中医药数据库构建的思路与方法研究进展

张馨月[1,2#], 孟鸿腾[1,2#], 胡亚宣[1,2], 吴明扬[1,2], 张晶晶[1,2], 陈一萱[1,2], 李鑫举[3], 张 晗[1,2], 吕 明[1,2*]

(1. 天津中医药大学, 现代中药创制全国重点实验室, 天津 301617; 2. 天津中医药大学, 现代中医药海河实验室, 天津 301617; 3. 天津中医药大学中医学院, 天津 301617)

摘要: 中医药数据库为深化中医理论、探索病证关系、揭示配伍机制、开发组分中药提供了重要平台支撑。中医药数据库构建以“证候−方剂−草药−成分−靶点−通路−功能−疾病”的关系网络为核心, 并随着计算机科学和高通量实验技术的发展, 进一步开始呈现与人工智能、多模态多组学技术有机结合的发展趋势。本研究介绍了目前中医药数据库的研究现状, 从数据的收集、分析及可视化三个方面综述了相关中医药数据库的研究进展, 系统地讨论了构建数据库的策略和方法。此外, 还揭示了当前中医药数据库所面临的挑战, 并提出了未来可探索的方向和潜在的发展趋势, 旨在为后续相关数据库的构建提供技术支撑, 并为相关领域的研究提供思路。

关键词: 中医药数据库; 网络药理学; 人工智能; 数据收集; 数据分析; 数据可视化

中图分类号: R917 **文献标识码**: A **文章编号**: 0513-4870(2025)09-2679-11

Research progress on the ideas and methods of traditional Chinese medicine database construction

ZHANG Xin-yue[1,2#], MENG Hong-teng[1,2#], HU Ya-xuan[1,2], WU Ming-yang[1,2], ZHANG Jing-jing[1,2], CHEN Yi-xuan[1,2], LI Xin-ju[3], ZHANG Han[1,2], LÜ Ming[1,2*]

(*1. National Key Laboratory of Chinese Medicine Modernization, Tianjin University of Traditional Chinese Medicine, Tianjin 301617, China; 2. Haihe Laboratory of Modern Chinese Medicine, Tianjin University of Traditional Chinese Medicine, Tianjin 301617, China; 3. College of Traditional Chinese Medicine, Tianjin University of Traditional Chinese Medicine, Tianjin 301617, China*)

Abstract: Traditional Chinese medicine (TCM) databases provide a crucial platform for deepening TCM theory, exploring the relationship between disease and syndrome, revealing compatibility mechanism, and developing component-based Chinese medicine. The construction of TCM databases is centered around the multi-level "syndrome-formula-herb-component-target-pathway-function-disease" network. With the advancement of computer science and high-throughput experimental technologies, there is a growing trend of integration with artificial intelligence and multi-omics strategies. This paper reviews the current state of TCM databases, and summarizes the research progress from three aspects of data collection, data analysis, and network visualization. It systematically discusses the strategies and methods for building TCM databases. Additionally, it also reveals the challenges faced by the current TCM databases, and puts forward the future exploration direction and potential development trend,

收稿日期: 2025-01-21; 修回日期: 2025-06-22.

基金项目: 天津科技创新基地建设 (24ZYJDSS00320); 国家自然科学基金青年项目 (82104431); 国家自然科学基金重点项目 (82430121); 现代中药新质生产力科技创新工程专项 (24ZXZKSY00030); 京津冀基础研究合作专项 (23JCZXJC00380); 贵州省科技计划项目 (黔科合支撑[2023]一般100); 天津市教委科研计划项目 (2022KJ148); 组分中药国家重点实验室青苗基金项目 (CBCM2023205).

#共同第一作者.

*通讯作者 Tel: 86-22-59596113, E-mail: minglyu@tjutcm.edu.cn

DOI: 10.16438/j.0513-4870.2025-0080

aiming at providing technical support for the construction of subsequent related databases and offering insights for research in related fields.

Key words: traditional Chinese medicine database; network pharmacology; artificial intelligence; data collection; data analysis; data visualization

随着中医药现代化进程加快, 其理论体系与实践经验获得广泛关注, 当下中医药文化振兴呈现出良好形势, 其实用性探索也在不断纵深发展。然而, 中医药仅靠固化的传统知识体系与理论框架难以支撑现代化发展需求[1]。随着计算机技术的应用, 中医药领域开始探索数据库建设的可能性。中医药数据库发展初期, 主要侧重于对古籍文献的数字化整理和零散的科研数据收集, 虽然规模较小、功能有限, 但为后续的数据库建设奠定了基础。后期则依托信息技术实现临床疗效指标的规范化存储与共享, 助力各指标的合理应用, 以提升中医药研究的科学度与可信度[2]。本研究通过系统分析现有数据库的构建模式, 进一步探寻数据库建立的潜在规律。剖析中医药数据库构建过程中所存在的实际问题, 针对性提出具体解决措施, 以期为中医药数据库的发展提供参考借鉴, 促进中医药数据库的纵深发展。

1 中医药典型数据库

ETCM v2.0[3]数据库是由中国中医科学院中药研究所及中国科学院上海药物研究所合作构建的中医药综合数据库。ETCM v2.0[3]在数据收集和整合方面做出了巨大努力, 确保了数据的全面性和准确性, 是一个权威算法集成、智能网络关联的分析平台。ETCM v2.0[3]的核心竞争力在于其权威数据源、多维分析功能和开源生态。其运用了中药成分靶点预测方法(D3CARP[4])、处方药物相似性评价方法等, 全面解析中药复杂系统与机体分子网络之间的相互作用关系。同时, 与 PubChem、UniProt、GeneCards 等国际权威化学和生物学数据库建立外部链接, 进一步提高中医药的国际影响力。

HERB[5]数据库是由北京中医药大学、中国科学院计算技术研究所和四川大学华西医院合作创建的中药高通量实验和参考数据库。HERB[5]提供了全面的中药药物转录组数据库, 通过重新分析所有可用的高通量实验数据, 为研究人员和药物开发者提供了一个宝贵的资源。2024年11月, HERB[5]数据库更新为HERB 2.0[6], 该版本实现了临床证据整合, 对高通量实验数据进行扩展, 提供基因表达谱自动分析接口辅助用户个性化研究, 构建知识图谱支持高效的数据挖掘, 并优化可视化与分析功能。同时, 手动整理过去十年内发表的参考文献, 给出与中药/成分相关的高可信度靶点和疾病信息。

TCMbank[7]数据库是由中山大学通过自然语言处理 (natural language processing, NLP) 和知识图谱技术建立的数据库。TCMBank[7]数据库是当时最全面、可下载且最大的非商业中医药数据库, 提供了草药成分的3D结构, 并与其他公共数据库如CAS、PubChem等有交叉引用链接。区别于其他中医药数据库, TCMbank[7]将大量的人工智能算法入局中医药发展, 通过3DGT-DDI[8]等AI模型预测中药复方在治疗慢性疾病中的多靶点协同机制, 揭示了中药“多成分−多靶点−多疾病”的复杂作用网络。

BATMAN-TCM 2.0[9]是由国家蛋白质科学中心(北京) 推出的数据库, 这是一个专门为中药药理学研究设计的集成数据库, 有助于中药分子机制的研究以及复杂疾病新药的开发。BATMAN-TCM 2.0[9]提供了当时最大规模中药成分−靶标相互作用数据集, 并能通过疾病组学数据筛选潜在中药, 对于剖析中药原理、找寻复杂疾病潜在药物具有重要意义。

此外, 常用的中医药数据库还有: TM-MC 2.0[10]、TCMSID[11]、LTM-TCM[12]、SoFDA[13]、TMNP[14]、HIT 2.0[15]、superTCM[16]、TCMM[17]、SymMap[18]、TCMID 2.0[19]、YaTCM[20]、CEMTDD[21]、TCMSP[22]、TCM-ID[23]等。

随着现代计算机技术以及互联网信息的不断普及, 数据库逐渐被广泛应用于中医药信息的保存和查询。现代信息技术的普及推动中医药数据库的广泛应用, 目前约20种专业数据库通过整合科学信息、促进全球共享, 推动了中医药现代化进程。然而我国中医药数据库建设仍呈现一定的延迟性, 大量有用的中药信息被淹没在传统典籍中, 对中药学现代化的发展产生影响。本研究基于典型的中医药相关数据库做详细的介绍。中医药数据库的构建主要分为数据收集、分析以及可视化3个部分, 如图1所示[24-30]。

2 数据收集

数据库构建中的数据收集是一个复杂而重要的过程, 涉及数据来源、收集方法、收集流程等多个方面。中医药数据库的数据来源选择需兼顾多样性与权威性。具体而言, 目前已发表的数据库中数据来源主要包括: 文献挖掘, 数据库整合, 临床和基础实验数据及关联分析与预测。在质量评估方面, 建议采用“三

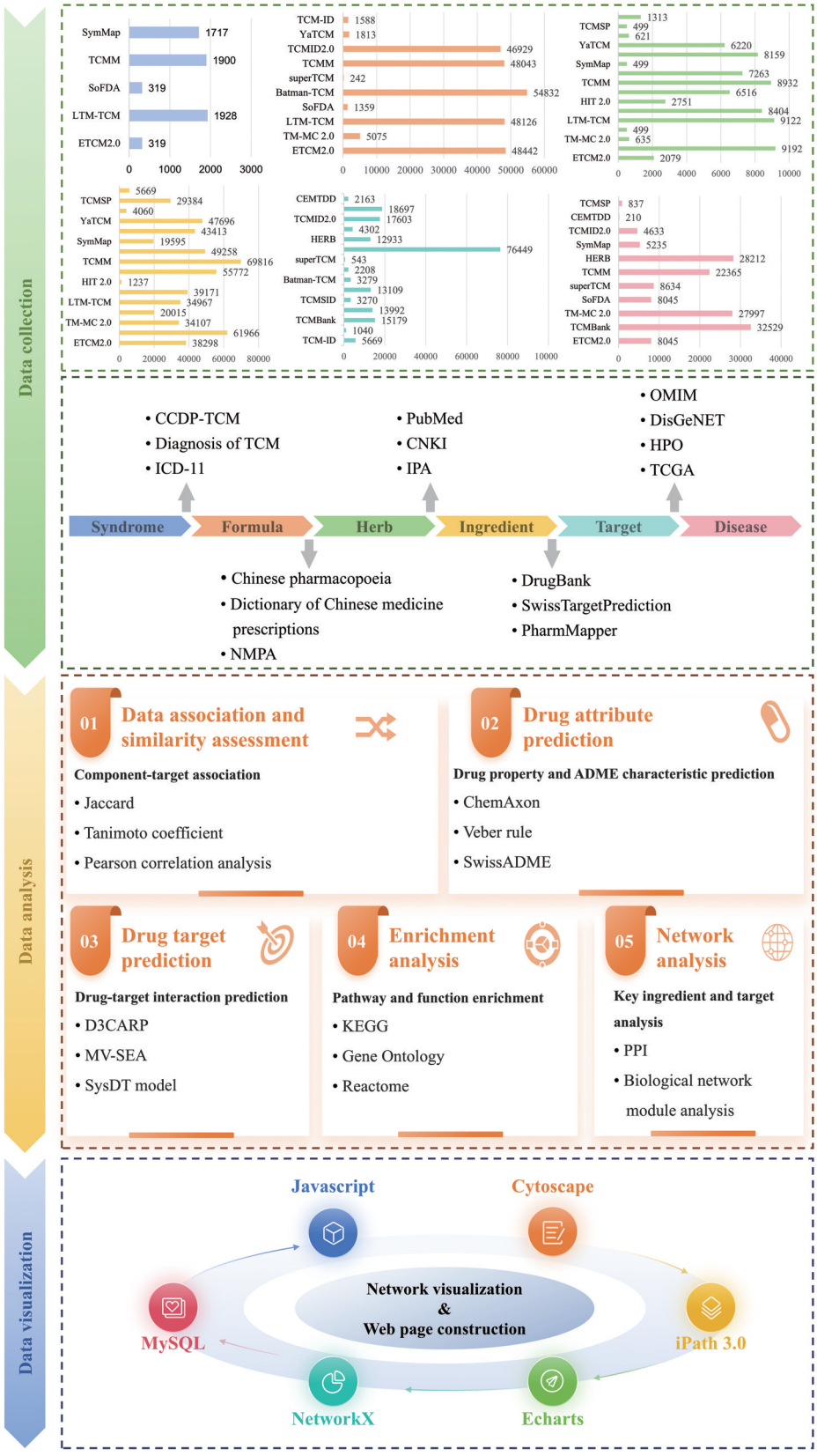


Figure 1 Ideas and methods of traditional Chinese medicine (TCM) database construction. CCDP-TCM: Classification and codes of diseases and patterns of traditional Chinese medicine; ICD: International classification of diseases; IPA: Ingenuity pathway analysis; NMPA: National medical products administration; MV-SEA: Multi-voting similarity ensemble approach; KEGG: Kyoto encyclopedia of genes and genomes; GO: Gene ontology; PPI: Protein-protein interaction

步法”: 可靠性确认、完整检查和权威性验证[31]。随着技术进步, 各数据库逐渐引入了靶点预测、蛋白质相互作用 (protein-protein interaction, PPI) 等数据形式。网络分析技术的提升使得计算科学在中医研究中发挥显著的作用, 各个数据库逐渐形成了“证候−方剂−中药−成分−靶点−通路−功能−疾病”的核心链。在数据量方面, 中医药数据库的规模不断扩大, 数据量从早期的数千个逐步增长至数万甚至更多。数据量的不断增长, 可能会带来数据缺失与低质量问题, 需采用多层次技术策略, 如数据清洗清除异常值, 并删除冗余记录; 填补缺失值, 预测缺失成分的概率权重; 修正低质量数据, 对置信度低的数据启动专家评审流程, 最终提升数据可用率。中医药数据库具体数据来源见表 1[3,6,7,9-13,15-22]。

Table 1 Data sources of TCM database. NATCM: The National Administration of TCM; DO: Disease ontology; NLP: Natural language processing; TTI: Target-target interaction; NCBI: National center of biotechnology information; UMLS: Unified medical language system; GNPS: Global natural products social molecular networking

Database	Category	Counts	Source	Ref.
ETCM v2.0 (2023)	Syndromes	319	Clinical expert consensus and manual verification	[3]
	Formulas	48 442	TCM formulas	
	Chinese patent drugs	9 872	NMPA	
	Herbs	2 079	Chinese Pharmacopoeia (2020 edition), The Fourth National Survey on Chinese Materia Medica Resources, Authoritative Chinese medical books and dictionaries	
	Ingredients	38 298	Chinese Pharmacopoeia (2020 edition), literatures	
	Targets	1 040	D3CARP platform based on BindingDB database	
	Diseases	8 045	MalaCards v5.0, HPO, OMIM, DisGeNET v5.0, ORPHANET v5.49.0	
HERB 2.0 (2024)	Herbs	6 892	The national database for herbs in TCM, iPlant database, CNKI Plant Register database, YaoZhi database	[6]
	Formulas	6 743	Chinese Pharmacopoeia (2020 edition), NMPA, NATCM	
	Ingredients	44 595	PubChem, NPASS	
	Targets	15 515	HIT 2.0, GenBank, GeneCards, TTD	
	Diseases	30 170	DisGeNET, OMIM, HPO, DO	
TCMBank (2023)	Herbs	9 192	NLP, knowledge graph intelligent recognition	[7]
	Ingredients	61 966	NLP, knowledge graph intelligent recognition	
	Targets	15 179	OMIM, HGNC, MeSH, ENsembl, DO, HPO, DisGeNET v5.0	
	Diseases	32 529	OMIM, HGNC, MeSH, ENsembl, DO, HPO, DisGeNET v5.0	
BATMAN-TCM 2.0 (2021)	Formulas	54 832	TCMID, TCM-Suite, LTM-TCM, ITCM	[9]
	Herbs	8 404	TCMID, TCM-Suite, HERB, ITCM	
	Ingredients	39 171	TCMID, TCM-Suite, HERB, ITCM	
	Targets	9 927	SWISS-PROT	
	Pathways	217	KEGG	
	Functions	11 931	GO	
	TTIs	17 068	DrugBank, KEGG, TTD, BATMAN-TCM, HIT, HERB	
TM-MC 2.0 (2023)	Formulas	5 075	Two prescription textbooks and five internal medicine text books used at Korean medicine universities	[10]
	Herbs	635	Korean, Chinese, and Japanese Pharmacopoeias	
	Ingredients	34 107	PubMed	
	Targets	13 992	STITCH v5.0	
	Diseases	27 997	DisGeNET v7.0	
TCMSID (2022)	Herbs	499	Chinese Pharmacopoeia (2015 edition), TCMSP, SymMap	[11]
	Drugs	10 487	DrugBank	
	Ingredients	20 015	TCMSP, SymMap	
	Herb-ingredient associations	50 053	TCMSP, SymMap	
	Targets	3 270	SEA, SwissTargetPrediction, HitPickV2, PPB, PPB2, ChEMBL	
LTM-TCM (2022)	Symptoms	1 928	Public databases, literature mining	[12]
	Formulas	48 126	TCMID, TCM-ID, Chinese Pharmacopoeia (2020 edition), NMPA, GJMLZY	
	Herbs	9 122	CANCERHSP, TCMID, TCMIP, Chinese Pharmacopoeia (2020 edition)	
	Ingredients	34 967	TCM-ID, TCMID, TCM-MESH, TCM-IP, TM-MC	
	Targets	13 109	BATMAN-TCM, ChEMBL, STITCH	
SoFDA (2022)	Syndromes	319	CCDP-TCM, ICD	[13]
	Formulas	1 359	Diagnostics of Chinese medicine, ETCM v2.0	
	Diseases	8 045	MalaCards v5.0, HPO, OMIM, DisGeNET v5.0, ORPHANET v5.49.0	

Continued

Database	Category	Counts	Source	Ref.
HIT 2.0	Herbs	2 751	TCM-ID	[15]
(2021)	Ingredients	7 375	TCM-ID, PubMed	
	Targets	2 208	PubMed	
	Herbal ingredient-target activity pairs	10 031	PubMed	
superTCM	Formulas	242	Book: Chinese Herbal Medicine: Formulas & Strategies	[16]
(2021)	Herbs	6 516	Published books, TCMID 2.0, CMAUP, TCM-Mesh, TM-MC	
	Ingredients	55 772	Published books, TCMID 2.0, CMAUP, TCM-Mesh, TM-MC	
	Targets	543	ChEMBL	
	Diseases	8 634	ICD-10-CM	
	Pathways	254	KEGG	
	Botanical synonyms	11 794	MPNS, NCBI	
TCMM	Formulas	48 043	TCMID, CPMCP, TCMBank, SymMap	[17]
(2021)	Herbs	8 932	TCMID, CPMCP, TCMBank, SymMap	
	Ingredients	69 816	CPMCP, TCMBank, SymMap, PharMeBINet	
	Targets	76 449	CPMCP, TCMBank, SymMap, PharMeBINet, PrimeKG	
	Diseases	22 365	CPMCP, TCMBank, SymMap, PharMeBINet, PrimeKG	
	MM Symptoms	17 079	PharMeBINet, PrimeKG	
	TCM Symptoms	1 900	TCMID, CPMCP, TCMBank, SymMap	
	Syndromes	146	PharMeBINet, PrimeKG	
	Pathways	3 704	PharMeBINet, PrimeKG	
SymMap	TCM Symptoms	1 717	Chinese Pharmacopoeia (2015 edition)	[18]
(2019)	MM Symptoms	961	UMLS	
	Herbs	499	Chinese Pharmacopoeia (2015 edition)	
	Ingredients	19 595	TCMID, TCMSP, TCM-ID	
	Targets	4 302	HIT, TCMSP, HPO, DrugBank, NCBI	
	Diseases	5 235	OMIM, MeSH, Orphanet	
TCMID 2.0	Formulas	46 929	Document mining	[19]
(2018)	Drugs	8 182	DrugBank	
	Herbs	8 159	TCM-ID	
	Ingredients	43 413	TCM-ID, HIT	
	Prescription ingredients	1 045	CNKI	
	Targets	17 603	HIT, STITCH, OMIM, DrugBank	
	Diseases	4 633	OMIM	
	Herbal mass spectra	778	CNKI	
	Mass spectrometry of ingredients	3 895	GNPS	
YaTCM	Formulas	1 813	Formulas of Chinese medicine, Chinese Pharmacopoeia (2020 edition)	[20]
(2018)	Herbs	6 220	TCMID, TCMSP, PubChem	
	Ingredients	47 696	TCMID, TCMSP, PubChem	
	Targets	18 697	TCMID, TTD, OMIM	
	Predicted targets	1 907	MV-SEA	
	Pathways	390	KEGG	
CEMTDD	Herbs	621	Kazakh medicine blog, Uygur medicine blog, Chinese Pharmacopoeia, Chinese Materia Medica	[21]
(2015)	Ingredients	4 060	CNKI, SCI finder, Web of Science, Google	
	Targets	2 163	SEA Dock, Google	
	Diseases	210	PubMed, CNKI, SCI finder, Web of Science, Google	
TCMSP	Herbs	499	Chinese Pharmacopoeia (2020 edition)	[22]
(2014)	Ingredients	29 384	PubChem, ChEMBL, ChemSpider, ISIS Draw 2.5, Sybyl 6.9, Open Babel, InChIKey	
	Targets	3 311	DrugBank, HIT, SysDT	
	Diseases	837	TTD, PharmGKB	

3 数据分析

随着中医药数据库的发展, 中药复方有效成分组学被引入中医药数据库的构建, 由“单一靶点、单一药物、单一疾病”的研究思路向“多靶点、多途径”相互协同、相互制约的网络思想转变[32]。构建中医药数据库逐步形成以“证候–方剂–中药–成分–靶点–通路–功能–疾病”关联网络为基本的研究流程, 对收集到的数据进行分析。但各数据库对数据的分析方法各有侧重, 包括研究中药活性成分作用靶点的数据库 TCMSP[22], 中医药整合药理学的数据库 BATMAN-TCM 2.0[9], 中医证候类数据库 SoFDA[13], 结合高通量实验数据的数据库 HERB[5], 民族药理学相关数据库 CEMTDD[21]等。

除各数据库之间的差异外, 在研究方法上又存在相同的研究思路。本研究总结了各数据库的分析方法, 提炼了关键的数据分析环节。

目前中医药数据库的数据分析涉及以下几个技术环节: ① 多维关联和相似性评估: 包括化学相似性、生物网络关联两部分, 用于评估多个数据集间的关联性或分子相似性。ETCM v2.0[3]、SoFDA[13]数据库进行相似性分析时使用了Jaccard及Cosine相似度评分。Jaccard相似性主要适用于基因或蛋白质等二元数据的存在性分析; Cosine相似性主要适用于数值型向量如基因表达量、蛋白质序列特征向量等。二者各有其使用的范围, 随着多模态数据的融合, 混合相似性评估方法可能会成为相似性分析的重要工具。② 药物属性预测: 通过 ChemAxon、SwissADME[30]、ADMETlab 2.0[33]等工具预测成分药物动力学和毒理学属性, 计算每种成分的关键药物样性质。ETCM v2.0[3]、TCMSID[11]等数据库在药物属性预测上使用了 ADMETlab, 该工具支持对化合物的物理化学性质、药代动力学参数等进行全面预测。平台采用多种机器学习与深度学习技术, 可用于大规模的药物属性预测。③ 药物靶点预测及相互作用: 利用分子对接工具, 如SysDT模型、D3CARP平台[4]识别药物潜在作用靶点, 揭示多靶点协同机制。靶点预测的主要环节包括构建靶点数据库、反向对接、虚拟筛选。若靶点范围较广, 可对大规模靶点库进行高通量虚拟筛选, 快速缩小候选靶点范围[34]。④ 富集分析: 这一环节主要用于识别在特定数据集中显著的生物学功能、通路或特征。包括超几何分布检验、基因集富集分析 (gene set enrichment analysis, GSEA) 及多重检验校正。⑤ 网络分析: 通过构建PPI网络, 对关键节点进行筛选, 完成系统验证, 从而从全局视角解析药物−靶点−疾病的复杂互作关系。PPI网络推动从“单一靶点”向“网络分析”的研究转变, 未来也可结合人工智能驱动进行相互作用预测。数据库具体数据分析方法见表2[3,6,7,9-16,18-22]。

Table 2 Data analysis methods and tools of TCM database. QED: The quantitative estimate of drug-likeness; MS: Mass spectrometer; EL: Ensemble learning; HNN: Hybrid neural network; BH: Benjamini-Hochberg; FDR: False discovery rate; GSEA: Gene set enrichment analysis

Category	Database	Analysis tool	Application	Advantage	Limitation	Ref.
Multidimensional association and similarity assessment	ETCM v2.0, SoFDA	Jaccard similarity assessment	Quantization of the correlation between datasets	Simple and convenient algorithm	Ignore weight considerations and collection size constraints	[3,13]
	HERB 2.0	Data-driven connectivity mapping pipeline	Pairwise similarity assessment between the datasets and CMap gene expression profiles	Support complex relationship modeling and can process data efficiently	Data quality risks and high maintenance costs	[6]
	TCMSID	ECFP4	Molecular precise structural similarity search	Accurate capturing of molecular structural differences and broad applicability	Prone to false negatives with functionally similar compounds	[11]
	TCMSID	FCFP6	Functional group combinations similarity search	Strong functional generalization ability and broad neighborhood coverage	Overgeneralization leads to overlooking critical structural features resulting in false positives	[11]
	TCMSID	Tanimoto coefficient	Molecular similarity assessment	Intuitive quantitative evaluation and efficient computing	Neglecting feature weights introduces feature bias	[11]
	SoFDA	Cosine similarity assessment	Evaluation of indirect correlation between datasets	Consider word's frequency and weight	Inadequate sparse data responsiveness	[13]
	SoFDA	Feature shuffling	Relational reliability assessment	Applicable to multiple models and capable of directly reflecting features	Neglecting synergistic feature interactions introduces bias	[13]
	TMNP	Pearson correlation coefficient	Evaluation of correlations between variables	Sensitive to linear relationships and broad applicability	Relying on the normality assumption leads to an inability to infer causality	[14]
	superTCM	KEGG global maps	Ingredient-target-pathway association	Broad coverage and manual annotation support	Large-scale pathways with high complexity	[16]
	superTCM	RDKit	Fingerprint generation and similarity calculation	Optimize underlying algorithms and efficient molecular data processing	Risk of generating erroneous structures for molecules with atypical configurations	[16]

Continued

Category	Database	Analysis tool	Application	Advantage	Limitation	Ref.
Medicinal attributes prediction	TM-MC 2.0	ChemAxon	Acquisition of pharmaceutical chemical properties	Support full-process chemical information	Non-open-source platform	[10]
	TM-MC 2.0	QED equation	Drug similarity assessment	Rapid assessment and intuitive quantitative scoring	Neglecting synthetic accessibility introduces design risks	[10]
	TM-MC 2.0	Veber's rule	Oral bioavailability assessment	Augmented with molecular rigidity and polarity descriptors	Overly restrictive rules fail to encompass all critical factors	[10]
	TCMSID	ADMETlab 2.0	ADME/T-related properties prediction	Data-driven modeling and multi-task learning framework	Dynamic updating requirements and data dependencies	[11]
	LTM-TCM	SwissADME	ADME/T-related properties prediction	Integration of experimental data and machine learning models for enhanced predictive accuracy	Predictions may be influenced by complex factors and require experimental validation	[12]
	TCMID 2.0	MS spectra	Analysis of molecular structure and composition of compound	High sensitivity and specificity	High operational complexity	[19]
Drug-target prediction	ETCM v2.0	D3CARP	Target prediction, virtual screening	Computationally efficient without requiring conformational sampling	Fails to capture the effects of spatial arrangement or stereochemistry on activity	[3]
	TCMBank	Ligand-based EL model	Molecular activity prediction	Cost-effective for early-stage drug discovery	Neglect of target dynamics	[7]
	TCMBank	HNN-based EL model	Drug-target affinity prediction	High-accuracy prediction of complex molecular interactions	Requires sufficient data and computational resources	[7]
	LTM-TCM	AutoDock Vina	Molecular docking, virtual screening	High computational speed with enhanced accuracy	Limited handling of receptor flexibility, complex conformational changes require integration with other methods	[12]
	YaTCM	MV-SEA	Target prediction	Integration of multi-source similarity data with high predictive accuracy	High computational complexity, requires optimized algorithms and hardware support	[20]
	CEMTDD	LibDock score	Evaluation of intermolecular binding capacity	Rapid processing optimized for high-throughput screening	Simplified treatment of conformational flexibility and solvation effects	[21]
	CEMTDD	SEA dock	Molecular docking, target prediction	Feature fusion enhance model prediction accuracy	Computationally intensive	[21]
	CEMTDD	Reverse docking	Reverse docking, target prediction	Identification of small molecule multi-target mechanisms of action	Dependency on target library data integrity, high false positive rate	[21]
	TCMSP	SysDT	Prediction of potential targets of compounds	Integration of computational predictions and experimental validation	Dependency on high-quality target databases	[22]
Enrichment analysis	ETCM v2.0	Reactome enrichment analysis	Pathway enrichment analysis	Fine-grained pathway annotation supporting cross-omics analysis	Infrequent updates, limited species coverage	[3]
	ETCM v2.0, YaTCM, BATMAN-TCM 2.0	GO enrichment analysis	Functional enrichment analysis	Unified terminology and comprehensive functionality	Complex hierarchical structure may lead to redundant enrichment	[3,9,20]
	HERB 2.0, SymMap	Fisher's exact test	Obtain reliable associations with statistical significance	High precision and flexibility with distribution-free characteristics	Limited to small-sample data or sparse data	[6,18]

Continued

Category	Database	Analysis tool	Application	Advantage	Limitation	Ref.
	HERB 2.0, SymMap	BH multiple test correction	Control the FDR	Computationally efficient for high-throughput data processing	Sensitivity to threshold selection, inability to fully eliminate false positives	[6,18]
	TMNP, BATMAN-TCM 2.0	GSEA	Evaluation of gene expression differences	Function-oriented association with established biological functions or pathways	Dependency on the quality of predefined gene sets	[9,14]
	YaTCM, BATMAN-TCM 2.0, HIT 2.0	KEGG enrichment analysis	Pathway enrichment analysis	Well-structured interaction network covering multi-species pathway data	Limited to known canonical pathways	[9,15,20]
	LTM-TCM	clusterProfiler	Identify target enrichment pathways	Comprehensive functionality with robust visualization and diverse enrichment approaches	Limited support for non-model organisms, poor performance with large-scale data	[12]
Network analysis	CEMTDD, TCMSP, TCMID 2.0, SoFDA	PPI network	Screen for relationships among herbs, ingredients, targets and diseases	Systems-level analysis of complex biological processes	Experimental data processing is prone to false positives and false negatives	[13,19, 21,22]

4 数据可视化

数据可视化是一种将复杂的网络数据通过图形化手段进行直观展示的技术。它利用节点、边、颜色、大小等视觉元素来表示网络中的实体及其相互关系, 使得用户能够更容易地理解和分析网络数据的结构、特征以及动态变化。常见的网络可视化的工具如Gephi、Cytoscape等, 提供了丰富的网络可视化功能和算法支持。在中医药数据库的可视化步骤中, 常使用D3.js、ECharts等JavaScript库, 通过提供强大的底层应用程序编程接口和丰富的配置选项, 允许开发者灵活控制网络图的每个细节, 且支持高度定制化的网络可视化效果开发。然而JavaScript的灵活性和动态性可能导致代码难以理解和维护。特别是在大型项目中, 如果没有良好的代码结构和规范, 可能会出现代码混乱不堪的情况[35]。

通过对本研究选取的部分数据库进行分析发现, 数据可视化的方法呈逐步多样化趋势。早期 (如2014年的TCMSP[22]) 的数据库, 数据可视化方法相对简单, 主要依赖于Cytoscape Web等网络工具生成视觉网络, 并允许用户下载和进行进一步的拓扑分析。Cytoscape Web凭借其轻量级、易用性和跨平台特性使其成为早期中医药数据库构建中常用的可视化工具。其强项在于生物网络, 最吸引人的是大量的插件和极强的数据融合与互动能力[36]。然而, 它的性能在处理大规模数据时可能受限, 相对适合中小规模网络的可视化需求。随着时间推移, 数据可视化方法逐渐丰富, 如2018年的YaTCM[20]开始使用ECharts和NetworkX等更先进的可视化技术, 提供交互网络的可视化。其中, ECharts建立在轻量级的Canvas库ZRender之上, 提供了丰富的图表类型, 如折线图、柱状图、散点图等, 适用于数据分析和展示场景[37]。然而, ECharts对复杂网络图 (如生物网络、社交网络) 的支持较弱且需要一定的前端开发经验。NetworkX的优势则在于其强大的网络分析功能、易用性以及与Python科学计算生态 (如NumPy、Matplotlib) 的无缝集成。但其可视化功能较弱, 通常需要依赖Matplotlib或其他库 (如Gephi、Cytoscape) 进行图形展示。到2023年, 一些数据库 (如ETCM v2.0[3]、TCMBank[7]) 已经提供了基于JavaScript的网络可视化工具, 允许用户创建、修改和探索生物网络, 并且支持动态布局定制和多种格式的导出。同时, 数据库与外部资源的整合能力增强。早期数据库主要关注于中医药内部数据的整合和可视化, 与外部资源的链接相对较少。随着时间的推移, 越来越多的数据库开始提供对外部公共数据库的引用链接, 如CAS、PubChem等, 这有助于用户获取更全面的信息。例如, TCMBank[7]不仅提供了跨数据库链接, 还允许用户根据研究目的自定义网络的动态布局, 这进一步增强了数据库的实用性和灵活性。除此之外, 数据可视化与数据分析的结合更加紧密。前期数据库的数据可视化主要侧重于展示数据关系, 而数据分析功能相对较弱。技术进步促使部分数据库开始将数据可视化与数据分析紧密结合, 如ETCM v2.0[3]提供的网络可视化工具不仅支持可视化展示, 还允许用户进行网络分析和修改。数据库具体数据可视化方法见表3[3,6,7,11,12,16,17,19-22]。

5 讨论和展望

构建中医药数据库第一步是尽可能完整地获取“证候−方剂−中药−成分−靶点−通路−功能−疾病”的详细信息。在这一过程中主要存在以下问题: ① 由于收集

Table 3 Data visualization methods of TCM database. CSS: Cascading style sheets; DBMS: Database management system; NER: Named entity recognition; PDO: PHP data objects

Database	Data visualization method	Ref.
ETCM v2.0	JavaScript	[3]
HERB 2.0	Neo4j graph database, Apache ECharts	[6]
TCMBank	Dynamic layout customization of multi-level biological networks, JSME molecule editor	[7]
TCMSID	Python based web framework Django, HTML5, CSS, JavaScript, automatic generation of network analysis diagrams	[11]
LTM-TCM	uWSGI web, MySQL DBMS, MySQL database AnalyticDB, NER	[12]
superTCM	ipath3.0, Typescript, React JS, Ant Design, PDO, docker, ChemDoodle	[16]
TCMM	RxGen	[17]
TCMID 2.0	vis.js, TCM components-target network visualization	[19]
YaTCM	Django, Angula, Nginx, JSME, RDKit, EChart, NetworkX, PostgreSQL, Interactive network visualization technology	[20]
CEMTDD	Cytoscape web	[21]
TCMSP	Cytoscape web, NetworkAnalyzer, PHP, HTML, CSS, Apache 2.2.22	[22]

数据过程中数据来源单一, 对于中医药没有一个统一标准, 收集到的数据缺乏完整性和规范性。② 中药的用量改变也会影响中药的疗效, 缺乏对于中药量效关系的考虑。③ 目前中医药发展迅速, 缺乏对数据的及时更新。构建数据库的第二步是对收集到的数据进行数据分析, 包括: 成分–靶点预测、药物相似性评估和药物属性预测、富集分析、网络分析等。这一过程常结合计算机技术和人工智能等, 但也存在使用技术单一, 分析结果存在假阳性, 缺乏临床试验验证的问题, 因此需不断结合新理论新技术, 多学科交叉使分析结果更具说服力。构建中医药数据库的第三步是完善数据检索与数据可视化。现存中医药数据库利用网络可视化工具提供交互网络可视化技术, 允许用户自定义网络并供用户下载。但是由于各个数据库各有偏重, 存在局限性, 因此可加强各个数据库间的关联, 提高数据质量。面对当前中医药数据库存在的挑战, 未来中医药数据库的发展可以聚焦于以下方面, 以实现更高水平的数据整合、共享与应用。首先, 加强数据标准化与质量控制。可以借鉴数据库 IPA (ingenuity pathway analysis) 的构建范式, 该数据库是基于专家收集和管理的知识库, 将文献中已证实的发现通过机器学习结合专家审核的方式生成交互式网络。通过人工提取和验证实验观察结果, 确保数据的准确性[24]。建议未来中医药数据库建设可构建“人机协同”的混合架构, 在人工把控数据质量的基础上引入 AI 训练模型, 实现古籍文献与实验数据的智能对齐, 最终形成统一数据标准体系。其次, 促进数据共享与开放。积极构建开放平台, 鼓励和支持数据资源的跨领域共享。此外, 实现数据实时更新与维护。建立完善的数据更新机制, 确保数据库内容的时效和前沿。

在中医药数据库建立过程中, “成分–靶点”关系是其核心环节。因为中药多成分、多靶点的作用特点, 网络靶点预测技术是一种快速预测成分靶标的有效手段, 根据预测原理的不同, 大致可分为: ① 基于配体的预测方法, 如 ETCM v2.0[3]使用的 D3CARP[4]平台; ② 基于受体的靶标预测, 如 TCM-ID[23]使用的 INVDOCK[38]工具; ③ 机器学习的方法, 如 TCMBank[7]使用的神经网络技术; ④ 组合应用的靶点筛选, 目前结合多种技术组合应用逐渐成为一种趋势, 如 TCMSID[11]数据库结合了多重靶点预测工具进行靶点预测。这些技术与中医药数据库结合推动了中医药数据库的发展, 但技术本身仍有一定的局限性, 药物靶标预测仍存在预测不精准等问题, 解决这些问题仍需不断结合新理论新技术。多层次网络深度解析与高通量多模态多组学验证已成为网络靶标发展新趋势。

目前, 随着网络药理学广泛应用与标准化制定, 中药新药研发的技术平台被相继研制应用。由李梢教授团队历时 20 余年研发而成的 UNIQ 系统则是中医药与人工智能、大数据等前沿信息技术交叉创新的标志性成果。本研究所提及的 TCMSP[22]平台、ETCM v2.0[3]平台、BATMAN-TCM 2.0[9]数据库、SymMap[18]数据库以及 TCMID 2.0[19]数据库均为网络药理学用于中药新药研发提供重要平台与技术手段。2024 年 10 月 12 日发布的《网络药理学应用于中药新药研发专家共识总论》[39]是《网络药理学评价方法指南》[40]制定后的首部应用专家共识, 标志网络药理学在中药新药研发领域迈出了重要一步。网络药理学方法促进了计算科学围绕“证候–方剂–中药–成分–靶点–通路–功能–疾病”为核心整理收集大量数据, 建立相关中医药数据库。

在中医药数据库的现代化进程中, 人工智能技术的深度融合正成为突破数据异构性、提升知识发现效率的核心驱动力, 可以促进中医药复杂作用原理解析[41]。PubTator 3.0 是一个生物医学文献资源平台, 使用了深度学习转换器模型 AIONER 进行实体命名, 同时使用统一关系提取模型 BioREx 完成关系提取[42]。以 ChatGPT 为典型代表的生成式人工智能已广泛应用

于数据库构建中, 但ChatGPT在训练时经常会产生推理幻觉[43,44]。PubTator 3.0增强了GPT-4大型语言模型, 将ChatGPT4.0与PubTator API集成, 可以显著提高其响应的真实性和可验证性[42]。DeepSeek大模型本地化部署应用是人工智能在中医药数据库构建中的具体实践方向之一, 该模型具备开源性、高性能与低成本优势, 能够给使用主体带来效率提高、辅助决策等功用[45]。未来, 在数据库构建环节中搭建DeepSeek局域网络与开源模型, 有望提升数据信息以及智能系统运行的安全性与可控性[46]。PubTator等大语言模型的深度应用可以为数据库的关键环节提供更优的方案, 推动数据库现代化转型和高质量发展。这些技术为中医药数据库构建结合更大规模、更多功能、更精准的人工智能提供了可能。

作者贡献: 吕明负责提供选题与撰写思路、修改指导; 张馨月、孟鸿腾负责撰写及修改; 陈一萱、胡亚宣、吴明扬、张晶晶负责查阅和整理文献资料; 李鑫举、张晗参与文章内容的讨论。

利益冲突: 本文所有作者声明不存在利益冲突关系。

References

[1] Zhou GS, Wang C, Zhong TW, et al. Visualization analysis of "internet plus Chinese medicine" research based on citespace [J]. Guid J Tradit Chin Med Pharm (中医药导报), 2022, 28: 164-170.

[2] Wei M, Dai GH, Gao WL. Construction thoughts of traditional Chinese medicine clinical efficacy evaluation index database [J]. Shandong J Tradit Chin Med (山东中医杂志), 2024, 43: 680-685.

[3] Zhang Y, Li X, Shi Y, et al. ETCM v2.0: an update with comprehensive resource and rich annotations for traditional Chinese medicine [J]. Acta Pharm Sin B, 2023, 13: 2559-2571.

[4] Shi Y, Zhang X, Yang Y, et al. D3CARP: a comprehensive platform with multiple-conformation based docking, ligand similarity search and deep learning approaches for target prediction and virtual screening [J]. Comput Biol Med, 2023, 164: 107283-107290.

[5] Fang S, Dong L, Liu L, et al. HERB: a high-throughput experiment- and reference-guided database of traditional Chinese medicine [J]. Nucleic Acids Res, 2021, 49: D1197-D1206.

[6] Gao K, Liu L, Lei S, et al. HERB 2.0: an updated database integrating clinical and experimental evidence for traditional Chinese medicine [J]. Nucleic Acids Res, 2025, 53: D1404-D1414.

[7] Lv Q, Chen G, He H, et al. TCMBank-the largest TCM database provides deep learning-based Chinese-Western medicine exclusion prediction [J]. Signal Transduct Target Ther, 2023, 8: 127.

[8] He H, Chen G, Chen C. 3DGT-DDI: 3D graph and text based neural network for drug-drug interaction prediction [J]. Brief Bioinform, 2022, 23: bbac134.

[9] Kong X, Liu C, Zhang Z, et al. BATMAN-TCM 2.0: an enhanced integrative database for known and predicted interactions between traditional Chinese medicine ingredients and target proteins [J]. Nucleic Acids Res, 2024, 52: D1110-D1120.

[10] Kim SK, Lee MK, Jang H, et al. TM-MC 2.0: an enhanced chemical database of medicinal materials in Northeast Asian traditional medicine [J]. BMC Complement Med Ther, 2024, 24: 40.

[11] Zhang LX, Dong J, Wei H, et al. TCMSID: a simplified integrated database for drug discovery from traditional Chinese medicine [J]. J Cheminform, 2022, 14: 89.

[12] Li X, Ren J, Zhang W, et al. LTM-TCM: a comprehensive database for the linking of traditional Chinese medicine with modern medicine at molecular and phenotypic levels [J]. Pharmacol Res, 2022, 178: 106185.

[13] Zhang Y, Wang N, Du X, et al. SoFDA: an integrated web platform from syndrome ontology to network-based evaluation of disease-syndrome-formula associations for precision medicine [J]. Sci Bull (Beijing), 2022, 67: 1097-1101.

[14] Li P, Zhang H, Zhang W, et al. TMNP: a transcriptome-based multi-scale network pharmacology platform for herbal medicine [J]. Brief Bioinform, 2022, 23: bbab542.

[15] Yan D, Zheng G, Wang C, et al. HIT 2.0: an enhanced platform for herbal ingredients' targets [J]. Nucleic Acids Res, 2022, 50: D1238-D1243.

[16] Chen Q, Springer L, Gohlke BO, et al. SuperTCM: a biocultural database combining biological pathways and historical linguistic data of Chinese materia medica for drug development [J]. Biomed Pharmacother, 2021, 144: 112315.

[17] Ren Z, Ren Y, Li Z, et al. TCMM: a unified database for traditional Chinese medicine modernization and therapeutic innovations [J]. Comput Struct Biotechnol J, 2024, 23: 1619-1630.

[18] Wu Y, Zhang F, Yang K, et al. SymMap: an integrative database of traditional Chinese medicine enhanced by symptom mapping [J]. Nucleic Acids Res, 2019, 47: D1110-D1117.

[19] Huang L, Xie D, Yu Y, et al. TCMID 2.0: a comprehensive resource for TCM [J]. Nucleic Acids Res, 2018, 46: D1117-D1120.

[20] Li B, Ma C, Zhao X, et al. YaTCM: yet another traditional Chinese medicine database for drug discovery [J]. Comput Struct Biotechnol J, 2018, 16: 600-610.

[21] Huang J, Zheng Y, Wu W, et al. CEMTDD: the database for elucidating the relationships among herbs, compounds, targets and related diseases for Chinese ethnic minority traditional drugs [J]. Oncotarget, 2015, 6: 17675-17684.

[22] Ru J, Li P, Wang J, et al. TCMSP: a database of systems pharmacology for drug discovery from herbal medicines [J]. J Chemin-

form, 2014, 6: 13.

[23] Chen X, Zhou H, Liu YB, et al. Database of traditional Chinese medicine and its application to studies of mechanism and to prescription validation [J]. Br J Pharmacol, 2006, 149: 1092-1103.

[24] Kramer A, Green J, Pollard JJ, et al. Causal analysis approaches in ingenuity pathway analysis [J]. Bioinformatics, 2014, 30: 523-530.

[25] Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018 [J]. Nucleic Acids Res, 2018, 46: D1074-D1082.

[26] Daina A, Michielin O, Zoete V. SwissTargetPrediction: updated data and new features for efficient prediction of protein targets of small molecules [J]. Nucleic Acids Res, 2019, 47: W357-W364.

[27] Amberger JS, Bocchini CA, Scott AF, et al. OMIM.org: leveraging knowledge across phenotype-gene relationships [J]. Nucleic Acids Res, 2019, 47: D1038-D1043.

[28] Pinero J, Sauch J, Sanz F, et al. The DisGeNET cytoscape app: exploring and visualizing disease genomics data [J]. Comput Struct Biotechnol J, 2021, 19: 2960-2967.

[29] Kohler S, Gargano M, Matentzoglu N, et al. The human phenotype ontology in 2021 [J]. Nucleic Acids Res, 2021, 49: D1207-D1217.

[30] Daina A, Michielin O, Zoete V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules [J]. Sci Rep, 2017, 7: 42717.

[31] Li XR, Wang CJ, Wang XY, et al. Development and applications of traditional Chinese medicine databases [J]. Chin Tradit Herb Drugs (中草药), 2025, 56: 293-304.

[32] Liu AL, Du GH. Network pharmacology: new guidelines for drug discovery [J]. Acta Pharm Sin (药学学报), 2010, 45: 1472-1477.

[33] Xiong G, Wu Z, Yi J, et al. ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties [J]. Nucleic Acids Res, 2021, 49: W5-W14.

[34] Tang JN, Sun Y. Application of computational biology in pharmaceutical research [J]. Acta Pharm Sin (药学学报), 2024, 59: 2192-2198.

[35] Zhou CY. Research on the application of JavaScript technology in web front-end development [J]. Comput Eng Software (软件), 2024, 45: 175-177.

[36] Wang S, Zhang Y, Wu YD. Survey on network topology visualization [J]. Chin J Netw Inf Secur (网络与信息安全学报), 2018, 4: 1-17.

[37] Fu TD, Li WY, Wang SX, et al. Visualization analysis of recruitment information data based on python crawler technology [J]. Comput Knowl Technol (电脑知识与技术), 2024, 20: 77-82.

[38] Ji Z L, Wang Y, Yu L, et al. *In silico* search of putative adverse drug reaction related proteins as a potential tool for facilitating drug adverse effect prediction [J]. Toxicol Lett, 2006, 164: 104-112.

[39] Li S, Xiao W. General expert consensus on application of network pharmacology in research and development of new traditional Chinese medicine drugs [J]. China J Chin Mater Med (中国中药杂志), 2024, 49: 5055-5068.

[40] Li S. Network pharmacology evaluation method guidance-draft [J]. World J Tradit Chin Med, 2021, 7: 165-166.

[41] Zhang N, Wang XY, Han B, et al. Research progress of artificial intelligence technology in pharmacology of traditional Chinese medicine [J]. Acta Pharm Sin (药学学报), 2025, 60: 550-558.

[42] Wei CH, Allot A, Lai PT, et al. PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge [J]. Nucleic Acids Res, 2024, 52: W540-W546.

[43] Jin Q, Leaman R, Lu Z. Retrieve, summarize, and verify: how will ChatGPT affect information seeking from the medical literature? [J]. J Am Soc Nephrol, 2023, 34: 1302-1304.

[44] Tian S, Jin Q, Yeganova L, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health [J]. Brief Bioinform, 2024, 25: bbad493.

[45] Yang J, Li CJ. DeepSeek large model localized deployment applications: functional review, emerging risks, and governance paths [J]. J Unit Front Sci (统一战线学研究), 2025, 27: 1-14.

[46] Wei YM, Jia K, Zeng RX, et al. Artificial intelligence innovation development and governance reform under DeepSeek breakthrough effect [J]. E-Gov (电子政务), 2025, 22: 2-39.