邻接矩阵  $A$  和加权矩阵  $B$ 。邻接矩阵  $A$  依据靶点之间是否存在功能交集获得,其矩阵元素满足

$$A_{ij} = \begin{cases} 0, & \text{if } F(i) \cap F(j) = \emptyset \\ 1, & \text{if } \exists f \in F(i) \cap F(j) \end{cases} \quad (3)$$

其中:  $F(i)$  为靶点  $i$  参与的功能集合,  $F(j)$  为靶点  $j$  参与的功能集合,  $A_{ij}$  为邻接矩阵的第  $i$  行和第  $j$  列。

加权矩阵  $B$  依据靶点间的功能相似性获得,计算公式为

$$B_{ij} = \frac{|F(i) \cap F(j)|}{|F(i) \cup F(j)|} \quad (4)$$

其中:  $|\cdot|$  表示集合中元素的个数,  $B_{ij}$  表示加权矩阵的第  $i$  行和第  $j$  列。

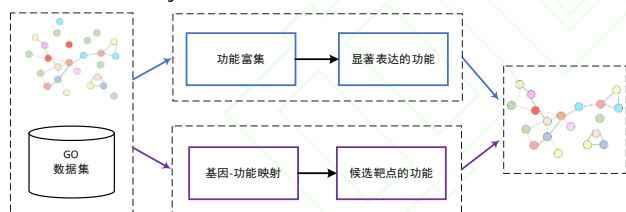


图 2 任务相关靶点获取流程

Fig. 2 Task-related target acquisition process

## 2.2 基于 GCN 的靶点社区划分方法

GCN 被提出用于将卷积运算从传统数据集推广

到图数据集<sup>[14]</sup>。在实践中,通常使用多层图卷积神经网络获得更好的结果。在 TS-TCM 框架中引入 GCN,以实现靶点的重叠社区划分,从而更符合生物学上靶点的多功能性和多通路参与性。采用两层 GCN 结构,其卷积函数为

$$H = \rho \left( D^{-\frac{1}{2}} (A+I) D^{-\frac{1}{2}} \rho \left( D^{-\frac{1}{2}} (A+I) D^{-\frac{1}{2}} X \theta^{(1)} \right) \theta^{(2)} \right) \quad (5)$$

其中:  $\rho$  为 ReLU 激活函数,  $I$  为单位矩阵,  $D$  的功能是归一化  $A+I$ ,  $X$  为靶点的特征向量,  $\theta^{(1)}$  和  $\theta^{(2)}$  为网络需要学习的参数。

在经过 GCN 的传播与变换后,每个靶点均被映射到低维嵌入空间,进而通过嵌入向量之间的关系,刻画靶点之间的连接性与社区结构。嵌入向量的点积越大,向量之间的夹角越小,表明它们越相似。因此,对于连接靶点而言,其嵌入向量之间的点积应尽可能大;对于未连接靶点而言,其嵌入向量之间的点积应尽可能小。划分的社区尽可能多地包含正边(倾向于实际存在的连接),尽可能少地包含负边(倾向于实际不存在的连接)。据此建立模型的损失函数,为

$$L_2 = -\sum_{i=1}^n \log(1 - e^{-eps - E(x_i) \times E(x_2)}) + \alpha \sum_{j=1}^m \log(1 - e^{E(y_1) \times E(y_2)}) \quad (6)$$

其中:  $E$  为靶点的嵌入函数,  $eps$  为一个极小正数,  $x$  为正边,  $y$  为负边,  $n$  和  $m$  分别为正边和负边的数量,  $\alpha$  为平衡正负样本损失的超参数。

GCN 模型基于网络的连通性训练,每次取一部分分子节点,采样子节点中实际存在的边作为正边,随机生成等量实际不存在的边作为负边。由于 PPI 网络通常是稀疏的,因此框架中选择奇异值分解(SVD)<sup>[23]</sup>作为特征向量生成方法。训练方式为构建指定数量的社区,在训练中提高正边在社区内出现的概率,降低负边在社区内出现的概率。在完成模型训练之后,使用  $X$  (邻接矩阵  $A$  的特征向量)作为输入,可以获取靶点对每个社区的隶属度。基于 GCN 的社区划分网络如图 3 所示。

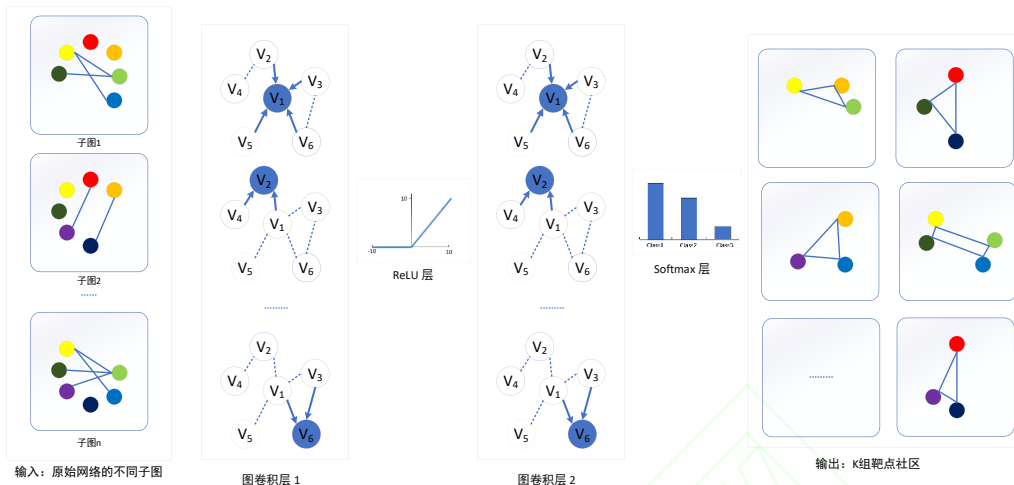


图 3 基于 GCN 的社区划分网络

Fig. 3 GCN-based community partitioning network

### 2.3 可信超图构建方法

相较于普通图而言, 超图结构能够提供更为全面的节点间关系表征<sup>[24-25]</sup>。普通图(图 4)的每条边只能连接两个顶点, 通常采用  $N \times N$  矩阵表示, 第  $i$  行第  $j$  列表示节点  $i$  和节点  $j$  是否连接。超图(图 5)允许每条边包含任意个数的顶点, 通常采用  $N \times E$  矩阵表示, 其中  $N$  为节点数量,  $E$  为超边数量, 第  $i$  行第  $j$  列表示节点  $i$  是否在超边  $j$  上。

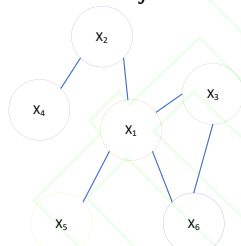


图 4 普通的图典型结构

Fig. 4 Typical structure of general graph

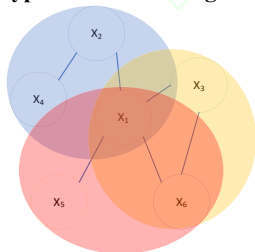


图 5 超图的典型结构

Fig. 5 Typical structure of hypergraph

在中药作用机理的特殊性背景下, 本框架假设同一超边的靶点具有相似性质, 但重要性不同。通过基于 GCN 的靶点社区划分网络, 可以获得靶点对不同社区的隶属度, 其中相似的靶点倾向于出现在相似的

社区中。然而, 社区的数量无法准确定义。基于密度的聚类算法, 使用邻域参数确定要生成的簇的数量, 消除了需要预先设置簇数量的限制, 可以在不精确定义聚类数量的情况下实现良好的聚类划分<sup>[26]</sup>。受此启发, 本文提出一种基于扩展密度聚类的超图生成算法。根据社区距离的高斯分布建立靶点的连通性矩阵  $C$ , 以供密度聚类算法使用, 其计算公式为

$$C_{ij} = \begin{cases} 1, & D(i, j) < \mu - \beta\sigma \\ 0, & D(i, j) \geq \mu - \beta\sigma \end{cases} \quad (7)$$

其中:  $C_{ij}$  为靶点  $i$  和靶点  $j$  之间的连通性,  $D$  为靶点之间的距离,  $\mu$  和  $\sigma$  为高斯分布的参数,  $\beta$  为控制靶点连通性的超参数。

在靶点社区划分的基础上, 通过基于扩展的密度聚类算法生成超图。扩展的密度聚类算法修改了聚类的迭代过程和终止条件。起始点选择策略不同, 允许在已经聚类的点中选择起点, 每次从起始点开始绘制一条超边, 并将其插入超图结构中。此外, 聚类终止条件修改为在大多数靶点完成聚类后停止, 而不是只留下噪声节点。该算法的伪代码见算法 1。

**算法 1** 基于扩展密度聚类的超图生成算法。

**输入** 靶点连通性矩阵  $C$ , 靶点的近邻矩阵  $K$ , 选择概率  $SP$ , 最小超边尺度  $Minpts$ , 靶点覆盖率  $Cover$ 。

**输出** 精简的超图结构  $H$ 。

- (1) 初始化  $H \leftarrow \emptyset$ , 当前覆盖率为 0。
- (2) while 当前覆盖率  $< Cover$  do:

(3) 选择起始点  $S_i$  //以  $1-SP$  的概率选择已经聚类的点为起始点

(4) 初始化当前超边尺度为 1, 新点集合  $\leftarrow \{S_i\}$ , 超边  $\leftarrow \{S_i\}$

(5) while 当前超边尺度  $< Minpts$  且新点集合  $\neq \emptyset$  do:

(6) 初始化邻域

(7) for 每个新点  $\in$  新点集合 do:

(8) 将新点的邻接节点加入邻域

(9) end for

(10) 临时变量  $\leftarrow$  邻域-超边 //记录当前增加的新点

(11) 超边  $\leftarrow$  超边  $\cup$  临时变量 //更新当前超边包含的节点

(12) 更新当前超边尺度, 新点集合  $\leftarrow$  临时变量

(13) end while

(14) if 当前超边尺度  $< Minpts$  then:

(15) 使用近邻矩阵  $K$  补全超边

(16) end if

(17) 将超边放入  $H$

(18) 更新当前覆盖率

(19) end while

(20) return  $H$

上述超图由扩展的密度聚类算法生成。最小超边尺度主要影响单个超边的内聚性与功能覆盖范围, 设置过小易导致捕捉相似功能靶点不全面, 而设置过大可能掩盖局部网络的紧密关联, 从而模糊真实功能结构。靶点覆盖率则决定超图对全局网络的覆盖完整度, 设置过小重要靶点会被遗漏, 导致信息不全, 设置过大则会为满足覆盖稀疏靶点而被迫生成低质量超边, 易引入噪声。然而, TS-TCM 框架中需要超边具有较高的可信度, 并尽量保留网络中的大部分靶点。针对此, 框架中使用基于动态表示和静态表示之间差异的网络进行超边筛选<sup>[27]</sup>。筛选依据是, 良好的超边更不具偶然性, 更不易受高斯分布参数的变化或部分节点的丢失而发生扰动。如果超边的动态表示和静态表示存在显著差异, 则意味着该超边上的点更倾向于形成超边, 评估超边形成概率的公式为

$$p = \frac{1}{K} \sum_{i=1}^K \sigma \left( W_o^T \left( d_i - \tanh(W_s^T x_i) \right) \odot^2 + b \right) \quad (8)$$

其中:  $p$  为超边由给定节点集构成的概率,  $K$  为超边的节点数,  $d_i$  为节点聚合其他节点信息的动态表示,  $\tanh(W_s^T x_i)$  为节点本身的静态表示,  $\sigma$  为 Sigmoid 激活函数,  $\odot$  为哈达玛积,  $w$  和  $b$  为网络待学习参数。

## 2.4 处方评价网络方法

可信超图结构描述了药物与疾病之间关系的微观视图, 从传统药效评价的角度来看, 该结构可以直接用于距离的计算。然而, 在实验中发现, 当随机处方中的中药都是治疗该疾病常用中药时, 基于距离计算的方法往往无法有效区分不同处方的有效性。其根本原因在于, 该类方法未能考虑不同超边所具有的重要性差异。鉴于此, 在本框架中设计了一种能够学习处方有效性与靶点结构之间潜在模式的网络模型, 以实现不同超边赋予合理权重, 其结构如图 6 所示。

在中药的应用场景中, 针对特定疾病的处方数据集规模通常有限, 若采用过于复杂的模型, 极易出现过拟合问题<sup>[28]</sup>。因此, TS-TCM 框架中的处方评价网络采用一个 4 层的简单神经网络, 以在保证表达能力的同时降低过拟合风险。处方评价网络的损失函数为

$$L_{BCE} = -Y \log(y) - (1-Y) \log(1-y) \quad (9)$$

其中:  $y$  为疗效预测值,  $Y$  为处方疗效真实值。

为了训练网络, 使用专家针对特定疾病给出的若干有效处方作为正样本, 并随机选择 4~6 种中药组合来生成负样本。在样本稀缺的情况下, 传统机器学习方法容易过拟合, 而该问题可以通过使用受先验知识约束的机器学习方法进行缓解<sup>[29]</sup>。针对此, 本文在处方评价网络中引入超图结构约束网络, 超图结构提供了中药处方疗效的微观视角, 现有中药处方和随机中药处方之间的疗效差距作为宏观视角, 通过将处方显式映射到超图结构连接宏观和微观尺度。映射函数为

$$M_{ij} = \frac{|D_i \cap H_j|}{|H_j|} \quad (10)$$

其中:  $D_i$  为药物  $i$  包含的靶点集合,  $H_j$  为超边  $j$  包含的靶点集合; 分子为药物  $i$  和超边  $j$  共有的靶点数量, 分母为药物  $i$  和超边  $j$  涉及的靶点数量。

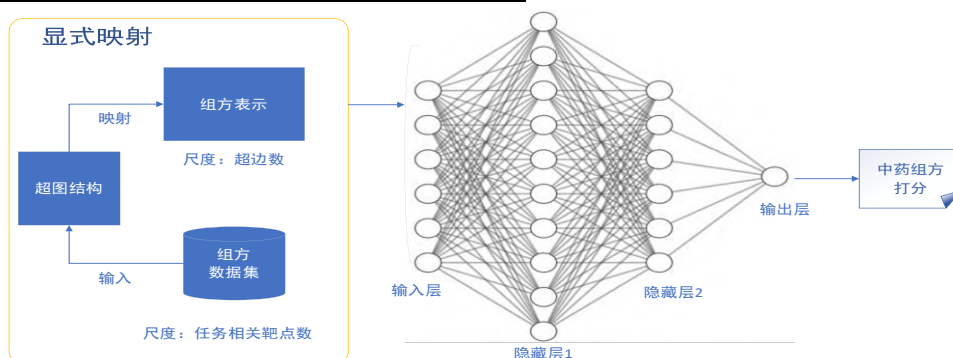


图 6 处方评价网络

Fig. 6 Prescription evaluation network

### 3 结果与讨论

#### 3.1 可视化方法比较

任务相关靶点在 PPI 网络、社区、超图中的分布如图 7 所示。

原始子 PPI 网络关于帕金森症的任务相关靶点如图 7(a)所示。使用帕金森症靶点作为疾病背景靶点, 将 48 种现有中药处方中包含的靶点作为药物背景靶点。使用 0.0005 作为 GO 富集的阈值, 小于此值表示对应功能显著表达。筛选当前靶点集合中具有显著表达功能的靶点, 作为后续的任务相关靶点。

为了尽可能地保持靶点的相互作用关系, 框架中选择了较大的社区规模, 获得由任务相关靶点组成的 250 个社区, 部分社区结构如图 7(b)所示, 其中红色靶点表示该靶点参与了多个社区的构建, 矩形框标记部分为孤立作用的靶点集。在完成社区构建之后,

利用本文的扩展密度聚类算法获取精简超图。在超边筛选部分, 使用基于表示差异的网络对超边打分, 使用基于梯度的方法, 以任务导向的自适应筛选方式过滤超边, 直到达到预设的覆盖要求。在框架中, 默认要求最终结构包括 95% 的靶点, 最终从 300 条超边中选择约 200 个置信度较高的超边。超图的部分结构如图 7(c)所示, 其中红色靶点表示该靶点参与了多条超边的构建, 圈出部分为孤立作用的靶点集。

通过可视化结果可以看出, 原始子 PPI 网络有许多孤立的靶点, 无法准确表示靶点之间的相互作用关系。在使用 GCN 划分靶点社区后, 靶点之间的相互作用关系更加清晰, 但有少量靶点集与其他靶点联系不密切。超图生成后, 孤立的靶点集减少, 靶点之间的相互作用关系体现的更加全面。可见, 基于精简可信超图生成的方法可以显著增强靶点层面的可解释性。

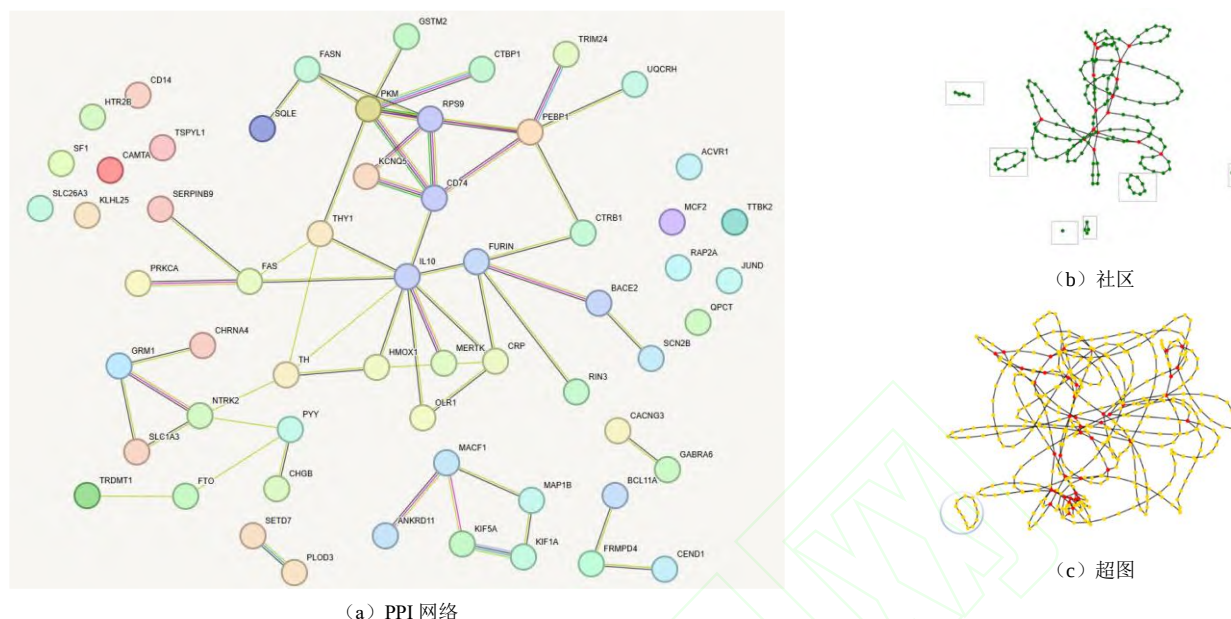


图 7 任务相关靶点在 PPI 网络、社区、超图中的分布

Fig. 7 Task-related target distribution in PPI network, community and hypergraph

### 3.2 性能评价指标比较

通过在帕金森症、哮喘和中风疾病上生成处方的准确率、特征曲线下面积 (AUC) 和筛选效率, 验证 TS-TCM 框架的有效性, 对比算法包括: 基于 PPI 网络距离的方法<sup>[6]</sup>、基于谱聚类 (spectral clustering) 的方法<sup>[30]</sup>、基于 GCN 社区划分的方法<sup>[10,14,31]</sup>、基于任务相关靶点社区划分的方法<sup>[20,31]</sup>、基于 TS-TCM 框架的方法。其中, 以基于 PPI 网络距离的方法作为原始方法, 基于 GCN 靶点社区划分的方法和基于任务相关靶点社区划分的方法为 TS-TCM 方法的消融。此外, 引入谱聚类算法进行对比。在数据划分方面, 选择原始有效中药处方数据集的 30% 作为测试集的正样本, 剩余 70% 作为训练集和验证集的正样本, 而负样本通过从 TCM 背景数据集中采样 4~6 种中药随机生成。由于正样本获取成本高昂, 训练集和验证集中负样本数量多于正样本; 测试集中样本均衡, 以便更公正地评估模型性能。实验过程采用五折交叉验证的方式划分训练集和验证集, 选择在验证集上平均性能最优的数据划分作为最终训练与评估的数据集。所有实验在一台配备 Xeon E5 V3 处理器、RTX 2080Ti 显卡和 64 GB 内存的服务器上完成。

3 种疾病处方、药物及靶点统计分析结果见表 1。不同方法在 3 种疾病上的性能对比结果见表 2。

表 1 3 种疾病处方、药物及靶点统计分析

Tab. 1 Statistical analysis of prescriptions, drugs, and

targets for three diseases

| 项目         | 帕金森病 | 哮喘  | 中风  |
|------------|------|-----|-----|
| 处方/种       | 48   | 48  | 49  |
| 中药成分/种     | 106  | 103 | 102 |
| 处方涉及靶点/个   | 1010 | 281 | 288 |
| 疾病涉及靶点/个   | 683  | 316 | 47  |
| 任务相关靶点靶点/个 | 1582 | 502 | 285 |

准确率表示模型在给定测试数据集上正确分类的样本数与总样本数之比。本框架采用动态调整阈值优化策略, 以平衡验证集上的分类准确率为优化目标选择最优阈值。在测试集中, 将预测概率高于最优阈值的归为正样本, 低于该阈值的归为负样本。实验结果表明, 对于哮喘和中风这类已经广泛研究的疾病, 本文方法仅使用单一模型就可实现 100% 的准确率。然而, 对于帕金森症, 即使尝试了多组不同参数设置, 模型的最佳结果也低于 90%。将其归因于当前对帕金森症的研究相对不足, 用于训练的专家处方也尚未经过临床检验, 这些正样本虽然据专家估计具有一定疗效, 但其中编码的信息可能有限或过于混乱而难以学习<sup>[32]</sup>。相比之下, 另外两种疾病的正样本处方原本即为针对特定疾病开具, 其功能主治明确包含相关症状, 这些样本潜藏的编码信息倾向于满足某个模式, 处方评价网络更容易学习到这些模式。

AUC 数值反映了不依赖于特定阈值的处方疗效排序性能。从 AUC 数值来看, 基于 PPI 网络距离的

方法表现最差, 在哮喘数据集上低至 0.19, 同哮喘病治疗涉及的部分麻醉类药物缺乏相关靶点有关<sup>[33]</sup>。该方法性能较差的原因, 在于其评估依据为药物靶点集与疾病靶点集在整个 PPI 网络中分布的偏移程度。由于随机处方中单种药物均出自常用于指定疾病的中药, 即使单独使用, 也可能表现出疗效, 因此即便是随机组合的处方, 其靶点分布也优于在整个 PPI 网络中完全随机的分布。在这种情况下, PPI 网络距离方法几乎无效。与之对应, 尽管谱聚类的方法通常落后于 GCN 社区划分, 但在哮喘数据集上二者性能表现接近。这可能原因是基于谱聚类的方法在分类过程中使用拉普拉斯矩阵聚合邻域目标信息, 同 GCN 基于拉普拉斯算子的思想具有一定相似性, 但谱聚类的过

程中丢失了靶点特异性, 因而丧失了进一步优化结果的能力<sup>[34]</sup>。因此, 需要在 GCN 社区划分的基础上才能获得更好的结果。在帕金森症上执行任务相关靶点社区划分优于在整个 PPI 网络上执行社区划分, 原因是帕金森症任务相关靶点集同整个 PPI 网络相比, 它精简了全局网络中的冗余信息, 从而更有效地提取出与具体疾病相关的有效信息; 在其他两种疾病上对整个 PPI 网络执行社区划分优于在任务相关靶点上执行社区划分, 原因是对于广泛研究的疾病, 其任务相关靶点较少, 仅依赖微观信息存在不足。本文 TS-TCM 框架受益于蕴含了宏观层面药物疗效信息, 在所有方法中实现了最高的 AUC, 表明其在处方疗效排序任务中具备最强的判别能力。

表 2 不同方法在 3 种疾病上的性能对比结果

Tab. 2 Performance comparison results of different methods on three diseases

| 方法名        | 帕金森病         |             |              | 哮喘       |          |              | 中风       |          |              |
|------------|--------------|-------------|--------------|----------|----------|--------------|----------|----------|--------------|
|            | 准确率          | AUC         | 执行时间/s       | 准确率      | AUC      | 执行时间/s       | 准确率      | AUC      | 执行时间/s       |
| PPI 网络距离   | 0.5          | 0.48        | $> 10^4$     | 0.5      | 0.19     | $> 10^4$     | 0.556    | 0.58     | $> 10^4$     |
| 谱聚类        | 0.611        | 0.52        | 41.535       | 0.722    | 0.72     | 38.889       | 0.667    | 0.70     | 39.922       |
| GCN 社区划分   | 0.556        | 0.54        | 18.506       | 0.778    | 0.73     | 17.422       | 0.944    | 0.96     | 17.406       |
| 任务相关靶点社区划分 | 0.5          | 0.64        | 30.389       | 0.5      | 0.43     | 17.050       | 0.5      | 0.62     | 3.006        |
| TS-TCM 框架  | <b>0.889</b> | <b>0.94</b> | <b>0.027</b> | <b>1</b> | <b>1</b> | <b>0.003</b> | <b>1</b> | <b>1</b> | <b>0.003</b> |

从执行时间来看, 基于 PPI 网络距离的方法效率较低, 通常仅适用于对极少数精选处方的评价; 基于谱聚类和社区划分的方法效率有所提升, 在一定算力支持下能够实现对大规模中药处方的筛选; 本文的 TS-TCM 方法具备较高的执行效率, 适用于大规模处方筛选。基于 PPI 网络距离的方法计算速度最慢, 归因于 PPI 网络尺度较大, 每次距离计算的时间复杂度为  $O(N^2)$ , 其中  $N$  为 PPI 网络的尺度, 约 18000 维; 基于谱聚类和社区划分的方法减小了计算规模, 例如帕金森症为 768 维, 因此显著提升了计算速度; TS-TCM 方法则无需重复使用 PPI 网络进行复杂度较高的计算, 仅依赖于处方评价网络的尺度, 以帕金森症为例, 网络的两层隐藏层尺寸分别为 250 和 25, 因而在执行效率上具有明显优势。

### 3.3 医学实验验证

利用 TS-TCM 框架, 筛选出两个对帕金森症具有较高潜在疗效的处方, 处方 1 为当归、桃仁、赤芍、红花, 处方 2 为郁金、香附、丹参、泽兰; 以上两个处方的评价网络预测输出分数分别为 0.970 和 0.901,

显著优于随机处方的分数均值 0.15。为此, 针对帕金森症, 使用小鼠脑组织切片进行了酪氨酸羟化酶 (tyrosine hydroxylase, TH) 染色实验, 结果如图 8 所示。

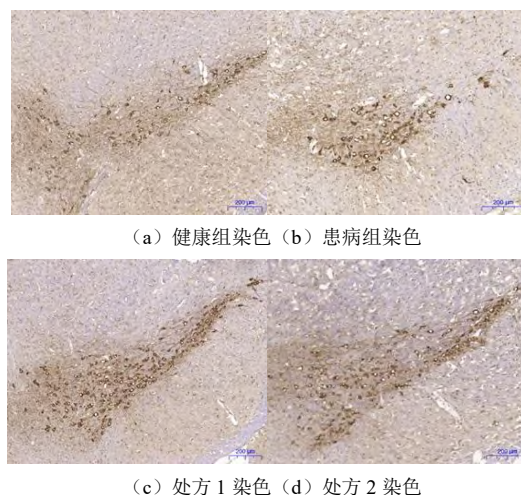


图 8 TH 染色实验结果

Fig. 8 TH staining experimental results

由图 8 可知: 帕金森症患病组 TH 表达比健康组降低; 使用处方 1 和处方 2 治疗, TH 表达比帕金森

症组有所升高,说明帕金森症状有所减轻。TH 染色医学实验表明,利用 TS-TCM 框架筛选出的两个处方对帕金森症有效。

#### 4 结语

本文提出一种高效准确的基于精简可信超图的中药处方筛选方法,即靶点特异性中药(target-specificity traditional Chinese medicine, TS-TCM)框架。从微观视角来说,为了解决靶点同质性问题,使用 GCN 网络生成靶点社区,构建一个高斯分布重建基于靶点社区距离的邻接矩阵,使用扩展的密度聚类算法生成初步的超图结构并利用基于表示差异的超图神经网络校验该结构,以校验后的超图结构替换 GCN 网络生成的靶点社区。从宏观视角来说,为了解决可解释性问题,针对具体疾病构建处方数据库,使用一个处方评价网络捕获处方疗效与靶点微观结构之间的模式,最终将其应用于中药处方筛选。在中药多靶点、多组分背景下,本文方法在区分特定疾病常用中药随机组合与有效处方时,与传统方法相比,准确率、稳定性和执行效率等方面均有显著提升,可在节省算力的基础上,实现针对特定疾病的大规模中药处方精准筛选。TS-TCM 框架为中医药现代化和精准化提供了新的技术手段,有利于未来基于网络药理学的大规模中药处方筛选和优化研究。

当前的 TS-TCM 框架仍然存在一些缺点。首先,框架目前只关注药物组合的治疗效果,而没有考虑药物组合可能产生的副作用<sup>[35]</sup>。如果未来引入中药处方在实际应用中的负面效果的数据集,模型结果的可信度将进一步提高。此外,PPI 网络当前并不完善,Deepmind 的 AlphaFold 实现了对蛋白质结构的准确预测,蛋白质的结构又将决定其功能<sup>[36]</sup>。因此,未来借助 AlphaFold 补全 PPI 网络有望在中医药的医疗实践中取得更准确的结果。

#### 参考文献:

- [1] 何飘,朴美虹,谢丽华,等.从肝治心组方调控铁自噬保护缺血/再灌注损伤心肌细胞机制研究[J].中国中西医结合杂志,2024,44(2):183-192.
- [2] 陈寅清,姚天文,卢嫣,等.桃红四物汤外敷改善血液透析动静脉内瘘通畅性的随机,双盲,安慰剂对照临床研究[J].上海中医药杂志,2023,57(11):33-37.
- [3] Chen Xin, Zhou Hao, Liu Yaobin, et al. Database of traditional Chinese medicine and its application to studies of mechanism and to prescription validation[J]. British Journal of Pharmacology, 2006, 172(23): 1708-8869.
- [4] Zhang Cheng, Chen Guoming, Tang Guoyi, et al. Multi-component Chinese medicine formulas for drug discovery: state of the art and future perspectives[J]. Acta Materia Medica, 2023, 2(1): 106-125.
- [5] Gan Xiao, Shu Zixin, Wang Xinyan, et al. Network medicine framework reveals generic herb-symptom effectiveness of traditional Chinese medicine[J]. Science Advances, 2023, 9(43): eadh0215.
- [6] Guney E, Menche J, Vidal M, et al. Network-based in silico drug efficacy screening[J]. Nature Communications, 2016, 7: 10331.
- [7] Cheng Feixiong, Kovacs I A, Barabasi A L. Network-based prediction of drug combinations[J]. Nature Communications, 2019, 10: 1197.
- [8] 陈聪聪,秦雪梅,杜冠华,等.中药有效组分“配伍艺术”的研究基础——组分配比研究策略[J].药学报,2019,54(5):808-817.
- [9] 字磊,李艳娟,李艳芹,等.中药复方配伍机制研究方法策略进展[J].中国药房,2023,34(11):1393-1398.
- [10] Szklarczyk D, Kirsch R, Koutrouli M, et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest[J]. Nucleic Acids Research, 2023, 51(D1): D638-D646.
- [11] Ma Jun, Wang J, Ghorai L S, et al. A comparative study of cluster detection algorithms in protein-protein interaction for drug target discovery and drug repurposing[J]. Frontiers in Pharmacology, 2019, 10: 109.
- [12] Liu Enze, Zhang Zhuangzhuang, Cheng Xiaolin, et al. SCNrank: spectral clustering for network-based ranking to reveal potential drug targets and its application in pancreatic ductal adenocarcinoma[J]. BMC Medical Genomics, 2020, 13(Suppl 5): 50.
- [13] Yuan Yingdan, Zhang Bo, Tang Xinggang, et al. Comparative transcriptome analysis of different dendrobium species reveals active ingredients-related genes and pathways[J]. International Journal of Molecular Sciences, 2020, 21(3): 861.

- [14] Zhang Si, Tong Hanghang, Xu Jiejun, et al. Graph convolutional networks: a comprehensive review[J]. Computational Social Networks, 2019, 6(1): 1-23.
- [15] Xiao Ze, Deng Yue. Graph embedding-based novel protein interaction prediction via higher-order graph convolutional network[J]. PloS One, 2020, 15(9): e0238915.
- [16] Qiu Jiajun, Chen Kui, Zhong Chunlong, et al. Network-based protein-protein interaction prediction method maps perturbations of cancer interactome[J]. PLoS Genetics, 2021, 17(11): e1009869.
- [17] Barrett T, Wilhite S E, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update[J]. Nucleic Acids Research, 2012, 41(D1): D991-D995.
- [18] Zhou Ying, Zhang Yintao, Zhao Donghai, et al. TTD: therapeutic target database describing target druggability information[J]. Nucleic Acids Research, 2024, 52(D1): D1465-D1477.
- [19] Ru Jinlong, Li Peng, Wang Jinan, et al. TCMSP: a database of systems pharmacology for drug discovery from herbal medicines[J]. Journal of Cheminformatics, 2014, 6(1): 1-6.
- [20] GeneOntologyConsortium. The gene ontology resource: 20 years and still GOing strong[J]. Nucleic Acids Research, 2019, 47(D1): D330-D338.
- [21] Kim H Y. Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test[J]. Restorative Dentistry & Endodontics, 2017, 42(2): 152-155.
- [22] Jafari M, Ansari-Pour N. Why, when and how to adjust your P values?[J]. Cell Journal, 2019, 20(4): 604.
- [23] KeZ T, Wang Minzhe. Using SVD for topic modeling[J]. Journal of the American Statistical Association, 2024, 119(545): 434-449.
- [24] Feng Yifan, You Haoxuan, Zhang Zizhao, et al. Hypergraph neural networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: Association for the Advancement of Artificial Intelligence, 2018: 4924-4931.
- [25] Gao Yue, Feng Yifan, Ji Shuyi, et al. HGNN+: general hypergraph neural networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(3): 3181-3199.
- [26] Bhattacharjee P, Mitra P. A survey of density based clustering algorithms[J]. Frontiers of Computer Science, 2021, 15(1): 1-27.
- [27] Zhang Ruochi, Zou Yuesong, Ma Jian. Hyper-SAGNN: a self-attention based graph neural network for hypergraphs[EB/OL]. arXiv (2019-11-11) [2025-11-20]. <https://doi.org/10.48550/arXiv.1911.02613>.
- [28] Lü Qiuji, Chen Guanxing, Yang Ziduo, et al. Meta learning with graph attention networks for low-data drug discovery[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 35(8): 11218-11230.
- [29] von Rueden L, Mayer S, Beckh K, et al. Informed machine learning—a taxonomy and survey of integrating prior knowledge into learning systems[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 35(1): 614-633.
- [30] Berahmand K, Nasiri E, Li Yuefeng. Spectral clustering on protein-protein interaction networks via constructing affinity matrix using attributed graph embedding[J]. Computers in Biology and Medicine, 2021, 138: 104933.
- [31] 王志杰. 基于图神经网络的现代中药配伍机制的研究[D]. 天津: 天津科技大学, 2024: 18-38.
- [32] 陆潭丽. 基于生物信息学方法寻找帕金森病的关键基因[D]. 南宁: 广西医科大学, 2021: 9-13.
- [33] 翟芬芬, 刘毓菲, 张传涛, 等. 基于网络药理学和生物信息学的小青龙汤治疗哮喘和慢性阻塞性肺疾病的“异病同治”机制分析[J]. 湖南中医药大学学报, 2022, 42(06): 958-965.
- [34] Kipf T N, Welling M. Semi-Supervised Classification with Graph Convolutional Networks[EB/OL]. arXiv (2016-09-27) [2025-11-20]. <https://doi.org/10.48550/arXiv.1609.02907>.
- [35] 颜兴琼, 刘妍臻, 胡永翠, 等. 浅谈中药制剂安全性评价的建立[J]. 中国畜禽种业, 2023, 19(3): 137-139.
- [36] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with alphafold[J]. Nature, 2021, 596(7873): 583-589.