基于多源知识图谱的疾病潜在药物发现研究

陈星羽[1] 侯跃芳[1] 赖书兰[2] 梅佳月[3] 李梓萌[1] 韩玛蔓[1]

([1] 中国医科大学健康管理学院 沈阳 110122
[2] 中国医学科学院/北京协和医学院医学信息研究所 北京 100020
[3] 大连医科大学附属第一医院 大连 116021)

〔摘要〕 **目的/意义** 构建多源知识图谱，提出基于知识图谱的药物发现方法，为疾病新药研发提供参考依据。**方法/过程** 设计知识图谱架构，优选 7 种异构数据库，经知识抽取、数据处理与清洗、知识融合构建知识图谱。基于图谱和 SemMedDB 建立推理规则及关联路径，改进加权链路预测方法，提出可用于药物发现的综合知识发现方法。**结果/结论** 以阿尔茨海默病为例，构建含 67 780 个实体、282 870 个三元组的多源异构疾病知识图谱，预测潜在治疗药物 184 种。

〔关键词〕 疾病知识图谱；药物发现；知识推理；路径发现；链路预测
〔中图分类号〕 R－058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673－6036.2025.07.007

Study on Potential Drug Discovery for Diseases Based on Multi－source Knowledge Graphs

CHEN Xingyu[1], HOU Yuefang[1], LAI Shulan[2], MEI Jiayue[3], LI Zimeng[1], HAN Yuman[1]

[1] School of Health Management, China Medical University, Shenyang 110122, China; [2] Institute of Medical Information, Chinese Academy of Medical Sciences&Peking Union Medical College, Beijing 100020, China; [3] The First Affiliated Hospital of Dalian Medical University, Dalian 116021, China

〔**Abstract**〕 **Purpose/Significance** To construct multi－source knowledge graphs, and to propose drug discovery methods based on knowledge graphs, so as to provide reference basis for the research and development of new drugs for diseases. **Method/Process** The architecture of the knowledge graph is designed, and seven heterogeneous databases are selected. The knowledge graph is constructed through knowledge extraction, data processing and cleaning, and knowledge fusion. Based on the graph and SemMedDB, inference rules and association paths are established, and the weighted link prediction method is improved. A comprehensive knowledge discovery method applicable to drug discovery is proposed. **Result/Conclusion** Taking Alzheimer's disease as an example, a multi－source disease knowledge graph containing 67 780 entities and 282 870 triples is constructed, and 184 potential drugs are pvedicted.

〔**Keywords**〕 disease knowledge graph; drug discovery; knowledge reasoning; path discovery; link prediction

1 引言

知识图谱是基于图的数据结构，本质是知识存储库，由代表实体的节点和代表实体间语义关系的边组成[1]。知识发现是从大量零散信息中提取有价

〔修回日期〕 2025－05－13
〔作者简介〕 陈星羽，硕士研究生；通信作者：侯跃芳，博士，教授，硕士生导师。
〔基金项目〕 辽宁省教育厅高校基本科研项目（项目编号：LJ112410159061）。

值新知识的过程[2]。生物医学知识图谱能连接各类生物医学实体及其关系，提供有效组织、存储和查询生物医学知识的方式。利用知识图谱管理海量医学数据，可融合与关联分散、异构的医学知识，提升数据利用价值[3]，利于知识发现，尤其是疾病潜在药物发现。

医学知识图谱构建中，数据筛选、知识抽取、知识表示、知识融合等方法至关重要。蔡妙芝等[4]选取 PubMed 收录的高水平糖尿病专题文献，基于统一医学语言系统（unified medical language system，UMLS）进行知识表示，利用 SemRep 抽取“主语－谓语－宾语”（subject－predicate－object，SPO）三元组。Zhu Q 等[5]综合 34 个权威数据库中罕见病的基因、蛋白质、遗传表型等关键信息，构建罕见病知识图谱。张君冬等[6]以 5 个国内公开大型医学数据库为数据源，通过知识重组、实体对齐、关系融合构建医学知识图谱并开展药物预测实证研究。但多数知识图谱研究数据来源局限，对实体间语义关系与权重表达不充分。因此，本研究优选多源异构生物医学数据，设计知识图谱架构，充分表达实体间语义关系及权重，通过图数据库存储并可视化，构建疾病知识图谱。

随着药物信息学数据的不断累积，基于知识图谱发现疾病潜在治疗药物成为研究热点。基于知识图谱的知识发现方法主要包括推理规则、路径发现和链路预测等[7]。张晗等[8]获取多数据库知识关联，以药物重定位为实证，采用路径搜索和链路预测推理药物在肿瘤治疗中的新用途。Nian Y 等[9]利用知识图谱补全算法预测疾病与药物等实体的新关系，发现有可能预防或延缓神经退行性病变的药物。对比发现，推理规则准确且可解释；路径发现能发现深层隐含知识，但准确性和可解释性较差；链路预测简单高效，但易忽略节点间语义关系。本研究优化改进基于知识图谱的知识发现方法，为提高结果可解释性和有效性，拟综合采用各种策略开展疾病潜在药物发现研究。

本研究基于多源知识图谱进行疾病潜在药物发现，并以阿尔茨海默病为例进行实证分析。以非结构化文本数据和结构化知识库数据为数据源，构建疾病知识图谱，综合运用推理规则、路径发现及链路预测 3 种知识发现方法实现潜在药物发现，该方法可为疾病新药研发及临床治疗提供参考依据。

2 多源知识图谱构建及药物发现

本研究构建多源知识图谱，并基于其发现疾病潜在治疗药物，见图 1。

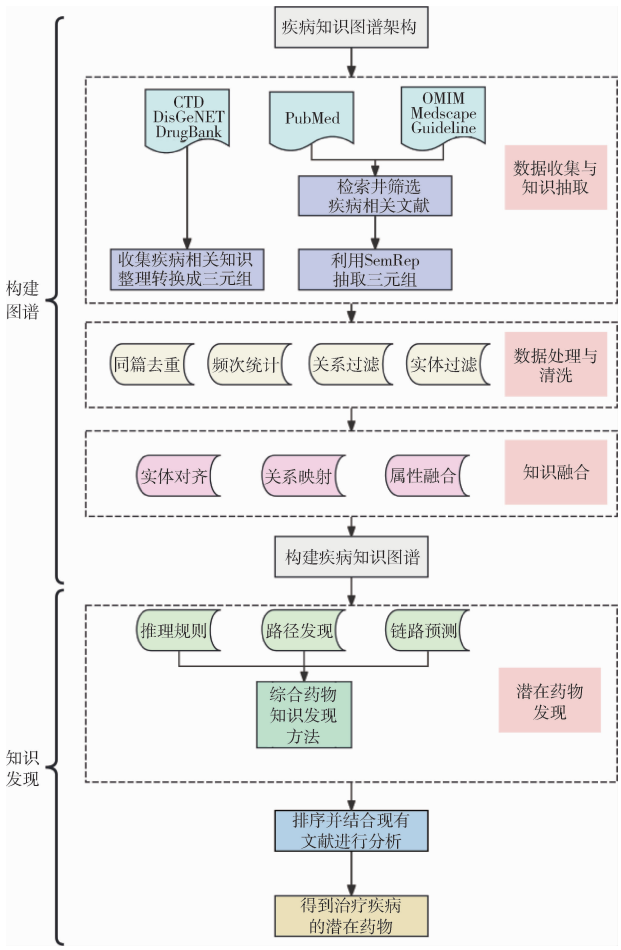

图 1 基于多源知识图谱的疾病潜在药物发现流程

2.1 疾病知识图谱构建

2.1.1 多源疾病知识图谱架构 基于 UMLS 词表设计知识图谱架构，定义 SPO 三元组中实体、关系的命名规范及属性，见图 2。实体名称、实体编码、语义类型及关系名称均参照 UMLS；定义 7 种疾病相关实体类别，包括解剖、化学物质及药物、紊乱、基因及分子序列、生物、生理和人工干预过

程；关系来源指 SPO 三元组来源数据库名称；文献频次指 SPO 三元组来源的 PubMed 文献数量；审查状态定义为 3 个级别，从高到低依次为人工审查型（curated）、已知型（known）和推断型（inferred）。

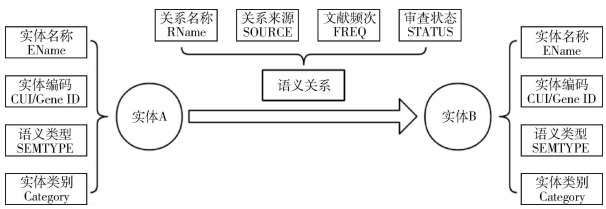

图 2 多源疾病知识图谱架构

2.1.2 数据收集与知识抽取 优选多源异构数据，非结构化文本数据来自 PubMed、OMIM、Guideline、Medscape 和疾病相关临床指南，结构化知识库数据来自 CTD、DisGeNET 和 DrugBank。在 PubMed 检索某疾病相关文献，下载标题和摘要；从 OMIM 和 Medscape 下载该疾病全文介绍；在多种数据库查询并下载相关临床指南。利用 SemRep 从非结构化文本数据（如 PubMed 文献标题和摘要）中抽取 SPO 三元组，存入 SemMedDB。由于其中未涵盖最新文献，须持续使用 SemRep 抽取更新。从 CTD 知识库下载某疾病相关的化学物质、基因及其之间的相互作用；从 DisGeNET 知识库下载基因－疾病关联、疾病－疾病关联数据；从 DrugBank 下载疾病相关的药物数据、药物－靶点关联数据，并整理为 SPO 三元组。

2.1.3 数据处理与清洗 （1）同篇去重。针对同一篇 PubMed 文献中可能抽取出相同 SPO 三元组的情况，对同一文献中的重复 SPO 三元组进行归并。（2）频次统计。为量化评估 SPO 三元组在不同文献中的影响力和可信度，统计每个 SPO 三元组的出现次数，记录为文献频次。（3）关系过滤。为提高知识图谱专指度，剔除相关度不高的语义关系和否定谓词。（4）实体过滤。在 SemRep 抽取结果中，剔除少量范畴宽泛的概念。

2.1.4 知识融合与知识存储 对不同来源数据进行融合处理，实现异构数据源间语义互通。（1）实体对齐。以 UMLS 为标准，利用 Postman 软件规范不同来源数据的实体名称和编码，通过编程实现实体语义类型与类别的归属对齐。（2）关系映射。依据 UMLS 中的语义关系，对来自结构化知识库的三元组关系进行映射，见表 1。（3）属性融合。对不同来源语义关系的“审查状态”属性进行定义和融合，PubMed 定义为 Known，Guideline、OMIM、Medscape 和 DrugBank 定义为 Curated，CTD 和 DisGeNET 根据原库标识分别定义为 Curated、Known 和 Inferred。整合多个来源于不同数据库的相同三元组的审查状态属性，取级别最高的作为最终审查状态属性。

表 1 来源于结构化知识库的三元组关系映射

数据来源	关系类型	原关系名称	映射后关系名称
CTD	化学物质－疾病	marker/mechanism	CAUSES
		therapeutic	TREATS
		inferred	ASSOCIATED_WITH
	基因－疾病	marker/mechanism	CAUSES
		therapeutic	CAUSES
		inferred	ASSOCIATED_WITH
	化学物质－基因	increases	STIMULATES
		decreases	INHIBITS
		affects	AFFECTS
DisGeNET	基因－疾病	无	ASSOCIATED_WITH
	疾病－疾病	无	ASSOCIATED_WITH
DrugBank	药物－疾病	无	TREATS
	药物－靶点	无	AFFECTS

2.2 疾病潜在药物发现

基于 SemMedDB[10] 中存储的三元组，参考生物医学知识，构建推理规则和关联路径，改进链路预测方法，提出综合方法体系，以发现疾病潜在药物。

2.2.1 构建推理规则 构建 66 条药物可能治疗疾病的推理规则，见图 3，其中 C 代表化学物质，G 代表基因，D 代表疾病，L 代表生物，P 代表生理。例如，如果 SemMedDB 三元组显示药物 C_1 能抑制基因 G_1，而基因 G_1 增强疾病 D_1，则可推理出药物 C_1 有可能治疗疾病 D_1。

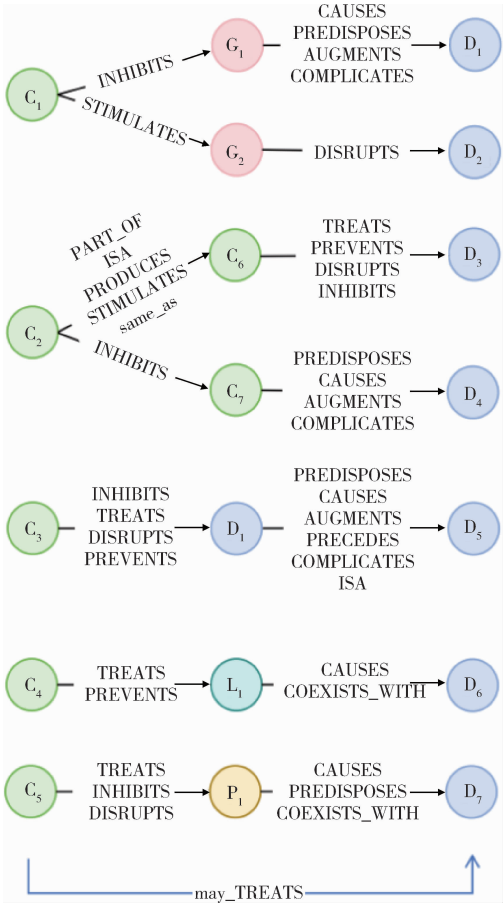

图 3 药物可能治疗疾病的推理规则（部分）

2.2.2 构建关联路径 本研究团队[11]已构建 506 条药物可能治疗疾病的 3 – hop 关联路径，每条路径含两个中间节点。例如，如果药物 A 是（ISA）化学物质 B，化学物质 B 抑制（INHIBITS）基因 C，基因 C 增强（AUGMENTS）疾病 D，则药物 A 可能治疗（may_TREATS）疾病 D，见图 4。

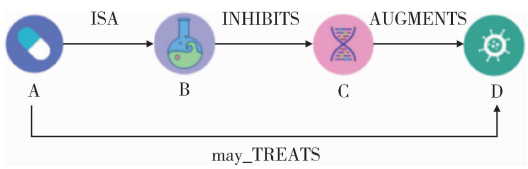

图 4 药物可能治疗疾病的路径发现规则示例

2.2.3 改进链路预测方法 采用并改进加权 adamic – adar 算法（weighted adamic – adar，WAA）[12]进行链路预测计算。考虑实体间语义关系，将有明确指向的语义谓词的权重设为频次 ×1，无明确指向的语义谓词的权重设为频次 ×0.5。

2.2.4 提出综合方法体系 对比推理规则、路径发现及链路预测 3 种方法发现[11]，推理规则预测效果较好，路径发现预测效果不如推理规则，但能增强发现深度，二者可互补。推理规则与路径发现依据规则预测，链路预测基于网络中间节点数量和权重预测，两类方法原理不同。因此融合 3 种方法提出综合药物发现策略：先融合推理规则与路径发现结果，取并集整合；再将并集结果与改进权重的 WAA 链路预测结果结合，取交集精炼，作为最终疾病潜在药物发现结果，以提高预测结果准确性和全面性。

3 实证研究

3.1 阿尔茨海默病知识图谱构建

以阿尔茨海默病为例进行实证研究。按前述数据收集方法收集阿尔茨海默病相关多源异构数据。非结构化文本数据方面，在 PubMed 检索到 122 797 篇文献，以该疾病规范名称在其他数据库查询并下载数据。结构化知识库数据方面，在 DisGeNET 知识库获取 22 476 条疾病 – 疾病关联数据、3 398 条基因 – 疾病关联数据；在 DrugBank 知识库获取药物 11 种、阿尔茨海默病药物 – 靶点数据 100 条。按前述方法框架，对收集的阿尔茨海默病相关多源异构数据进行知识抽取、处理与清洗、融合，构建阿尔茨海默病知识图谱。该知识图谱含 282 870 个三元组、67 780 个实体和 28 种语义关系。将三元组数据导入 Neo4j 图数据库存储并可视化，局部示例，见图 5，右侧列有节点“阿尔茨海默病”的属性。

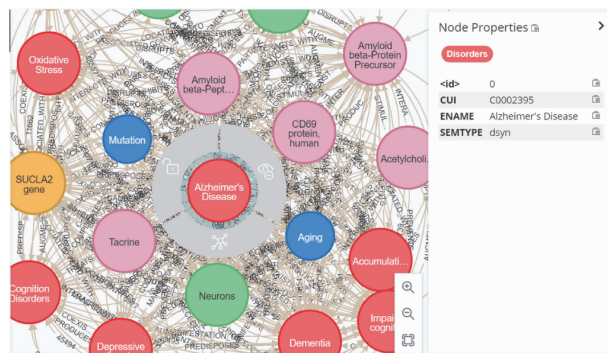

图 5 阿尔茨海默病知识图谱局部展示

注：不同颜色表示不同的实体类别，红色表示疾病（紊乱），黄色表示基因及分子序列，浅粉色表示化学物质及药物，浅绿色表示解剖学，蓝色表示生理过程，深粉色表示人工干预过程，深绿色表示生物种类。

3.2 阿尔茨海默病潜在药物发现结果

应用前述方法共预测出可能治疗阿尔茨海默病的药物 3 759 种，包含了阿尔茨海默病知识图谱中 Curated 药物 104 种、Known 药物 2 861 种，均为已知药物，占比 79.08%，体现了该方法的可靠性。

排除已知药物，进一步分析其他药物。经人工审核，去除结果中泛指的药物类。针对阿尔茨海默病，筛选推理规则与路径发现得到的 SPO 频次均大于等于 3 的化学物质及药物，共得到 184 种。将潜在药物先按推理规则发现频次降序排列，再依次按路径发现频次、链路预测得分降序排列，排名前 10 位的潜在治疗药物或化学物质，见表 2。

表 2 预测的阿尔茨海默病潜在治疗药物（排名前 10 位）

序号	药物名称	推理规则频次（次）	路径发现频次（次）	链路预测得分（分）
1	NR4A2 protein, human	9	752	0.36
2	Isoaspartate	6	397	0.19
3	Glimepiride	6	270	12.70
4	Platelet – Derived Growth Factor	5	194	3.33
5	Anakinra	5	184	0.17
6	Acetoacetates	5	93	0.23
7	Selenocysteine	5	88	0.31
8	alpha – Linolenic Acid	5	82	1.29
9	beta – Actin	5	27	1.56
10	Nociceptin	4	18	0.21

3.3 阿尔茨海默病潜在药物分析

结合现有文献及知识库，对预测的阿尔茨海默病潜在治疗药物进行分析，其中治疗可能性较高的 3 种药物如下。一是人类 NR4A2 蛋白（NR4A2 protein, human），对于中脑多巴胺能神经元的发育、功能和维持至关重要。研究[13]发现，敲除 NR4A2 会显著加重阿尔茨海默病病理症状，而激活 NR4A2 可减缓年龄相关记忆衰退，减少小鼠大脑神经炎症。二是格列美脲（glimepiride），能降低血浆血糖和糖化血红蛋白水平。研究[14]表明，其可通过抑制 BACE1 活性阻止淀粉样蛋白产生，有望成为治疗 2 型糖尿病相关阿尔茨海默病的潜在药物。三是乙酰乙酸酯（acetoacetates，AcAc），研究[15]发现，其通过促进脑源性神经营养因子和抑制炎症来改善患阿尔茨海默病小鼠的记忆力，表明其可能有益于阿尔茨海默病患者的正常脑功能，对阿尔茨海默病具有治疗作用。

4 结语

本研究设计疾病知识图谱架构，优选生物医学多源异构数据，经知识抽取、数据处理与清洗、知识融合构建疾病知识图谱。随后构建推理规则及 3 – hop 关联路径，改进链路预测方法，提出融合推理规则、路径发现与链路预测方法的综合方法，以发现疾病潜在治疗药物。以阿尔茨海默病为例开展实证研究，构建多源阿尔茨海默病知识图谱，并利用 Neo4j 进行知识存储与可视化展示。采用综合知识发现策略，共获取 3 759 种可能治疗阿尔茨海默病的药物，经筛选得到 184 种潜在治疗药物，为阿尔茨海默病治疗药物深入研究提供线索。本研究存在一定局限性，方法体系有人工参与环节，后续将提升自动化水平。未来拟开放共享多源疾病知识图谱，构建疾病知识推荐平台，方便生物医学专业人员快速检索信息，为疾病诊断和治疗提供参考与辅助。

作者贡献：陈星羽负责数据处理与分析、论文撰写；侯跃芳负责研究设计、论文修订；赖书兰、梅佳月负责数据收集与处理、论文撰写与修订；李梓萌、韩玛蔓负责数据处理。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 漆桂林，高桓，吴天星．知识图谱研究进展［J］．情报工程，2017，3（1）：4 – 25.

2 杜建．面向复杂决策和知识发现的医学知识不确定性计算方法［J］．医学信息学杂志，2022，43（7）：32 – 38.

（下转第 52 页）

－frame and multi－scale fusing gate network for accurate segmentation of plaques in ultrasound videos [J]. Computers in biology and medicine, 2023, 163 (9): 107091.
6 BISWAS M, SABA L, CHAKRABARTTY S, et al. Two－stage artificial intelligence model for jointly measurement of atherosclerotic wall thickness and plaque burden in carotid ultrasound: a screening tool for cardiovascular/stroke risk assessment [J]. Computers in biology and medicine, 2020, 123 (8): 103847.
7 VILA M D M, REMESEIRO B, GRAU M, et al. Semantic segmentation with DenseNets for carotid artery ultrasound plaque segmentation and CIMT estimation [J]. Artificial intelligence in medicine, 2020, 103 (3): 101784.
8 RONNEBERGER O, FISCHER P, BROX T. U－Net: convolutional networks for biomedical image segmentation [C]. Munich: International Conference on Medical Image Computing and Computer－assisted Intervention, 2015.
9 马巧梅, 梁昊然, 郎雅琨. 融合残差模块的 U－Net 肺结节检测算法 [J]. 计算机工程与设计, 2021, 42 (4): 1058－1064.
10 肖慧, 方威扬, 林铭俊, 等. 基于两阶段分析的多尺度颈动脉斑块检测方法 [J]. 南方医科大学学报, 2024, 44 (2): 387－396.
11 GAN H, ZHOU R, OU Y, et al. A region and category confidence－based multi－task Network for carotid ultrasound image segmentation and classification [EB/OL]. [2024－07－02]. https://arxiv.org/abs/2307.00583.
12 XIE M, LI Y, XUE Y, et al. Two－stage and dual－decoder convolutional U－Net ensembles for reliable vessel and plaque segmentation in carotid ultrasound images [C]. Miami: 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), 2020.
13 CHEN L C. Rethinking atrous convolution for semantic image segmentation [EB/OL]. [2024－12－05]. https://arxiv.org/abs/1706.05587v3.
14 CHEN L C, ZHU Y, PAPANDREOU G, et al. Encoder－decoder with atrous separable convolution for semantic image segmentation [C]. Munich: The European Conference on Computer vision (ECCV), 2018.
15 ZHANG Z, LIU Q, WANG Y. Road extraction by deep residual U－Net [J]. IEEE geoscience and remote sensing letters, 2018, 15 (5): 749－753.
16 JHA D, SMEDSRUD P H, RIEGLER M A, et al. ResuNet++: an advanced architecture for medical image segmentation [C]. San Diego: 2019 IEEE International Symposium on Multimedia (ISM), 2019.

(上接第 44 页)

3 范媛媛, 李忠民. 中文医学知识图谱研究及应用进展 [J]. 计算机科学与探索, 2022, 16 (10): 2219－2233.
4 蔡妙芝, 李晓瑛, 赵嘉玮, 等. 基于 SPO 语义三元组的疾病知识发现 [J]. 数据分析与知识发现, 2022, 6 (1): 134－144.
5 ZHU Q, NGUYEN D T, GRISHAGIN I, et al. An integrative knowledge graph for rare diseases, derived from the genetic and rare diseases information center [J]. Journal of biomedical semantics, 2020, 11 (1): 13.
6 张君冬, 杨松桦, 严颖, 等. 跨医学体系下医疗知识图谱的构建与药物预测研究——以动脉粥样硬化为例 [J]. 情报理论与实践, 2024, 47 (2): 178－188.
7 胡正银, 刘蕾蕾, 代冰, 等. 基于领域知识图谱的生命医学学科知识发现探析 [J]. 数据分析与知识发现, 2020, 4 (11): 1－14.
8 张晗, 安欣宇, 刘春鹤. 基于多源语义知识图谱的药物知识发现: 以药物重定位为实证 [J]. 数据分析与知识发现, 2022, 6 (7): 87－98.
9 NIAN Y, HU X, ZHANG R, et al. Mining on Alzheimer's diseases related knowledge graph to identity potential AD－related semantic triples for drug repurposing [J]. BMC bioinformatics, 2022, 23 (S6): 407.
10 U.S. National Library of Medicine. Access to SemMedDB database download [EB/OL]. [2024－02－01]. https://lhncbc.nlm.nih.gov/temp/SemRep_SemMedDB_SKR/SemMedDB_download.html.
11 梅佳月. 基于疾病知识图谱的知识发现研究: 以小细胞肺癌为实证 [D]. 沈阳: 中国医科大学, 2024.
12 吕林媛, 周涛. 链路预测 [M]. 北京: 高等教育出版社, 2013.
13 YU H, WANG F, JIA D, et al. Pathological features and molecular signatures of early olfactory dysfunction in 3xTg－AD model mice [J]. CNS neuroscience & therapeutics, 2024, 30 (2): e14632.
14 LIU F, WANG Y, YAN M, et al. Glimepiride attenuates Aβ production via suppressing BACE1 activity in cortical neurons [J]. Neuroscience letters, 2013, 557 (12): 90－94.
15 WU X J, SHU Q Q, WANG B, et al. Acetoacetate improves memory in Alzheimer's mice via promoting brain－derived neurotrophic factor and inhibiting inflammation [J]. American journal of Alzheimer's disease and other dementias, 2022, 37 (9): 240484181.