基于 GEO 数据库的胃癌差异表达基因系统分析及潜在中药靶向筛选

沈翊康，庞华鑫，刘明睿，刘海煜，马鹏珍，李雅宁，王麒皓，谢晓霞，张小平，赵玉凤*

中国中医科学院 中医药数据中心，北京 100700

摘 要：目的 基于基因表达综合数据库（Gene Expression Omnibus，GEO）整合胃癌差异表达基因，系统鉴定肿瘤进展相关核心靶点，并通过网络距离预测具有治疗潜力的中药，为胃癌中西医结合精准干预提供分子依据。**方法** 从 GEO 下载 21 个胃癌数据集（胃癌 2 125 例、正常 367 例），构建表达矩阵。采用 limma 包筛选差异表达基因（differentially expressed genes，DEGs）；加权基因共表达网络分析（weighted gene co-expression network analysis，WGCNA）构建共表达网络，鉴定与疾病表型最相关模块；基因本体（gene ontology，GO）和京都基因与基因组百科全书（Kyoto encyclopedia of genes and genomes，KEGG）富集分析关键基因功能；构建 7 种机器学习模型，SHapley 可加性解释（SHapley additive exPlanations，SHAP）特征重要性；基于人类蛋白质相互作用（protein-protein interaction，PPI）网络计算中药靶点模块与胃癌关键基因的网络距离，筛选拓扑接近中药并统计四气、五味、归经与功效。**结果** 共获 455 个 DEGs，WGCNA 划分 31 个模块，浅黄色模块（r＝0.56，q＜0.01）含 194 个枢纽基因（hub genes），与 DEGs 交集得 177 个关键基因。富集分析显示 GO-生物过程（biological processes，BP）集中于细胞外基质组织与黏附，GO-细胞成分（cell component，CC）富集含胶原细胞外基质（extracellular matrix，ECM）与黏附斑，GO-分子功能（molecular function，MF）突出整合素/生长因子结合；KEGG 涵盖 actin 骨架调控、磷脂酰肌醇-3-羟激酶（phosphatidylinositol-3-hydroxykinase，PI3K）-蛋白激酶 B（protein kinase B，Akt）及白细胞介素-17（interleukin-17，IL-17）信号。随机森林（random forest，RF）模型准确率 0.991，SHAP 一致识别 SULF1、THY1、DNER、SPINK7 为首位贡献基因。网络距离筛选出牛蒡子、夏枯草、苍术、川贝、女贞子、地耳草、桃仁等前 15 味中药，四气以凉为主，五味以苦为先，归经肝、胃、肺，功效以清热为主、补虚次之。**结论** 系统揭示了胃癌的分子机制，预测清热解毒、滋阴活血类中药具多靶点抗胃癌潜力，为胃癌精准诊疗与中医药现代化提供新策略。

关键词：胃癌；差异基因；SULF1；THY1；DNER；SPINK7；清热解毒；滋阴活血

中图分类号：Q811.4；R285 **文献标志码：**A **文章编号：**0253 - 2670(2026)08 - 3099 - 11

DOI: 10.7501/j.issn.0253-2670.2026.08.022

Systematic analysis of differentially expressed genes in gastric cancer based on GEO database and targeted screening of potential traditional Chinese medicines

SHEN Yikang, PANG Huaxin, LIU Mingrui, LIU Haiyu, Ma Pengzhen, LI Yaning, WANG Qihao, XIE Xiaoxia, ZHANG Xiaoping, ZHAO Yufeng

Data Center of Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, No. 16, Nanxiaojie, Inside Dongzhimen, Dongcheng District, Beijing 100700, China

Abstract: Objective To integrate differentially expressed genes (DEGs) in gastric cancer (GC) from the Gene Expression Omnibus (GEO) database, systematically identify core targets associated with tumor progression, and predict therapeutic Chinese medicines via network distance, providing molecular evidence for integrated traditional Chinese and Western medicine precision intervention in GC. **Methods** A total of 21 GC datasets (2 125 GC, 367 normal samples) were downloaded from GEO to construct an expression matrix. DEGs were screened using the limma package ($|\log_2(FC)| > 1$, FDR < 0.05), weighted gene co-expression network analysis (WGCNA) was performed to identify modules most correlated with disease phenotype, gene ontology (GO) and Kyoto encyclopedia of genes and genomes (KEGG) enrichment analyses were conducted on key genes, seven machine learning models were built with SHapley additive

收稿日期：2025-11-13

基金项目：国家自然科学基金面上项目（82374621）；中国中医科学院创新工程项目（CI2021A05042）；国家自然科学基金面上项目（82575263）

作者简介：沈翊康（1995—），男，汉族，江苏无锡人，博士研究生，研究方向为真实世界中医药大数据分析与应用。

***通信作者：**赵玉凤（1978—），女，汉族，黑龙江五常人，研究员，博士研究生导师，博士，研究方向为真实世界中医药大数据分析与应用。E-mail: snowmanzhao@163.com

exPlanations (SHAP) for feature importance interpretation; network distance between Chinese medicine target modules and GC key genes was calculated based on the human PPI network to screen topologically proximal medicines, with statistics on four properties, five flavors, meridian tropism, and efficacy. **Results** A total of 455 DEGs were obtained. WGCNA yielded 31 modules, with the light-yellow module ($r = 0.56$, $q < 0.01$) containing 194 hub genes; intersection with DEGs produced 177 key genes. Enrichment analysis showed GO-biological processes (BP) focused on extracellular matrix organization and adhesion, GO-cell component (CC) on collagen-containing ECM and focal adhesion, GO-molecular function (MF) on integrin/growth factor binding, and KEGG on actin cytoskeleton regulation, phosphatidylinositol-3-hydroxykinase (PI3K)-protein kinase B (Akt), and interleukin-17 (IL-17) signaling. The random forest (RF) model achieved 0.991 accuracy, with SHAP consistently ranking SULF1, THY1, DNER, and SPINK7 as top contributors. Network distance screening identified *Arctii Fructus*, *Prunellae Spica*, *Atractylodis Rhizoma*, *Fritillariae Cirrhosae Bulbus*, *Ligustri Lucidi Fructus*, *Hypocreaceae*, and *Persicae Semen* among the top 15 medicines, characterized by cool/cold properties, bitter flavor, liver/stomach/lung tropism, and primarily heat-clearing with deficiency-tonifying efficacy. **Conclusion** This study systematically elucidates GC molecular mechanisms, predicting multi-target anti-GC potential of heat-clearing, *yin*-nourishing, and blood-activating Chinese medicines, and provides novel strategies for GC precision diagnosis/treatment and modernization of traditional Chinese medicine.

Key words: gastric cancer; differentially expressed genes; SULF1; THY1; DNER; SPINK7; heat-clearing and detoxifying; *yin*-nourishing and blood-activating

胃癌是一种以胃黏膜上皮细胞恶性增殖为特征的消化系统肿瘤。根据国际癌症研究机构（international agency for research on cancer，IARC）全球癌症观测站（Global Cancer Observatory，GLOBOCAN）2020 年全球癌症统计报告，胃癌为全球第 6 常见恶性肿瘤，年新发病例约 108.9 万，死亡约 76.9 万，占全部癌症死亡的 7.7%[1]。东亚地区（尤其是中国、日本、韩国）发病率和死亡率居全球首位，约占全球病例的 50%[2]。值得注意的是，胃癌在中年群体中负担日益突出，约 80%的患者初诊时已处于局部晚期或转移期[3]。尽管早期筛查可使 5 年生存率超过 90%，但总体 5 年生存率仍不足 40%，提示早诊早治依然是提高预后的关键[4]。

尽管我国已制定多学科综合治疗方案（包括手术切除、新辅助/围手术期化疗、靶向治疗及免疫治疗等），但对于晚期患者而言，复发率仍然较高、治疗效果有限[5]。《中国胃癌诊疗规范（2023 版）》及中国临床肿瘤学会（Chinese Society of Clinical Oncology，CSCO）2023 指南均推荐氟尿嘧啶＋亚叶酸钙＋奥沙利铂＋多西他赛（fluorouracil，leucovorin，oxaliplatin，docetaxel，FLOT）化疗方案作为局部进展期可切除胃癌的围手术化疗首选，但其 3～4 级中性粒细胞减少发生率可达 45%～50%[6]。人表皮生长因子受体 2（human epidermal growth factor receptor 2，HER2）靶向药曲妥珠单抗（Trastuzumab）仅适用于 15%～20%的 HER2 阳性患者[7]；免疫检查点抑制剂（如纳武利尤单抗）在程序性死亡配体 1 联合阳性评分（programmed death-ligand 1 combined positive score，PD-L1）≥5 或微卫星高度不稳定/错配修复缺陷（microsatellite instability-high/deficient mismatch repair，MSI-H/dMMR）亚型中的客观缓解率约 30%，而总体应答率不足 20%[8]。因此，晚期胃癌仍面临疗效有限与耐药频发的双重挑战，亟需开发兼具高效抗肿瘤作用与长期安全性的新策略[9]。

在中医理论体系中，胃癌并无完全对应的传统病名。根据其临床表现差异，可分属“胃痞”“噎膈”“积聚”“癥瘕”等范畴，其病机以本虚标实、脾胃虚弱、气滞血瘀、痰湿内结为主要特征[10]。治疗原则以扶正祛邪、健脾益气、活血化瘀为主。现代临床常用复方包括黄芪、白花蛇舌草、半枝莲、莪术、三棱等药物，部分研究显示联合化疗可改善患者生活质量，降低化疗后白细胞减少及胃肠道不良反应发生率[11]。中医药在延缓疾病进展、减轻化疗毒副反应方面具有辅助价值，但证据质量尚待高水平多中心随机对照试验验证[12]。

近年来研究揭示，胃癌在疾病早期即呈现显著的分子异质性，涉及肿瘤蛋白 p53 基因（tumor protein 53，TP53）、钙黏蛋白 1 基因（cadherin 1，CDH1）、AT 富集相互作用结构域 1A 基因（AT-rich interaction domain 1A，ARID1A）等基因突变及 Wnt、磷脂酰肌醇-3-羟激酶（phosphatidylinositol-3-hydroxykinase，PI3K）-蛋白激酶 B（protein kinase B，Akt）、丝裂原活化蛋白激酶（mitogen-activated protein kinase，MAPK）等信号通路异常重编程[13]。癌症基因组图谱（The Cancer Genome Atlas，TCGA）

提出基于分子分型的 4 亚型［爱泼斯坦-巴尔（Epstein-Barr，EB）病毒型、微卫星不稳定型、基因稳定型、染色体不稳定型］，揭示其侵袭性与治疗敏感性差异[14]。整合多组学数据与网络药理学分析，可识别与胃癌进展相关的关键基因模块，并筛选可能作用于这些靶点的中药活性成分。此类研究为精准医学及中西医结合治疗提供了新的思路，有望突破传统治疗的局限性。

1 资料与方法

1.1 数据收集与预处理

从基因表达综合数据库（Gene Expression Omnibus，GEO）[15]下载 21 个包含胃癌患者基因测序数据的公共数据集。检索词设定为“gastric cancer”和“homo sapiens”，筛选条件包括：数据集为未经干预的临床样本，每个数据集中至少包含 10 个样本。使用 R 语言“limma”包[16]对原始数据进行背景校正和归一化处理。最终纳入分析的数据集包括 GSE34942、GSE19826、GSE147043、GSE26942、GSE15459、GSE54129、GSE84787、GSE191275、GSE66222、GSE62254、GSE13911、GSE51105、GSE118916、GSE79973、GSE183136、GSE208099、GSE87666、GSE29998、GSE113255、GSE13861、GSE236522，共计 21 个数据集。

1.2 胃癌基因组数据库整合

为构建全面可靠的胃癌基因表达谱，对 21 个 GEO 数据集进行整合分析。首先，筛选在至少 90% 数据集中均检测到的基因，以获得共同基因集。其次，针对表达数据中的缺失值，采用中位数填充法进行填充。此外，为消除不同平台与批次效应，使用 R 包“ComBat-seq”[17]对整合数据进行批次效应校正，从而减少技术变异对下游分析的影响。

1.3 差异基因筛选

利用 R 语言“limma”包对整合后的表达矩阵进行差异表达分析。筛选标准为假发现率（false discovery rate，FDR）≤0.05 且$|\log_2(FC)|\geq 1$。通过 FDR 控制多重检验假阳性率，确保结果可靠性；$|\log_2 FC)|\geq 1$［FC 为差异倍数（fold change）］表示基因表达水平至少变化 2 倍。该筛选结果用于后续火山图、热图绘制、PPI 网络构建及加权基因共表达网络分析（weighted gene co-expression network analysis，WGCNA）。

1.4 WGCNA

基于 DEGs 运用 WGCNA 鉴定与胃癌表型显著相关的关键模块。构建加权基因网络前，过滤低表达基因及离群样本，设置最佳软阈值，模块内最少包含 50 个基因。共表达模块为高拓扑重叠相似性的基因集合，同一模块内基因通常具有更高共表达水平。使用模块特征基因（module eigengene，ME，模块第 1 主成分）描述各样本中模块表达模式，并以模块隶属度（module membership，MM，基因与 ME 相关系数）评估基因所属可靠性。根据 ME 与临床表型（胃癌/正常）相关性确定重要模块。高度模块化且与表型强相关的基因为潜在调控因子或生物标志物。

1.5 基因集富集分析

为深入挖掘关键基因的生物学意义，利用 Metascape 在线平台整合基因本体（gene ontology，GO）、京都基因与基因组百科全书（Kyoto encyclopedia of genes and genomes，KEGG）、Reactome 等多数据库，进行基因集富集分析（gene set enrichment analysis，GSEA）[18]。系统揭示基因在生物学过程（biological process，BP）、细胞组分（cellular component，CC）、分子功能（molecular function，MF）及信号通路中的富集情况，为功能解读提供多层次视角。

1.6 机器学习模型的开发与验证

基于 DEGs，利用 Python 3.11 搭建 7 种经典机器学习算法，包括逻辑回归（logistic regression，LR）、随机森林（random forest，RF）、支持向量机（support vector machine，SVM）、K 近邻算法（K-nearest neighbors，KNN）、高斯朴素贝叶斯（gaussian naive bayes，GNB）、决策树（decision tree，DT）和梯度提升（gradient boosting，GB）。

LR[19]通过线性方程预测分类概率，结构简单但对非线性关系有限；RF[20]为决策树集成算法，通过多树投票实现高准确率与泛化能力；SVM[21]寻找最优超平面，适用于高维非线性数据；KNN[22]基于距离最近的 k 个样本分类，简单但对噪声敏感；GNB[23]假设特征独立，计算高效但对相关性敏感；DT[24]基于 if-then 规则，易解释但易过拟合；GB[25]迭代优化弱分类器，准确率高且泛化良好。

将胃癌诊断视为二分类问题，以 DEGs 为预测变量，疾病状态为响应变量。采用 10 折交叉验证训练 2 492 个样本，评估准确率、特异性、敏感性和受试者工作特征（receiver operator characteristic，ROC）曲线下面积（area under urve，AUC）。

1.7　基于 SHapley 可加性解释（SHapley additive exPlanations，SHAP）的模型解释与特征重要性分析

采用 SHAP 方法[26]对 RF、GB、DT 模型进行解释，量化各基因对预测的贡献度，增强模型可解释性并揭示胃癌分子机制。取 3 模型贡献度前 10 基因去重，共 20 个用于中药筛选。

1.8　蛋白质-蛋白质相互作用（protein-protein interaction，PPI）网络与中药靶点数据

人类 PPI 网络整合多种实验验证数据，包括：（1）高通量酵母双杂交与三维结构分析的二元相互作用；（2）亲和纯化-质谱鉴定的复合物相互作用；（3）激酶-底物相互作用；（4）信号传导相互作用；（5）调控相互作用。最终网络含 18 505 个蛋白节点与 327 924 条边[27]。

中药靶点数据来源于（1）HIT 2.0 数据库[28]，提供中药-靶点映射；（2）TCMIO 数据库[29]，整合 TCMSP[30]、TCMID[31]、TCM-ID[32]，涵盖化学成分与治疗作用、成分经质谱与质量控制筛选[33-34]；靶点信息来自 STITCH 数据库[35]，仅保留实验支持数据。药物限定于《中药学》教材常用药。

1.9　网络距离计算

通过网络距离分析评估胃癌代表基因与中药作用靶点之间的拓扑接近性，其中网络距离越小表明中药靶点与疾病基因在 PPI 网络中的功能关联越紧密。在 PPI 网络中，蛋白质作为节点，蛋白质之间的相互作用作为无权边，任意 2 个蛋白节点之间的网络距离定义为其最短路径长度。网络距离计算所使用的原始数值即为 PPI 网络中蛋白质节点之间的最短路径长度。

中药通过其对应的靶点蛋白集合表示，胃癌通过其相关疾病基因集合表示。中药靶点集合 A 与疾病基因集合 B 之间的网络距离（D_{AB}）定义为 2 组节点之间所有节点对最短路径长度的平均值。

$$\langle D_{AB}\rangle = \frac{1}{|A||B|}\sum_{a\in A,\, b\in B} d(a,b)$$

其中，$d(a, b)$ 表示蛋白质 a 与蛋白质 b 在 PPI 网络中的最短路径长度。若某一节点对在网络中不存在连通路径，则该节点对不纳入距离计算。

在此基础上，进一步计算网络分离度（S_{AB}）[36-37]，用于衡量中药靶点集合与疾病基因集合在网络拓扑结构中的相对位置关系。

$$S_{AB} = \langle d_{AB}\rangle - \frac{\langle d_{AA}\rangle + \langle d_{BB}\rangle}{2}$$

其中，$\langle d_{AA}\rangle$ 和 $\langle d_{BB}\rangle$ 分别表示节点集合 A 和节点集合 B 内部所有节点对之间最短路径长度的平均值；$\langle d_{AB}\rangle$ 表示集合 A 与集合 B 之间节点对的平均最短路径长度。对于同时属于集合 A 和 B 的节点，其距离定义为 0。

基于 PPI，采用 Python 实现的网络分离度计算方法与 Menche 等[36]提供的代码逻辑一致。当 $S_{AB}<0$ 时，表明 2 组节点在网络中位于相同的功能邻域；当 $S_{AB}>0$ 时，表明 2 组节点在拓扑结构上相互分离。如图 1 所示，相较于中药 2，中药 1 的作用靶点与胃癌代表基因在网络中具有更高的接近性，提示其潜在治疗效果可能更优。

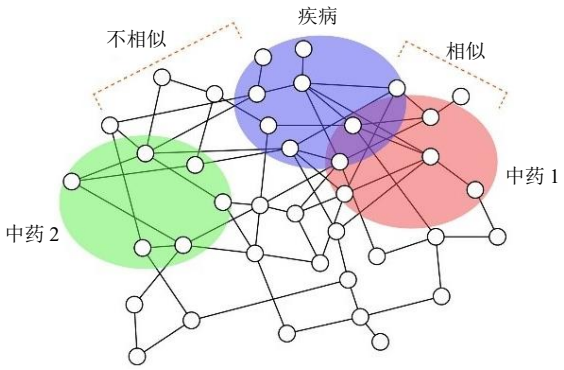


图 1　网络距离示意图

Fig. 1　Schematic diagram of network distance

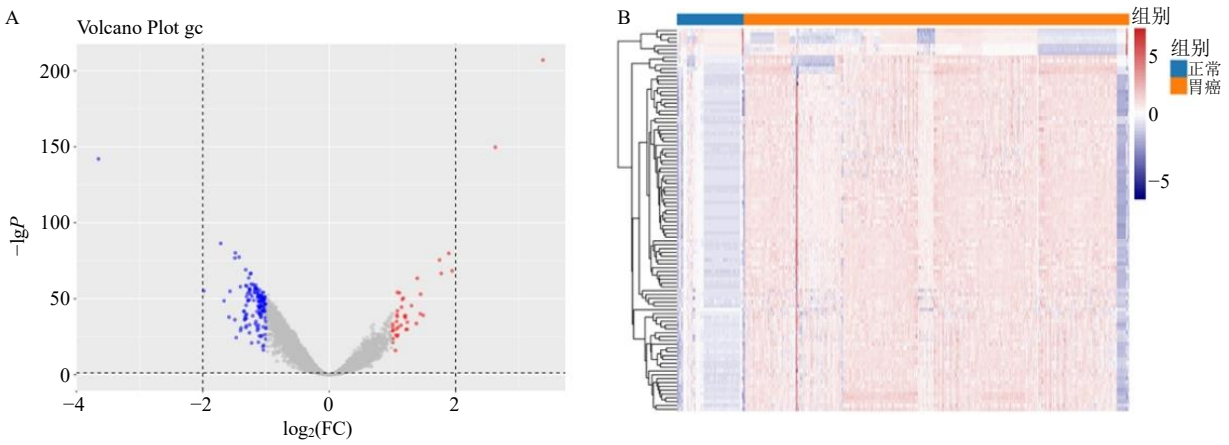
2　结果

2.1　差异基因分析

整合 21 个公共数据集中的胃癌患者基因测序数据，涵盖 2 125 个胃癌样本和 367 个正常样本，构建包含 13 772 个基因的表达矩阵。采用$|\log_2(FC)|>1.0$及调整后 P 值<0.05 作为筛选标准，共识别 455 个 DEGs，其火山图和热图见图 2。

2.2　WGCNA 分析结果

首先基于 STRING 数据库构建 455 个 DEGs 的 PPI 网络（图 3）。为强化强相关、弱化弱相关，WGCNA 采用幂函数加权基因共表达相关系数。对 2 492 个样本的 455 个基因进行层次聚类，划分模块并提取 ME，代表模块整体表达模式。根据 ME 与疾病表型相关性评估模块重要性，高度模块化且与表型强相关的基因为潜在关键调控因子或生物标志物。聚类分析确定最佳软阈值 β＝4，此时网络符合无尺度分布（图 4-A、B），共划分 31 个模块，其中浅黄色模块与胃癌进展相关性最强（r＝0.56，q＜0.01，图 4-C），含 194 个枢纽基因（hub genes）。



A-胃癌患者样本与正常样本差异基因火山图，红色点表示上调基因，蓝色点表示下调基因；B-胃癌组与正常组间差异基因热图（每行代表 1 个基因，红色为高表达，蓝色为低表达）。

A-volcano plot of differentially expressed genes (DEGs) between gastric cancer and normal samples, red dots indicate upregulated genes, and blue dots indicate downregulated genes; B-heatmap of DEGs between gastric cancer and normal groups (each row represents a single gene, red indicates high expression, and blue indicates low expression).

图 2　胃癌患者与正常组差异基因热图和火山图

Fig. 2　Heatmap and volcano plot of differentially expressed genes between gastric cancer patients and normal groups

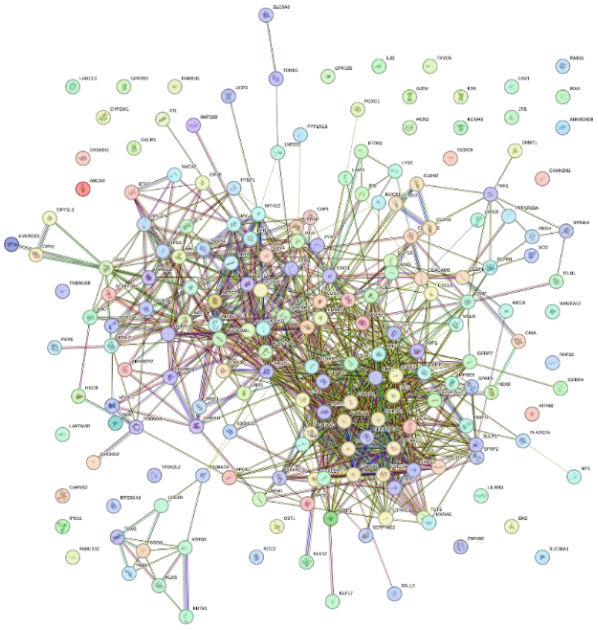


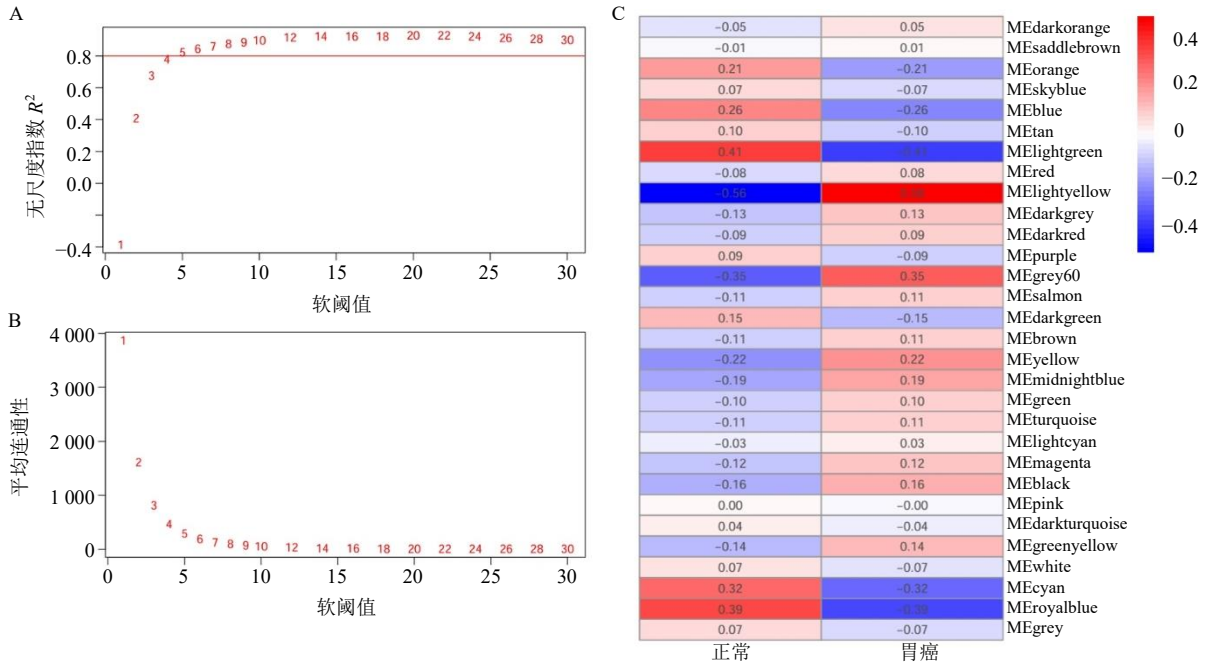
图 3　基于 STRING 数据库的差异表达基因蛋白互作网络

Fig. 3　Protein-protein interaction network of differentially expressed genes based on STRING database

采用 VENNY 2.1.0 取 DEGs 与枢纽基因的交集，获 177 个胃癌关键基因。

2.3　GO 功能及 KEGG 通路富集分析结果

为阐明 177 个关键基因的生物学意义，对其进行 GO 和 KEGG 富集分析（图 5）。GO-生物过程（biological processes，BP）方面：主要富集于外包结构组织（external encapsulating structure organization）、细胞外基质组织（extracellular matrix organization）、细胞外结构组织（extracellular structure organization）、细胞-基质黏附（cell-matrix adhesion）及细胞-底物黏附（cell-substrate adhesion）等，均与细胞外基质（extracellular matrix，ECM）形成与重塑相关。ECM 重构促进肿瘤侵袭、迁移与转移，并增强癌细胞-微环境交互[38]；黏附分子及细胞连接复合体异常与上皮-间质转化（epithelial-mesenchymal transition，EMT）及微环境重塑密切相关[39]，提示细胞外结构与黏附通路在胃癌进展中的核心作用。GO-细胞组分（cell component，CC）方面：主要富集于含胶原细胞外基质（collagen-containing extracellular matrix）、内质网腔（endoplasmic reticulum lumen）、黏附斑（focal adhesion）、基底膜（basement membrane）及胶原三聚体（collagen trimer）。这些条目反映 ECM 及其结构蛋白（尤其是胶原）异常；黏附斑调控细胞-基质交互，内质网腔与分泌颗粒腔富集提示蛋白分泌增强，共同驱动肿瘤间质重塑[40]。GO-分子功能（molecular function，MF）方面：主要富集于细胞外基质结构组分（extracellular matrix structural constituent）、整合素结合（integrin binding）、钙黏蛋白结合（cadherin binding）、生长因子结合（growth factor binding）及糖胺聚糖结合（glycosaminoglycan binding）等，与 ECM 组成及信号传导相关[41]。整合素介导的黏附-迁移、生长因子结合增强的促增殖信号（如血小板衍生生长因子）共同构成胃癌恶性进展

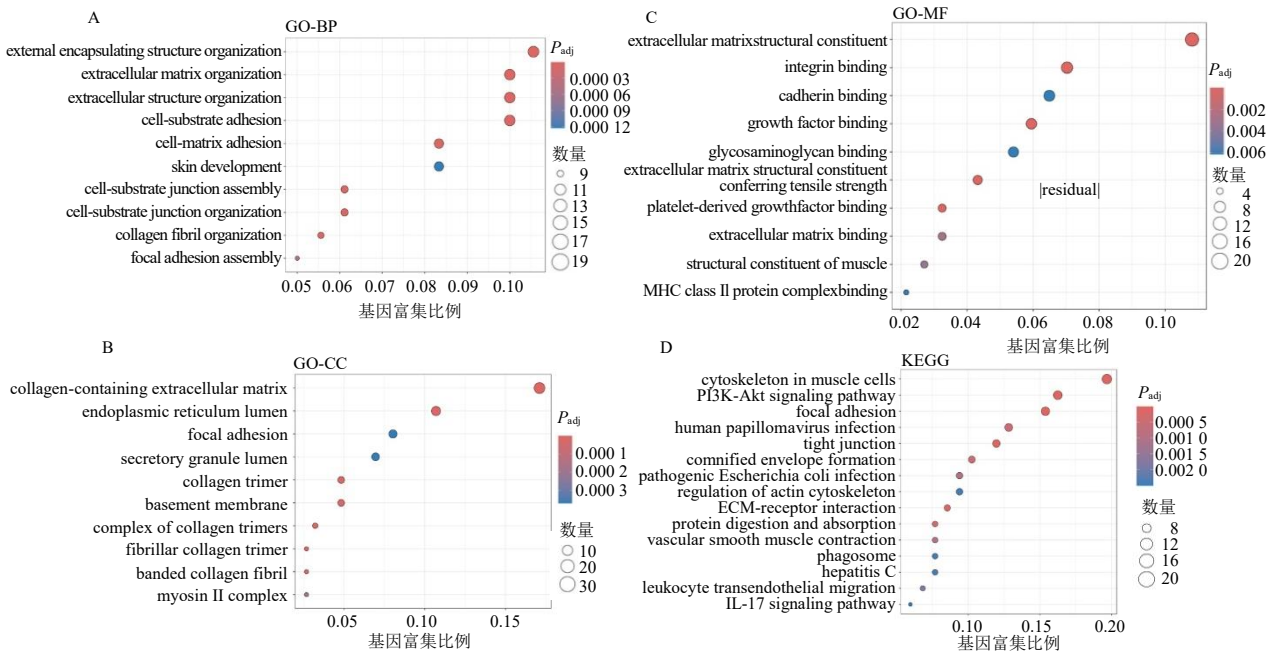


A-不同软阈值幂的无尺度指数分析；B-不同软阈值幂的平均连通性分析；C-模块特征基因与临床表型相关性热图。

A-scale-free topology fit index analysis for various soft-thresholding powers; B-mean connectivity analysis for various soft-thresholding powers; C-heatmap of correlation between module eigengenes and clinical traits.

图 4　基因共表达网络构建与模块划分可视化

Fig. 4　Visualization of co-expression network construction and module division



A～C-GO 功能富集分析结果；D-KEGG 通路富集分析结果。

A—C-results of GO functional enrichment analysis; D-results of KEGG pathway enrichment analysis.

图 5　GO 和 KEGG 富集分析结果

Fig. 5　Enrichment analysis results of GO and KEGG

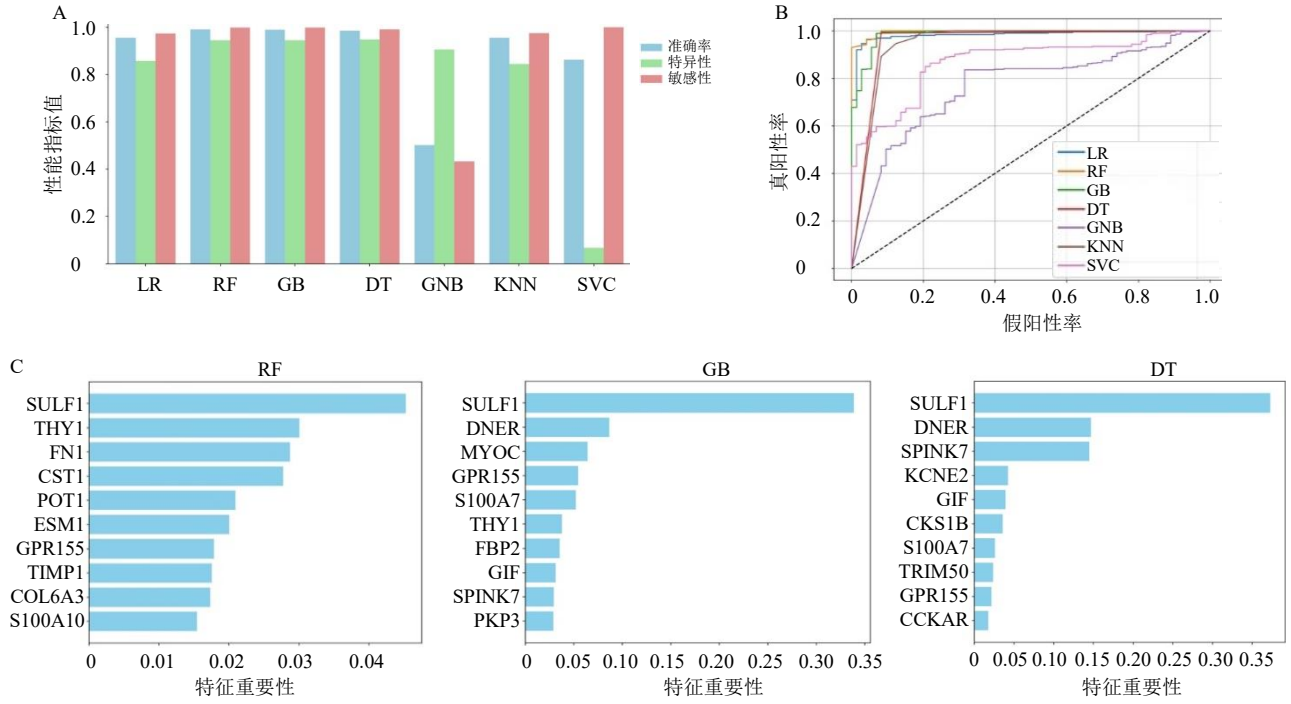
的分子基础[42-43]。

KEGG 通路方面：主要包括肌动蛋白细胞骨架调控（regulation of actin cytoskeleton）、紧密连接（tight junction）、黏附斑（focal adhesion）、磷脂酰

肌醇-3-羟激酶（phosphatidylinositol-3-hydroxykinase，PI3K）/蛋白激酶 B（protein kinase B，Akt）信号通路、ECM-受体相互作用（ECM-receptor interaction）及白细胞介素-17（interleukin-617，IL-17）信号通路等，涵盖细胞骨架重构、促增殖/抗凋亡、炎症及免疫逃逸机制[44]；蛋白质消化与吸收（protein digestion and absorption）、吞噬体（phagosome）及感染相关通路（如人乳头瘤病毒感染）进一步提示代谢、免疫与慢性炎症在胃癌中的协同作用[45]。

2.4 机器学习模型的构建与验证

基于 177 个 DEGs，利用 Python 3.11 实现 7 种经典机器学习算法，采用 10 折交叉验证评估 2 492 个样本的性能。比较准确率、特异性、敏感性（图 6-A）及 ROC 的 AUC（图 6-B）后，RF 模型最优（准确率 0.991、特异性 0.946、敏感性 0.999），GB 与 DT 次之（准确率 0.990/0.984、特异性 0.946/0.948、敏感性 0.997/0.991）。RF 的 AUC 最大，分类能力最强。综合性能与解释性选 RF、GB、DT 为最终模型，用于后续特征重要性分析。



A-7 种机器学习算法性能比较，LR-逻辑回归，RF-随机森林，SVM-支持向量机，KNN-K 近邻算法，GNB-高斯朴素贝叶斯，DT-决策树，GB-梯度提升；B-不同模型 ROC 曲线；C-RF、GB、DT 模型前 10 贡献基因。

A-performance comparison of seven machine learning algorithms, LR-logistic regression, RF-random forest, SVM-support vector machine, KNN-K-nearest neighbors, GNB-gaussian naive bayes, DT-decision tree, GB-gradient boosting; B-ROC curves of different models; C-Top 10 contributing genes for the RF, GB, and DT models.

图 6 机器学习模型构建与关键基因筛选

Fig. 6 Construction of machine learning models and screening of key genes

2.5 基于 SHAP 算法的机器学习模型解释与核心靶点筛选

采用 SHAP 算法解析 RF、GB、DT 模型特征重要性（图 6-C）。3 模型均识别硫酸酯酶 1（sulfatase 1，SULF1）为最重要基因。RF 模型中胸腺细胞表面抗原 1（Thy-1 cell surface antigen，THY1）、纤连蛋白 1（fibronectin 1，FN1）、半胱氨酸蛋白酶抑制剂 SN（cystatin SN，CST1）、端粒保护蛋白 1（protection of telomeres 1，POT1）贡献显著；GB 模型中 Delta/Notch 样表皮生长因子相关受体（Delta/Notch-like EGF-related receptor，DNER）、肌纤蛋白（myocilin，MYOC）、G 蛋白偶联受体 155（G protein-coupled receptor 155，GPR155）、S100 钙结合蛋白 A7（S100 calcium binding protein A7，S100A7）突出；DT 模型中 DNER、Kazal 型丝氨酸蛋白酶抑制剂 7（serine peptidase inhibitor Kazal type 7，SPINK7）、电压门控钾通道亚家族 E 调节亚基 2（potassium voltage-gated channel subfamily E

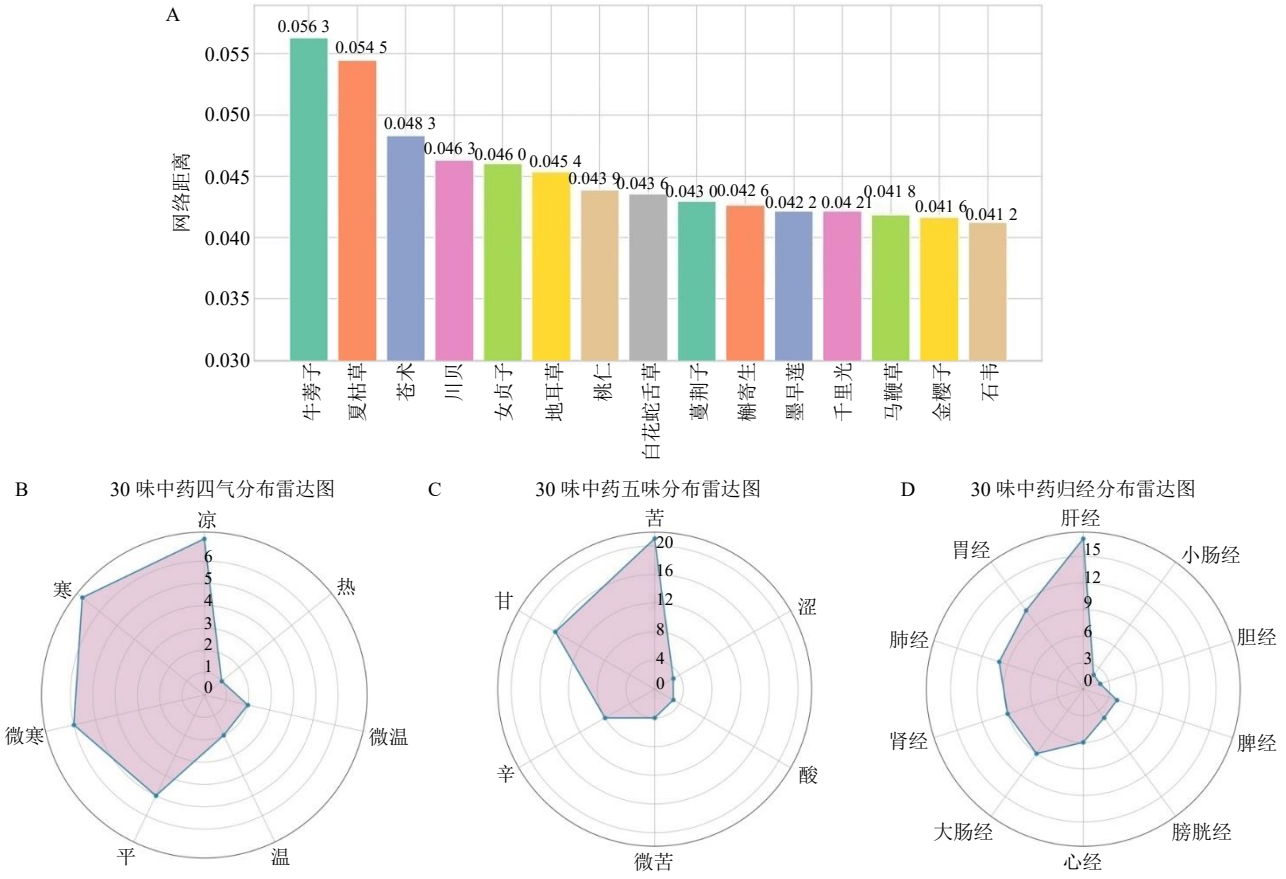
regulatory subunit 2，KCNE2）、胃内因子 GIF（gastric intrinsic factor，GIF）关键。取 3 模型前 10 基因去重，共 20 个用于药物筛选，这些基因依次为 SULF1、THY1、FN1、CST1、POT1、内皮细胞特异性分子 1（endothelial cell-specific molecule 1，ESM1）、GPR155、金属肽酶抑制剂 1（TIMP metallopeptidase inhibitor 1，TIMP1）、VI 型胶原 α3 链（collagen type VI alpha 3 chain，COL6A3）、S100A10、DNER、MYOC、果糖二磷酸酶 2（fructose-bisphosphatase 2，FBP2）、GIF、SPINK7、斑菲素蛋白 3（plakophilin 3，PKP3）、KCNE2、细胞周期蛋白依赖性激酶调节亚基 1B（CDC28 protein kinase regulatory subunit 1B，CKS1B）、三重基序蛋白 50（tripartite motif containing 50，TRIM50）、胆囊收缩素 A 受体（cholecystokinin A receptor，CCKAR）。

关键基因 SULF1、THY1、DNER、SPINK7 与胃癌发生发展密切相关：SULF1 调控肝素硫酸化，影响成纤维细胞生长因子（fibroblast growth factor，FGF）/血管内皮生长因子（vascular endothelial growth factor，VEGF）/转化生长因子-β（transforming growth factor-β，TGF-β）通路，高表达促进增殖、侵袭及血管生成，经 PI3K-Akt/MAPK 增强适应性[46]；THY1 上调驱动 EMT、侵袭及干细胞特性，与预后不良相关[47]；DNER 通过非典型 Notch 通路促进迁移与免疫逃逸[48]；SPINK7 下调破坏上皮屏障，诱发炎症与肿瘤发生[49]。4 种基因分别代表信号激活、黏附改变、免疫异常及上皮保护丧失，可作为潜在诊断标志物与治疗靶点。

2.6　基于网络距离的中药筛选

整合人类 PPI 网络与中药靶点数据，计算中药靶点模块与胃癌 DEGs 模块的网络距离（负值越大，拓扑越近）。牛蒡子、夏枯草等负距离最大，提示具有最高治疗潜力（图 7-A，前 15 位）。

对前 30 味中药统计（图 7-B～D）：四气以凉为主，寒、微寒次之；五味以苦为主，甘、辛次之；归经以肝经为主，胃经、肺经次之；功效以清热药为主，补虚药次之（表 1）。



A-与胃癌差异基因网络距离最小的前 15 味中药；B～D-前 30 味中药的四气、五味、归经分析。

A-top 15 taditional Chinese medicines with smallest network distance to gastric cancer DEGs; B—D-analysis of four *qi*, five flavors, and meridian tropism of the top 30 taditional Chinese medicines.

图 7　关键中药的筛选及其性味归经分析

Fig. 7　Screening of key taditional Chinese medicines and analysis of their properties, flavors, and meridian tropism

表 1 中药功效分类统计

Table 1 Classification statistics of traditional Chinese medicine efficacy

功效分类	中药味数	中药名称
解表药	2	牛蒡子、荆芥炭
清热药	11	夏枯草、川贝、地耳草、白花蛇舌草、千里光、穿心莲、天花粉、槐花、绿豆、小蓟、马鞭草
祛风湿药	2	苍术、草乌
活血化瘀药	2	桃仁、牛膝
补虚药	4	女贞子、墨旱莲、杜仲、续断
收涩药	2	金樱子、柿蒂
化痰止咳平喘药	3	瓜蒌、瓜蒌皮、白蔹
其他	4	石韦、大戟、蔓荆子、槲寄生

3 讨论

胃癌是一种以胃黏膜上皮恶性转化为核心的消化系统肿瘤，其临床表型复杂且异质性显著，涵盖上腹痛、消化道出血、进行性消瘦及腹腔/远处转移等多系统损害。目前尚无根治性疗法，临床干预以手术切除、新辅助化疗及靶向治疗为核心目标，但现有方案对晚期患者的疾病控制作用有限，5 年生存率不足 30%。值得注意的是，胃癌高发于 40～70 岁中老年群体，这一人群正处于家庭支柱与社会贡献的关键阶段，疾病导致的不可逆性器官功能缺损带来了巨大的社会负担。此外，全球医疗资源分配不均导致发展中国家患者难以获取 PD-1 抑制剂或 HER2 靶向药，凸显了开发低成本、多靶点疗法的迫切需求。如果能在中药中挖掘潜力方药开发新辅助方案，将给胃癌患者带来希望的曙光，为中医药现代化提供科学支撑。

本研究重点放在胃癌分子进展的精准干预上。虽然已有研究报道了胃癌相关基因及其信号通路，但本研究通过整合 21 个 GEO 数据集和 WGCNA 网络分析，同时结合机器学习（RF/GB/DT）及 SHAP 算法，系统识别了 177 个核心基因，并进一步筛选出 SULF1、THY1、DNER、SPINK7 等关键基因。相比已有研究，本研究不仅揭示了胃癌进展的分子网络动态特征，还将核心基因与潜在中药干预靶点直接关联，实现了从疾病分子机制到药物预测的跨层次整合分析。有影像学与病理研究表明，肿瘤微环境重塑是评估胃癌侵袭程度的 1 个核心指标。胃癌早期以慢性炎症-癌变过程中的 EMT 及 ECM 重构为特征，随疾病恶化，癌周间质逐渐纤维化并促进远处转移。近年研究表明，阻断胃癌患者 ECM 重塑与 EMT 进程是疾病早期一项重要的治疗策略[40]。通过整合 21 个 GEO 数据集（胃癌 2 125 例、正常 367 例），系统揭示了肿瘤基因动态变化的复杂性。DEGs 与 WGCNA 模块分析表明，胃癌进展不仅涉及细胞骨架调控失调，还与血管生成、PI3K-Akt 信号激活及免疫炎症微环境重塑密切相关。综合富集分析与 PPI 网络获得 177 个核心基因，其中浅黄色模块（r＝0.56，q＜0.01）含 194 个 hub 基因，与肿瘤侵袭表型高度匹配。GO-BP 富集于 ECM 组织、细胞-基质黏附等，GO-CC 集中于含胶原 ECM、黏附斑与内质网腔，GO-MF 突出整合素结合、生长因子结合；KEGG 通路涵盖 actin 骨架调控、局灶性黏附、PI3K-Akt 及 IL-17 信号。这些结果与临床病理一致：ECM 重构促进癌细胞侵袭，黏附斑介导迁移，PI3K-Akt 驱动抗凋亡，IL-17 募集炎症细胞，共同构成胃癌恶性进展的分子基础。

机器学习模型进一步验证了核心基因的诊断价值。RF 模型准确率达 0.991（AUC 最大），SHAP 分析一致识别 SULF1 为首位贡献基因，其次是 THY1、DNER、SPINK7 等。SULF1 高表达通过去硫酸化肝素增强 FGF/VEGF 信号，促进血管生成与侵袭[46]；THY1 上调驱动 EMT 并维持肿瘤干细胞特性，与预后不良相关[47]；DNER 经非典型 Notch 通路促进迁移与免疫逃逸[48]；SPINK7 下调破坏上皮屏障，诱发慢性炎症并为癌变提供土壤[49]。上述基因的异常表达不仅解释了胃癌的异质性侵袭模式，还为多靶点干预提供了分子把手。

通过网络距离策略，本研究从人类 PPI 网络中鉴定出牛蒡子、夏枯草、苍术、川贝、女贞子、地耳草、桃仁等负距离最大的中药（前 15 位）。这些中药功效以清热为主（凉、寒、微寒），五味苦为先，归经肝、胃、肺，精准对应胃癌“热毒-瘀血-阴虚”的中医病机。如《温热逢源》[50]言：“无邪不有毒，热从毒化，变从毒起，瘀从毒结也”，热毒贯穿胃癌全程，早中期以血瘀热毒为主，晚期转阴虚有热，故清热解毒（牛蒡子、夏枯草等）、补虚滋阴（女贞子等）、活血化瘀（桃仁等）3 法并用。现代药理实验验证了其抗肿瘤活性：牛蒡子苷元下调细胞周期蛋白 D1（cyclin D1）、上调 p21/p27，诱导 G_1 期阻滞[51]；夏枯草抑制 miR-155-Wnt/β-连环蛋白（β-

catenin）轴，逆转 EMT 并上调 E-cadherin[52]；女贞子齐墩果酸增强化疗敏感性[53]；桃仁苦杏仁苷促进肿瘤细胞铁死亡[54]。这些成分的作用路径涵盖细胞周期阻滞、EMT 逆转、血管生成抑制及凋亡诱导，提示中药复方可能通过多靶点协同实现更全面的抗肿瘤效应，同时缓解化疗引起的骨髓抑制、恶心呕吐等毒副反应。虽然这些预测仍需实验验证，但本研究为临床提供了潜在的辅助用药选择和多靶点干预策略，尤其适用于化疗耐药或晚期胃癌患者。研究结果为中药现代化应用提供了分子依据，为后续临床前实验和随机对照试验奠定了基础。

本研究通过整合组学数据与网络药理学揭示了胃癌进展的分子网络，并预测了多种具有干预潜力的中药方药。这些发现不仅为胃癌机制研究提供了新视角，也为中医药的精准化与国际化开辟了路径。然而，从计算预测到临床落地仍需跨越诸多科学技术鸿沟。未来需通过跨学科合作、技术创新与随机对照试验，逐步实现胃癌治疗从“减瘤控病”向“带瘤生存”的跨越，最终改善全球数百万胃癌患者的生活质量。

利益冲突 所有作者均声明不存在利益冲突

参考文献

[1] Sung H, Ferlay J, Siegel R L, *et al*. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries [J]. *CA Cancer J Clin*, 2021, 71(3): 209-249.

[2] World Cancer Research Fund. Stomach cancer statistics [EB/OL]. (2024-04-10) [2025-11-09]. https:// www.wcrf.org/preventing-cancer/cancer-statistics/stomach-cancer-statistics/.

[3] Morgan E, Arnold M, Camargo M C, *et al*. The current and future incidence and mortality of gastric cancer in 185 countries, 2020-2040 [J]. *EClin Med*, 2022, 47: 101406.

[4] 国家癌症中心. 中国恶性肿瘤流行病学报告（2023）[R]. 北京: 国家癌症中心, 2023.

[5] 国家卫生健康委员会. 中国胃癌诊疗规范（2023 版）[M]. 北京: 人民卫生出版社, 2023.

[6] Zeng H, Wang C, Song L Y, *et al*. Economic evaluation of FLOT and ECF/ECX perioperative chemotherapy in patients with resectable gastric or gastro-oesophageal junction adenocarcinoma[J]. *BMJ Open*, 2022, 12(11): e060983.

[7] Bang Y J, Van Cutsem E, Feyereislova A, *et al*. Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (*ToGA*): A phase 3, open-label, randomised controlled trial [J]. *Lancet*, 2010, 376(9742): 687-697.

[8] Janjigian Y Y, Ajani J A, Moehler M, *et al*. First-line nivolumab plus chemotherapy for advanced gastric, gastroesophageal junction, and esophageal adenocarcinoma: 3-year follow-up of the phase III CheckMate 649 trial [J]. *J Clin Oncol*, 2024, 42(17): 2012-2020..

[9] Bagheri M, Akrami H. Studying the non-coding RNA expression and its role in drug resistance mechanisms of gastric cancer [J]. *Pathol Res Practice*, 2025, 265: 155742.

[10] 张声生. 中华脾胃病学 [M]. 北京: 人民卫生出版社, 2016: 1895.

[11] Xie W J, Zhang Y S, Tang J Y, *et al*. Efficacy and safety of traditional Chinese medicines as a complementary therapy combined with chemotherapy in the treatment of gastric cancer: An overview of systematic reviews and meta-analyses [J]. *Integr Cancer Ther*, 2024, 23: 15347354231225961.

[12] Dai Z L, Tan C F, Wang J, *et al*. Traditional Chinese medicine for gastric cancer: An evidence mapping [J]. *Phytother Res*, 2024, 38(6): 2707-2723.

[13] Cristescu R, Lee J, Nebozhyn M, *et al*. Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes [J]. *Nat Med*, 2015, 21(5): 449-456.

[14] Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma [J]. *Nature*, 2014, 513(7517): 202-209.

[15] Barrett T, Wilhite S E, Ledoux P, *et al*. NCBI GEO: Archive for functional genomics data sets: Update [J]. *Nucleic Acids Res*, 2013, 41: D991-D995.

[16] Ritchie M E, Phipson B, Wu D, *et al*. Limma powers differential expression analyses for RNA-sequencing and microarray studies [J]. *Nucleic Acids Res*, 2015, 43(7): e47.

[17] Zhang Y Q, Parmigiani G, Johnson W E. ComBat-seq: Batch effect adjustment for RNA-seq count data [J]. *NAR Genom Bioinform*, 2020, 2(3): lqaa078.

[18] Subramanian A, Tamayo P, Mootha V K, *et al*. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles [J]. *Proc Natl Acad Sci USA*, 2005, 102(43): 15545-15550.

[19] LaValley M P. Logistic regression [J]. *Circulation*, 2008, 117(18): 2395-2399.

[20] Chen X, Ishwaran H. Random forests for genomic data analysis [J]. *Genomics*, 2012, 99(6): 323-329.

[21] Byvatov E, Schneider G. Support vector machine applications in bioinformatics [J]. *Appl Bioinformatics*, 2003, 2(2): 67-77.

[22] Zhang Z H. Introduction to machine learning: K-nearest neighbors [J]. *Ann Transl Med*, 2016, 4(11): 218.

[23] Kamel H, Abdulah D, Al-Tuwaijari J M. Cancer classification using Gaussian naive Bayes algorithm [A] // 2019 International Engineering Conference (IEC) [C]. Erbil: IEEE, 2020: 165-170.

[24] de Ville B. Decision trees [J]. *Wires Comput Stat*, 2013, 5(6): 448-455.

[25] Natekin A, Knoll A. Gradient boosting machines, a tutorial

[J]. *Front Neurorobot*, 2013, 7: 21.

[26] Lundberg S M, Lee S I. A unified approach to interpreting model predictions [A] // Neural Information Processing Systems [C]. Long Beach: NIPS, 2017.

[27] Morselli Gysi D, do Valle Í, Zitnik M, *et al*. Network medicine framework for identifying drug-repurposing opportunities for COVID-19 [J]. *Proc Natl Acad Sci USA*, 2021, 118(19): e2025581118.

[28] Yan D Y, Zheng G H, Wang C C, *et al*. HIT 2.0: An enhanced platform for Herbal Ingredients' Targets [J]. *Nucleic Acids Res*, 2022, 50(D1): D1238-D1243.

[29] Liu Z H, Cai C P, Du J W, *et al*. TCMIO: A comprehensive database of traditional Chinese medicine on immuno-oncology [J]. *Front Pharmacol*, 2020, 11: 439.

[30] Ru J L, Li P, Wang J N, *et al*. TCMSP: A database of systems pharmacology for drug discovery from herbal medicines [J]. *J Cheminform*, 2014, 6: 13.

[31] Huang L, Xie D L, Yu Y R, *et al*. TCMID 2.0: A comprehensive resource for TCM [J]. *Nucleic Acids Res*, 2018, 46(D1): D1117-D1120.

[32] Chen X, Zhou H, Liu Y B, *et al*. Database of traditional Chinese medicine and its application to studies of mechanism and to prescription validation [J]. *Br J Pharmacol*, 2006, 149(8): 1092-1103.

[33] Gao H M, Wang Z M, Li Y J, *et al*. Overview of the quality standard research of traditional Chinese medicine [J]. *Front Med*, 2011, 5(2): 195-202.

[34] Wang M X, Carver J J, Phelan V V, *et al*. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking [J]. *Nat Biotechnol*, 2016, 34(8): 828-837.

[35] Szklarczyk D, Santos A, Von Mering C, *et al*. STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data [J]. *Nucleic Acids Res*, 2016, 44(D1): D380-D384.

[36] Menche J, Sharma A, Kitsak M, *et al*. Disease networks. Uncovering disease-disease relationships through the incomplete interactome [J]. *Science*, 2015, 347(6224): 1257601.

[37] Cheng F X, Kovács I A, Barabási A L. Network-based prediction of drug combinations [J]. *Nat Commun*, 2019, 10(1): 1197.

[38] Winkler J, Abisoye-Ogunniyan A, Metcalf K J, *et al*. Concepts of extracellular matrix remodelling in tumour progression and metastasis [J]. *Nat Commun*, 2020, 11(1): 5120.

[39] Leggett S E, Hruska A M, Guo M, *et al*. The epithelial-mesenchymal transition and the cytoskeleton in bioengineered systems [J]. *Cell Commun Signaling*, 2021, 19(1): 32.

[40] Gordon-Weeks A, Yuzhalin A E. Cancer extracellular matrix proteins regulate tumour immunity [J]. *Cancers*, 2020, 12(11): 3331.

[41] Li S Y, Sampson C, Liu C H, *et al*. Integrin signaling in cancer: Bidirectional mechanisms and therapeutic opportunities [J]. *Cell Commun Signal*, 2023, 21(1): 266.

[42] Mao D L, Xu R, Chen H X, *et al*. Cross-talk of focal adhesion-related gene defines prognosis and the immune microenvironment in gastric cancer [J]. *Front Cell Dev Biol*, 2021, 9: 716461.

[43] Chu Y Q, Ye Z Y, Tao H Q, *et al*. Relationship between cell adhesion molecules expression and the biological behavior of gastric carcinoma [J]. *World J Gastroenterol*, 2008, 14(13): 1990-1996.

[44] Hałas-Wiśniewska M, Zawadka P, Arendt W, *et al*. From adhesion to invasion: Integrins, focal adhesion signaling, and actin binding proteins in cervical cancer progression-a scoping review [J]. *Cells*, 2025, 14(20): 1640.

[45] Wang J F, Wang Y, Zhang S W, *et al*. Expression and prognostic analysis of integrins in gastric cancer [J]. *J Oncol*, 2020, 2020: 8862228.

[46] Fang X C, Chen D M, Yang X Y, *et al*. Cancer associated fibroblasts-derived SULF1 promotes gastric cancer metastasis and CDDP resistance through the TGFBR3-mediated TGF-β signaling pathway [J]. *Cell Death Discov*, 2024, 10(1): 111.

[47] Rohan P, Dos Santos E C, Abdelhay E, *et al*. High expression of THY1 in intestinal gastric cancer as a key factor in tumor biology: A poor prognosis-independent marker related to the epithelial-mesenchymal transition profile [J]. *Genes*, 2023, 15(1): 28.

[48] Wang L J, Wu Q, Zhu S, *et al*. Delta/Notch-like epidermal growth factor-related receptor (DNER) orchestrates stemness and cancer progression in prostate cancer [J]. *Am J Transl Res*, 2017, 9(11): 5031-5039.

[49] Wali Z, Neha, Shamsi A, *et al*. The SPINK protein family in cancer: Emerging roles in tumor progression, therapeutic resistance, and precision oncology [J]. *Pharmaceuticals*, 2025, 18(8): 1194.

[50] 朱虹.《鼠疫汇编 温热逢源》临证精解 [M]. 北京: 中国医药科技出版社, 2024: 55.

[51] 杨振方, 王保卫, 刘大勇, 等. 牛蒡子苷元通过调控 SETDB1 表达对胃癌细胞增殖、迁移和侵袭的影响及其机制研究 [J]. 中国药学杂志, 2020, 55(19): 1596-1602.

[52] 李云川, 王刚, 李强, 等. 夏枯草抑制人胃癌细胞上皮间质化的机制 [J]. 西北药学杂志, 2024, 39(2): 93-98.

[53] 陈晓霞, 路晓庆, 刘向荣, 等. 女贞子-黄芪抗胃癌研究及作用机制的探讨 [J]. 中国药物与临床, 2021, 21(23): 3801-3804.

[54] 韦佳宋. 林才志教授运用桂枝茯苓丸加减治疗寒凝血瘀型消化道恶性肿瘤临床探讨 [D]. 南宁: 广西中医药大学, 2025.

[责任编辑 潘明佳]