基于知识流动的生物医学研究前沿识别：一种自监督图聚类框架

李永洁 王龙超 孙轶楠 唐小利
（中国医学科学院医学信息研究所 北京 100020）

[摘要] **目的/意义** 生物医学领域知识体量激增与载体异构化的复杂性，对传统前沿识别方法提出了挑战。本文旨在探讨构建一种量化“科学-技术”知识溢出路径的计算框架，以揭示从基础研究发现到技术应用之间的隐性关联与演化规律，从而为科研选题与研发布局提供数据驱动的决策支持。**方法/过程** 本文提出了一种融合隐性语义与异构拓扑关联的研究前沿识别框架。首先，整合论文与专利数据构建“科学-技术”二部知识流动网络，利用PubMedBERT抽取文本语义并基于K近邻方法构建语义增强边，形成稠密的语义-拓扑耦合结构。其次，设计了三模态门控编码器，以自适应融合节点的内容、时间与结构特征，生成统一的表征。最后，通过自监督联合优化策略协同学习节点表征与社区结构，并基于链路预测概率定义“前沿指数”，实现高转化潜力研究前沿的识别。 **结果/结论** 在乳腺癌领域的实证研究表明，本框架有效克服了传统二部引用网络的稀疏性问题，能够识别出三类典型的前沿社区：“理论爆发型”、“产业成熟型”与“科学-技术并行突破型”。可视化分析进一步揭示了从基础研究社区（如三阴性乳腺癌分子分型）向应用研究社区（如靶向联合用药、AI影像诊断）的隐性知识流动。研究结果表明，本框架在语义深度和前瞻性预警方面具有优势，可作为生物医学领域科技情报分析的有效手段之一。

[关键词] 前沿识别；知识流动；图聚类；链路预测；乳腺癌

Identifying Biomedical Research Frontiers via Knowledge Flow: A Self-Supervised Graph Clustering Framework

LI Yongjie, WANG Longchao, SUN Yinan, TANG Xiaoli
Institute of Medical Information, Chinese Academy of Medical Sciences, Beiing 100020

[Abstract] **Purpose/Significance** Driven by the exponential growth of biomedical information and the increasing heterogeneity of knowledge sources, conventional frontier identification techniques are struggling to maintain efficacy. This study develops a computational framework designed to quantify knowledge spillover pathways between science and technology, aiming to uncover the latent associations and evolutionary dynamics bridging basic research discoveries and technological applications.This framework offers instrumental decision support for optimizing research topic selection and R&D strategic planning.**Method/Process** This study develops an identification framework fusing implicit semantics and heterogeneous topological linkages. The process begins with the construction of a bipartite knowledge flow network spanning papers and patents. Leveraging PubMedBERT for semantic encoding and K-nearest neighbor for edge augmentation, we establish a dense semantic-topological structure. Subsequently, a tri-modal gated encoder is introduced to adaptively integrate nodes' textual, temporal, and structural attributes into a unified representation. The framework utilizes a self-supervised joint optimization strategy to concurrently optimize node representations and community detection. Ultimately, research frontiers with significant translational potential are identified using a custom ‘Frontier Index' derived from link prediction probabilities. **Result/Conclusion** Empirical results in the field of breast cancer demonstrate that the proposed framework

[修回日期] 2026-03-30
[作者简介] 李永洁，助理研究员，发表论文10余篇；通信作者：唐小利，研究馆员。
[基金项目] 中国医学科学院医学与健康科技创新工程项目(项目编号：2021-I2M-1-033)；国家科技图书文献中心2025年专项任务—重点领域文献信息监测与整合服务（项目编号：2025XM29)。

effectively mitigates the sparsity issues inherent in traditional bipartite citation networks. It successfully identifies three archetypal frontier communities: 'Theoretical Burst,' 'Industrial Maturation,' and 'Sci-Tech Resonance.' Visualization analysis further elucidates the latent knowledge flows emanating from basic research clusters (e.g., molecular subtyping of triple-negative breast cancer) toward applied research sectors (e.g., targeted combination therapies and AI-assisted imaging diagnosis). These findings suggest that the framework excels in semantic depth and proactive early-warning capabilities, serving as a robust computational tool for science and technology (S&T) intelligence analysis in the biomedical domain.

[Keywords] identifying frontiers；knowledge flow；Graph Clustering；link prediction ； breast cancer

1 引言

在人工智能与数据科学驱动下，全球生物医学研究正经历信息规模的爆发式增长，知识的复杂性、跨学科性与载体异构性日益凸显[1]。在此背景下，从海量、多元的科研资料中准确、及时地识别具有发展潜力的研究前沿，对于制定生物医学创新战略、优化资源配置及提升科技竞争力具有关键意义。生物医学创新的核心在于科学知识在不同主体与媒介间的动态流动与重组，其中科学发现（通常以学术论文为表征）与技术转化（通常以发明专利为表征）构成了深度互构、协同演进的关系[2, 3]。因此，量化测度科学知识向技术应用的流动强度与方向，成为评估创新生态健康度与活跃度的重要依据。

当前，识别科学研究前沿的主流方法主要依赖文献引用分析[4, 5]、文本内容挖掘[6]及专家研判[7]。这些方法在探测已形成共识的研究方向上成效显著，但在应对前沿识别的实时性与语义深度两方面仍存在一定的局限。引用分析方法受限于学术发表与引用累积的周期，存在时间滞后，难以及时捕捉正在涌现的研究动向[8]。文本内容挖掘方法（如主题模型）虽能直接分析文献内容，但其模型设定（如主题数的确定、先验分布的选择）常依赖假设，且与真实文本的复杂统计结构可能存在偏差[9]，导致结果稳健性不足。专家研判方法虽能提供深层洞察，但在处理海量文献时效率较低、成本高昂，且易受个人经验与主观判断影响[10]，难以实现持续、客观的动态监测。

生物医学是典型的高度知识密集型领域，其进步深度融合了多学科知识，并最终体现为医疗技术或新药的临床应用。然而，在该领域科学知识向技术转化的过程中，固有的引用网络稀疏性与知识载体的内容异质性，为准确识别潜在的研究前沿带来了挑战。为此，本研究提出了一种融合隐性语义与异构拓扑关联的研究前沿识别框架（Content-Time-Structure Finder Pro，CTS-Finder）。

2 研究总体框架

2.1 总体思路

本研究旨在通过协同自监督图聚类与链路预测机制，识别无标注异构知识载体中科学向技术的知识溢出路径， 并提供一套可量化的研究前沿测度方法，见图 1。本研究构建的 CTS-Finder 框架首先基于论文-专利引用关系构建异构知识流动网络，并利用 PubMedBERT 与 K 最近邻（K-Nearest Neighbors，KNN） 方法构建语义关联边，形成语义增强的拓扑结构。此基础上，通过三模态门控编码器，分别提取节点的内容、时间与结构特征，并自适应融合为统一表征。上述多模态表征空间的映射过程严格受控于四元自监督联合优化策略。通过基于相对熵（Kullback-Leibler divergence，KL 散度）的聚类损失促进社区结构的发现，利用信息噪声对比损失（Information Noise Contrastive Estimation Loss，InfoNCE 损失）增强表示判别性，并辅以模块正则化防止过拟合，实现节点表示与社区结构的协同学习。最后，基于学习到的节点表征进行研究前沿量化分析预测知识流动的潜在关系概率并聚合预测流入专利的强度。最终通过计算前沿指数(Frontier Index, FI)）对社区进行排名，识别出具有高增长潜力的生物医学研究前沿领域。

10.12201/bmr.202605.00070V1

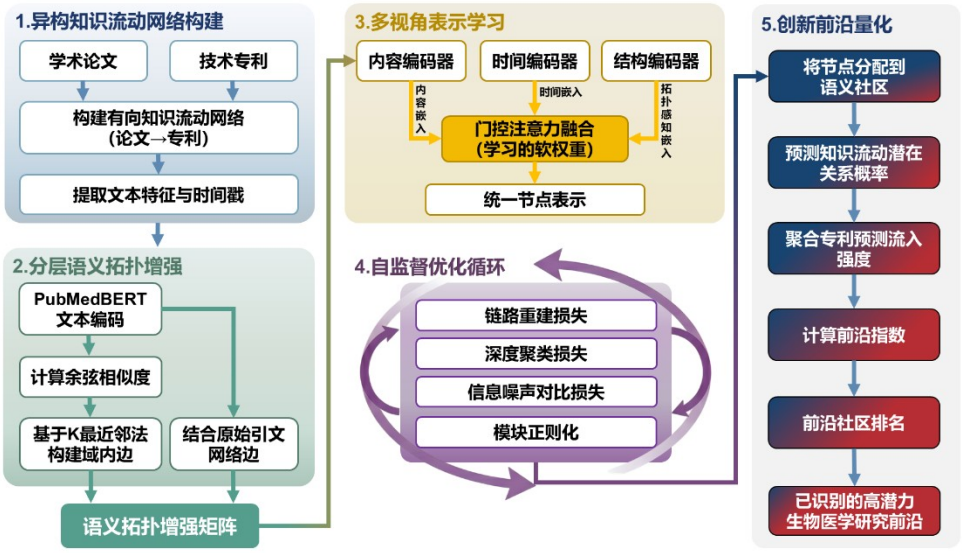

图 1　基于自监督图聚类的前沿识别方法框架

2.2 任务定义

聚焦于论文与专利引用关系的异构二部图，本文将生物医学研究前沿识别任务表征为自监督深度图聚类与潜在链路预测的联合寻优问题[11]。本文不仅关注已有知识流动关系的刻画，还进一步关注潜在“论文→专利”转化路径的预测。基于此，本文将论文节点与专利节点共同表示为异构知识流动网络中的节点集合，并将显性引用关系与后续构建的隐性语义关系共同纳入边集合。给定网络结构及节点特征后，模型的核心目标是学习一个低维鲁棒且可迁移的统一节点表征空间 Z ，使其能够同时支持知识社区划分与潜在链路预测。

2.3 分层语义拓扑增强

本研究构建了一种分层语义拓扑增强策略，通过将异构网络解构为反映引用关系的“显性转化流”和基于语义相似性的“隐性语义流”，重构出稠密且富含上下文信息的增强邻接矩阵 $\tilde{A}$ 。在该分层架构中，垂直维度由原始二部图引用矩阵 A_{cit} 构成，它保留了“基础研究 → 技术应用”的显性引用路径。与此互补，模型引入了基于语义同质性的 KNN 构图机制，以发掘潜在的隐性知识关联[12]。

利用 PubMedBERT 预训练模型对论文与专利的标题和摘要进行语义编码，并基于语义近邻关系构建域内隐性语义边。增强后的全局拓扑结构 $\tilde{A}$ 最终被形式化为显性跨域转化路径与双域隐性语义关联的线性叠加：

$$\tilde{A} = A_{cit} + A_{knn}^{(p)} + A_{knn}^{(t)} \tag{1}$$

其中， $A_{knn}^{(p)}$ 与 $A_{knn}^{(t)}$ 分别表示论文域与专利域内基于语义近邻关系构建的隐性增强子图。

2.4 三模态门控编码器

为全面表征节点的多维特征，本研究设计了包含内容（ $h_{content}^{(i)}$ ）、时间（ $h_{time}^{(i)}$ ）、结构（ $h_{struct}^{(i)}$ ）的三模态编码架构，并通过门控机制实现异构特征的动态融合。该设计主要考虑了研究前沿的形成同时受语义主题、时间演化和网络结构的共同影响，仅依赖单一模态难以揭示科学知识向技术应用扩散的过程。

在内容模态方面，本研究采用 PubMedBERT 提取领域语义特征 。 该过程为后续弥合网络结构中的拓扑断层提供语义基础[13]。在时间模态方面，模型引入了 Time2Vec 连续编码机制[14]，将节点时间戳映射为正弦周期特征，其允许模型敏锐地捕捉处于爆发期的新兴节点。基于语义增强后的拓扑网络 $\tilde{A}$ ，利用图注意力网络（Graph Attention

Network，GAT）对邻居节点进行加权聚合，从而差异化融合显性引用与隐性语义关联，量化知识流动强度。

另外，针对知识网络中节点角色与功能存在差异的问题（如部分节点侧重于基础科学创新，而另一些节点更贴近技术应用），本研究采用门控注意力融合机制（Gated Attention Fusion，GAF）将三种模态的特征进行拼接，随后通过多层感知机（Multilayer Perceptron，MLP）动态生成归一化的模态贡献权重实现特征的非线性融合[15]：

$$[\lambda_c,\lambda_t,\lambda_s]=Softmax\left(MLP_{gate}\left(h_{content}^{(i)}\oplus h_{time}^{(i)}\oplus h_{struct}^{(i)}\right)\right) \quad (2)$$

$$Z_i=\lambda_c h_{content}^{(i)}+\lambda_t h_{time}^{(i)}+\lambda_s h_{struct}^{(i)} \quad (3)$$

其中，$\oplus$ 表示向量拼接操作；MLP_{gate} 是一个多层感知机，用于将拼接后的特征映射为三个标量门控权重 $\lambda_c,\lambda_t,\lambda_s$，且满足 $\lambda_c+\lambda_t+\lambda_s=1$（通过 Softmax 归一化）。该机制使得模型能够根据节点的局部上下文动态调整语义、时序与结构信息的贡献比例，生成鲁棒且统一的节点嵌入 $Z\in R^{|V|\times d_{model}}$。

2.5 自监督联合优化策略

为了在无监督语境下实现多模态特征空间的高质量聚类与链路预测，本研究构建了一个包含图重构、分布优化、对比学习与模块度正则化四项损失约束的全局目标函数[16, 17]：

$$L=L_{clu}+\lambda_1 L_{recon}+\lambda_2 L_{con}+\lambda_3 L_{mod} \quad (4)$$

其中 $\lambda_1,\lambda_2,\lambda_3$ 为平衡各任务贡献的超参数，本研究中通过网格搜索确定其最优值。其中，图重构损失（L_{recon}）作为任务基础，引导模型在潜在表征中保留引用关系的方向性[18]。与重构任务互补，聚类分布优化（L_{clu}）遵循深度嵌入聚类算法（Deep Embedded Clustering，DEC）经典范式，通过最小化目标分布与学生氏分布（Student's t-Distribution，简称 t 分布）间的 KL 散度，促使节点平滑地向高置信度的语义中心聚拢[19]。模型同步引入了基于 InfoNCE 的异构图对比学习（L_{con}），增强节点表征的判别能力。本研究引入可微的稀疏模块度正则化（L_{mod}），进一步避免社区划分仅停留于语义层面而脱离实际知识流动结构[20, 21]。通过这种四元联合优化机制，CTS-Finder 能够在统一表征空间内实现表示学习、社区发现与潜在关系预测的协同优化。

2.6 基于预测流入的前沿识别量化规范

为克服传统文献计量学依赖历史引用累积所带来的滞后性局限，本研究提出了一种基于“预测技术流入强度”的动态量化规范。该方法旨在将模型的输出转化为可解释的前沿识别指标，通过语义社区划分与转化潜力测度的耦合，实现对生物医学研究前沿的准确定位[22]。

在社区划分阶段，模型依据深度聚类模块学习得到的软分配矩阵，将每个节点归入其后验概率最大的语义簇，从而完成从连续特征空间到离散知识社区的映射：

$$C_k=\{v_i\in V\mid K=arg\,max_j\, q_{ij}\} \quad (5)$$

其中，V 表示表示异构知识流动网络的全体节点集合，v_i 表示集合 V 中的任一节点，q_{ij} 为节点 i 属于第 j 个簇的概率，k 表示具体社区编号，K 为预设的社区总数。

为识别具有高转化潜力的“活性前沿”，本研究利用潜在表征空间中重构得到的的链路预测概率，对学术社区中论文节点未来流入技术域的总体强度进行累加。定义社区 C_k 的前沿指数（FI）为：

$$FI(C_k)=\sum_{v_p\in C_k\cap V_p}\sum_{v_t\in V_t}\widetilde{A}_{tp}\cdot I(v_t\in P_{target}) \quad (6)$$

其中，潜在引用概率 $\widetilde{A}_{tp}=\sigma(z_t^T z_p)$ 客观量化了论文节点 v_p 向技术节点 v_t 发生知识溢出的预测倾向。指示函数 $I(\cdot)$ 将计算范围严格限定于目标技术域 P_{target}（即全量专利空间）。该指数测度了科学知识向技术生态潜在溢出的总强度，从而识别那些当前引用频次不高、但具备较高技术耦合潜力的新兴研究方向。最终，依据 FI 值的排序选取 Top-N 社区即可稳健识别该领域的核心研究前沿。

3 实验设计与结果分析

3.1 数据集构建

本研究依托 Dimensions、incoPat、Web of Science 等数据库，构建了面向乳腺癌领域的数据集，经时间窗口约束与清洗后，包含 7985 项专利与 111444 篇论文。为保障领域一致性，通过关键词筛选确保所有文本均包含“breast”或“mammary”等核心术语。基于论文-专利引用关系构建单向知识流动网络，并以专利知识流入度与论文知识流出度作为连通性度量。如表 1 所示，全量图呈现高稀疏性与强冷启动特性：二部图密度仅为 9.90×10^{-6}，且零流入专利与零流出论文占比分别高达 90.01%与 94.31%。为缓解该结构对模型训练的影响，后续实验基于高活跃度节点构建稠密子图以进行有效学习。

为避免因节点规模过大导致网络结构失效，本研究基于节点活跃度选取 5000 篇高知识流出度的论文与 5000 项高知识流入度的专利，构成规模可控的关键子图。为缓解该截取可能引起的边断裂问题，引入了引用补齐机制，最终论文节点数被补齐至 11338 篇以确保引用关系的完整性(即确保关键知识转移路径在拓扑层面的完整性)，从而支持后续的知识溢出分析。网络拓扑结构统计数据的变化见表 1，验证了该方案的科学性。该实验子图基于活跃度筛选策略，在显著压缩数据规模的同时，完整保留了原始全网络中 9002 条跨域引用链路（专利引用论文的引用关系）。实验子图的二部图密度提升至 1.59×10^{-4}，专利平均知识流入度增至 1.8004，同时零知识流出的论文节点占比下降至 42.84%。

基于上述高密度实验子图，本研究采用严格按时间顺序的边级划分策略，以 8:1:1 的比例生成训练集（7202 条边）、验证集与测试集（各 900 条边）。通过严格按时间顺序划分训练集、验证集和测试集，确保模型在训练时只能学习到过去发生的引用关系。模型在测试阶段面对的是它从未见过的“未来”数据，其预测性能才能反映它对新知识的发现能力，避免了因混入未来信息而导致的虚假高分。该时间约束为基线对比确立了兼具严谨性与复现性的评估环境。

表 1 论文-专利引用全量图与实验子图拓扑结构统计数据

拓扑结构统计指标	全量图	实验子图
专利节点数	7985	5000
论文节点数	113846	11338
异构节点总数	121831	16338
有效引用关系数量	9002	9002
二部图密度	9.90×10^{-6}	1.59×10^{-4}
专利知识流入度均值	1.1274	1.8004
论文知识流出度均值	0.0791	0.7940
零知识流入专利占比	90.01%	84.04%
零知识流出论文占比	94.31%	42.84%

3.2 实验设计

3.2.1 对比基准模型

为了全面评估 CTS-Finder 在异构图表示学习与前沿识别任务中的有效性，本研究构建了涵盖 4 个维度的基准对比体系。在基础方法层面，选取 K 均值聚类算法（k-means clustering algorithm，K-Means）、图嵌入方法（Node2Vec）和异质图嵌入模型（Meta-Path-Based to Vector，Metapath2Vec），用于在高度稀疏的引文网络中建立性能基线。在图神经网络方法层面，引入具有归纳学习能力的图采样与聚合（Graph SAmple and aggreGatE，GraphSAGE）与时序图注意力网络（Temporal Graph Attention Networks，TGAT），以验证邻域采样与时间信息融合机制在动态网络数据上的适应性。在链路预测方法层面，将贝叶斯个性化排序矩阵分解（Bayesian Personalized Ranking- Matrix Factorization，BPR-MF）与轻量级图卷积（Light Graph Convolutional Network，LightGCN）纳入对比，评估基于协同过滤的推荐逻

辑对知识转化路径的预测能力。最后，深度聚类方法层面，实验选取了以 DEC 为代表的纯内容聚类模型，以及深度注意嵌入图聚类（Deep Attentional Embedded Graph Clustering，DAEGC）、结构化深度聚类网络（Structral Deep Clustering Network，SDCN）、自适应图卷积网络（Aggregate Graph Convolutional Neural Network，AGCN）等基于图结构的深度自监督聚类模型。这些模型分别体现了自编码重构与自监督对比聚类的技术路线，为本研究提出的三模态门控架构与联合优化策略提供了直接性能对照。

3.2.2 实验设置

本研究的所有实验均在 Python 3.10 与 PyTorch 2.9.1 环境下进行。经超参数调优，三模态门控编码器的隐藏层维度统一设置为 $d_{hidden}=512$，并辅以 Time2Vec 周期性映射作为时序特征编码。四元联合损失函数的权重分别校准为 $\lambda_{recon}=1.0$、$\lambda_{contrastive}=0.5$ 与 $\lambda_{mod}=0.05$；在深度聚类模块中，t 分布的自由度参数固定为 $\alpha=1.0$ 以保持聚类分布的致密性。为保证对比的公平性，所有基线模型的超参数均严格遵循其原始文献推荐的最优配置。

3.2.3 评价指标

针对前沿识别任务中潜在知识链路预测与前沿社区排序的双重目标，本研究构建了多维度量化评估体系。针对“论文-专利”潜在连接的预测任务，采用曲线下面积（Area Under Curve, AUC）评估模型全局分类性能，并使用平均精度（Average Precision, AP）衡量其在正负样本不均衡下对高置信度连接的识别能力。为评估高潜力社区的发现与排序质量，引入平均倒数排名（Mean Reciprocal Rank, MRR）、归一化折损累计增益（Normalized Discounted Cumulative Gain, NDCG）和基于超链接的主题搜索（Hyperlink-Induced Topic Search, Hits）等指标。重点关注模型将高转化潜力社区排于前列的能力，并通过精确率（Precision，P）与召回率（Recall,R）综合测度 Top-10 结果的查准与查全性能。

3.3 实验结果分析

为全面评估本研究方法在生物医学异构引文网络中的有效性，本研究将 CTS-Finder 与上述 11 种基线模型在生物医学异构引文网络上进行了对比实验，见表 2。表 2 汇总了各模型在链路预测与前沿热点识别任务上的核心指标得分。

表 2 实验模型在链路预测与前沿热点识别任务上的核心指标

模型	AP	AUC	MRR	NDCG@10	Hits@10	P@10	R@10
K-Means	0.2270	0.9030	0.3905	0.4503	0.6924	0.1363	0.3890
Node2Vec	0.0220	0.4997	0.0508	0.0446	0.0993	0.0082	0.0160
Metapath2Vec	0.0223	0.5032	0.0517	0.0451	0.0996	0.0085	0.0161
GraphSAGE	0.0557	0.7653	0.2403	0.3441	0.7173	0.0152	0.0544
TGAT	0.0856	0.7969	0.3506	0.4444	0.7630	0.0445	0.1069
LightGCN	0.0228	0.5370	0.0994	0.1104	0.2080	0.0115	0.0194
BPR-MF	0.0284	0.5934	0.1028	0.1110	0.2012	0.0137	0.0246
DEC	0.0928	0.7448	0.2025	0.2512	0.4833	0.0529	0.2190
DAEGC	0.2268	0.8783	0.3915	0.4854	0.8095	0.0665	0.2131
SDCN	0.2378	0.9054	0.3836	0.4796	0.8113	0.0682	0.2199
AGCN	0.2142	0.8905	0.3785	0.4731	0.8041	0.0689	0.2584
CTS-Finder	0.6626	0.9156	0.6843	0.7080	0.8204	0.2085	0.4161

实验结果表明，CTS-Finder 在处理冷启动节点占比达 82.7%的稀疏异构知识网络时表现出显著优势。其在潜在知识关联预测任务中取得 AP=0.6626、AUC=0.9156 的性能。相比之下，依赖显性拓扑结构的 Node2Vec、Metapath2Vec 等方法判别能力较弱（AUC≈0.5），而基于协同过滤的 LightGCN 与 BPR-MF 在 AP 指标上也表现有限。这证

实了引入分层语义拓扑增强机制的必要性：通过融合语义表征与隐性连接，模型有效缓解了冷启动问题，并在 P@10 与 R@10 指标上优于当前最优基线模型 SDCN。

在预测能力提升的基础上，模型在前沿社区的排序质量上也表现出较高的敏感度。数据表明，尽管 SDCN 与 DAEGC 在 Hits@10 维度表现较好，但在衡量排序质量的核心指标 MRR 与 NDCG@10 上，CTS-Finder 明显优于次优基线模型。这种排序优势主要得益于 Time2Vec 连续时序编码的引入。通过将静态年份戳映射为具备周期律的连续特征，模型具备了感知知识节点时间演化速率的能力，使其能够更准确地捕捉处于爆发初期的新兴节点并将其优先推荐，有效缓解了传统算法容易偏向陈旧、高频节点的偏差。

综合来看，各项指标的提升主要归功于三模态门控编码器对多源异构信息的自适应融合。例如，引入了时间注意力的 TGAT 由于缺乏语义增强，其 AP 仅为 0.0856；而纯内容的深度聚类模型 DEC 虽然具备语义判别力，但因缺乏拓扑推理能力，其 AUC 受到明显限制。CTS-Finder 通过门控机制动态调节语义、结构与时序特征的贡献比例，不仅保留了细粒度的概念分辨能力，也兼容了异构图的结构推理逻辑，最终在各项核心指标上实现了较好的平衡与鲁棒性。

3.4 乳腺癌领域研究前沿识别结果分析

本研究以乳腺癌领域为实证对象，对科学-技术知识溢出路径进行深入剖析。本节依据前沿指数（FI）的量化结果，从宏观热点趋势与微观网络结构两个维度，解析该领域知识流动的关键模式与演进规律。

3.4.1 基于 FI 的腺癌领域研究前沿分析

FI 的核心意义在于量化特定知识社区内基础研究与技术应用之间的结构耦合强度。本研究筛选出该领域内 FI 指数最高的 Top-10 研究前沿，见表 3，并引入归一化百分制 FI 指数以衡量不同前沿方向的相对热度差异。

表 3 乳腺癌领域内 FI 指数最高的 Top-10 研究前沿

排名	社区 ID	前沿主题名称	Top-5 核心关键词	FI	论文/专利数量
1	#4	三阴性乳腺癌分子分型与新辅助化疗预后标志物研究	breast cancer; cancer patients; gene expression; triple-negative breast; neoadjuvant chemotherapy	100.0	455 / 10
2	#10	靶向 CDK4/6 与新型内分泌联合治疗乳腺癌的临床疗效评估	breast cancer; metastatic breast; endocrine therapy; her2-positive breast; advanced breast	98.60	370 / 25
3	#7	基于多模态影像与人工智能模型的腋窝淋巴结转移无创预测	breast cancer; lymph node; axillary lymph; cancer patients; lymph nodes	96.85	285 / 13
4	#47	靶向 PI3K/AKT 通路的抗肿瘤小分子体外机制筛查	breast cancer; cancer cells; cell lines; cell line; cell cycle	96.19	262 / 10
5	#44	面向特定靶点的新型靶向抑制剂设计与合成	breast cancer; cancer cells; cell lines; triple-negative breast; sub sub	95.62	241 / 10
6	#19	绝经前激素受体阳性乳腺癌的卵巢功能抑制与临床管理	breast cancer; case report; metastatic breast; cancer patients; ovarian function	94.90	219 / 22
7	#46	靶向实体瘤的 CAR-NK 细胞工程与抗体偶联药物研发	breast cancer; sup sup; cancer cells; solid tumors; triple-negative breast	94.17	195 / 12
8	#49	融合超声与病理影像的乳腺良恶性肿瘤深度学习鉴别诊断	breast cancer; deep learning; neural network; breast tumor; machine learning	93.61	175 / 179
9	#2	面向体液微量生物标志物的超灵敏生物传感器研发	breast cancer; detection limit; early detection; cancer patients; cancer detection	89.39	99 / 13
10	#31	抗肿瘤组合药物、可药用盐制备及其联合治疗方案构筑	breast cancer; triple negative; negative breast; pharmaceutically acceptable; treatment breast	80.93	29 / 553

表 3 的数据分布揭示了乳腺癌领域研究前沿演进的非均匀性与阶段性特征。FI 排名领先的知识社区，如聚焦分子分型及预后生物标志物的“社区 4”，表现出极高的跨域知识流动强度，且在其内部实体构成上呈现显著差异。通过“论文/专利数量”这一指标，可进一步辨析不同前沿主题所处的产业化阶段。以“社区 4”和“社区 47”为代表，论文数量占绝对主导，呈现典型的理论爆发型，知识流动主要集中于学术网络内部，反映出早期基

础研究的活跃态势。相较之下，“社区 31”，其专利数量约为论文的 19 倍，表明该研究方向已进入以工艺开发和知识产权布局为核心的产业化成熟阶段。另外，值得注意的是“社区 49”，其论文与专利数量基本持平，体现出算法研究与影响诊断设备开发紧密互动，正处在科学与技术并行突破、相互驱动的关键发展阶段模式。

3.4.2 基于多关系流的前沿社区拓扑结构解析

为进一步在微观层面验证 FI 指数对隐性知识流动的捕获能力，研究依托 CTS-Finder 框架的多关系流表征能力，对排名前三的高潜能前沿社区——“社区 4”（三阴性乳腺癌分子分型与新辅助化疗预后标志物研究，蓝色）、“社区 10”（靶向 CDK4/6 与新型内分泌联合治疗乳腺癌的临床疗效评估，绿色）与“社区 7”（基于多模态影像与人工智能模型的腋窝淋巴结转移无创预测，橙色）进行了拓扑可视化解析，见图 2。

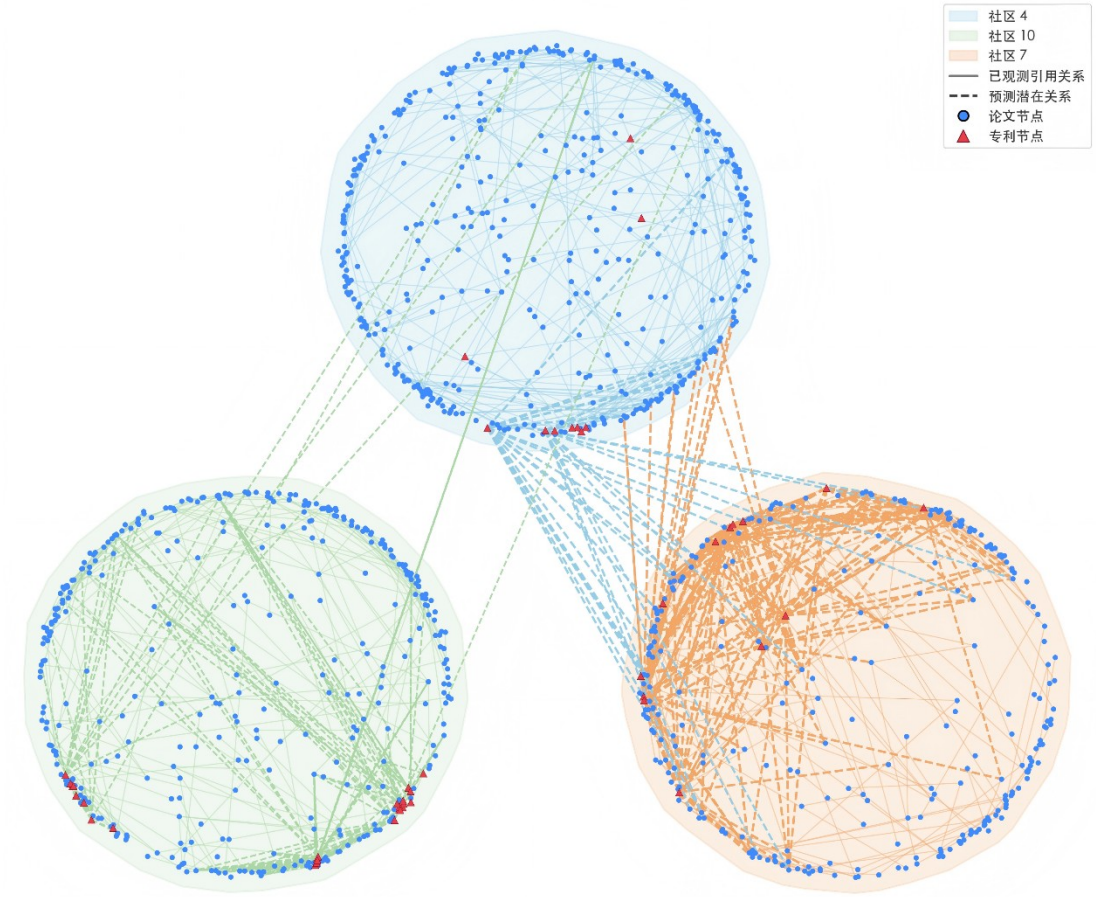

图 2 基于多关系流的高潜能前沿社区拓扑图

拓扑可视化分析显示，三大前沿社区内部连接紧密、边界清晰，呈现显著的高内聚“团块”结构。其中，位于拓扑图上方的“社区 4”表现出明显的“知识源头”特征：大量基础研究节点构成其稠密的核心（蓝色圆形），而少数技术应用节点（红色三角形）则分布相对边缘。该结构表明，该社区聚焦于三阴性乳腺癌分子分型与基因表达机制等基础研究，说明该社区仍处于基础理论密集产出的阶段。

位于拓扑图下方的“社区 10”与 “社区 7”则呈现出明显的“转化接收”特征。在这两个面向临床实际需求的前沿中，代表技术成果的专利节点（红色三角形）密集分布于社区边缘，形成了承接上游科学知识溢出的“转化接口”。该结构表明，相关研究已进入将科学发现快速固化为知识产权与解决方案的应用发展阶段，与前文所述“产业成熟型”社区的界定形成了逻辑呼应。

尤为重要的是，模型所揭示的知识流动模式超越了稀疏的显性引用关系。图中虚线所示的预测潜在关系揭示出显著的隐性知识流动方向：从“社区 4”向“社区 10”存在持续的语义辐射，“社区 4”与“社区 7”相互的知识流动交互。这种跨社区的强关联揭示出，表观遗传学等层面的分子标志物发现，正在深层次驱动临床靶向用药方案的优化与影像诊断模型的特征构建。

该隐性知识协同机制的捕捉与近期领域进展相吻合，印证了本框架识别隐性知识关联的有效性。近期研究显示，三阴性乳腺癌分型已从 mRNA 表达谱扩展至染色质可及性、单细胞转录组等多维度探索[23]，相关预后标志物不断被阐明[24]。这对应了本研究中社区 4 所呈现的“理论爆发”特征。同时，针对耐药性的靶向治疗策略（如 CDK4/6 抑制剂、新型抗体药物偶联物等）已显著改善临床疗效[25]，其中 CDK4/6 抑制剂联合疗法及 AI 辅助诊断工具逐步成熟与应用[26]标志着相关研究已进入临床转化阶段，与社区 10、7 的“转化接收”模式相符。尤为重要的是，表观遗传信息正被用于构建“影像-基因组学”多模态模型以指导精准治疗[27]，这一前沿方向印证了本研究所揭示的跨域知识流动路径。

4 结语

本研究从“科学-技术”知识流动的视角出发，构建了一个融合隐性语义与异构拓扑的量化分析框架。不同于传统基于共被引或关键词共现的静态同质网络分析，本研究通过引入领域预训练语言模型与自监督图学习，将稀疏的显性引用关系与稠密的隐性语义关联相融合，从而构建了一个能反映知识深层演化逻辑的“科学-技术”知识流动网络。另值得关注的是，当前研究框架虽引入时间编码，但本质是对连续动态过程的离散化采样。未来的工作可致力于构建连续时间动态图神经网络模型，以刻画知识流动随时间的产生、消亡、爆发与迁移等连续演化过程。尽管存在上述局限，本框架为生物医学研究前沿的识别与监测提供了一种高效、可量化且具有前瞻性的新策略，有助于进一步推动科技创新向数据驱动与知识引导的智能化决策支持转变。

作者贡献：李永洁负责研究设计、实验结果分析、论文撰写；王龙超、孙轶楠负责完善研究方案、数据集构建、数据实验；唐小利负责完善研究方案、论文修订。
利益声明：所有作者均声明不存在利益冲突。

参考文献

[1] JIA Y, CHEN H, LIU J, et al. Exploring network dynamics in scientific innovation: collaboration, knowledge combination, and innovative performance[J]. Frontiers in Physics, 2025, 12:1492731.

[2] JIANG D, LI N, WANG K, et al. Research hotspots and frontier trends in the field of 3D printing in medical education from 2010 to 2025: a bibliometric analysis[J]. 3D Print Med, 2025,11(1): 54.

[3] LI Y, ZHU J, WANG L, et al. Identifying Innovation Frontiers Based on Prediction of Citation Network Links Between Papers and Patents[C]//. 17th International Joint Conference on Knowledge Discovery,Knowledge Engineering and Knowledge Management, Marbella, Spain, 2025: 398-405.

[4] 罗瑞, 许海云, 董坤. 领域前沿识别方法综述[J]. 图书情报工作, 2018,62(23): 119-131.

[5] FAJARDO-ORTIZ D, LOPEZ-CERVANTES M, DURAN L, et al. The emergence and evolution of the research fronts in HIV/AIDS research[J]. PLoS One, 2017,12(5): e178293.

[6] 王灿灿. 基于 NSF 基金文本的前沿主题探测及分析研究[D]. 华中师范大学, 2023.

[7] FUNK P, DAVIS A, VAISHNAV P, et al. Individual inconsistency and aggregate rationality: Overcoming inconsistencies in expert judgment at the technical frontier[J]. Technological Forecasting and Social Change, 2020,155: 119984.

[8] CHOI J, YOON J. Measuring knowledge exploration distance at the patent level: Application of network embedding and citation analysis[J]. Journal of Informetrics, 2022,16(2): 101286.

[9] EGGER R, YU J. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts[J]. Frontiers in Sociology, 2022, 7:886498.

[10] 范旭辉, 穆智蕊. 融合 BERTopic 和大语言模型的研究前沿识别——以美国 NSF 人工智能领域资助为例[J]. 情报工程, 2025,11(01): 18-28.

[11] FAN J, YANG J, GU Z, et al. Path-aware multi-scale learning for heterogeneous graph neural network[J]. Neural Networks, 2025,191: 107743.

[12] GUO Z, WANG F, YAO K, et al. Multi-Scale Variational Graph AutoEncoder for Link Prediction[C]//. Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, State of Arizona, USA, 2022: 334-342.

[13] GU Y, TINN R, CHENG H, et al. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing[J]. ACM Trans. Comput. Healthcare, 2021,3(1): 2.

[14] BUI N K H, CHIEN N D, KOV'ACS P, et al. Transformer Encoder and Multi—Features Time2Vec for Financial Prediction [C]//. 2025 33rd European Signal Processing Conference (EUSIPCO), Palermo, Italy, 2025: 1682-1686.

[15] XIONG L, YUAN X, HU Z, et al. Gated Fusion Adaptive Graph Neural Network for

Urban Road Traffic Flow Prediction[J]. Neural Processing Letters, 2024,56(1): 9.

[16] 孙艳丰, 杜鹏飞. 基于图关系选择的深度聚类网络[J]. 北京工业大学学报, 2024,50(12): 1428-1436.

[17] 文茜琳, 杨旭华, 马钢峰. 基于邻域增强的图自监督推荐[J]. 计算机学报, 2025,48(10): 2278-2297.

[18] LIN M, WANG T, ZHU Y, et al. A Heterogeneous Directed Graph Attention Network for inductive text classification using multilevel semantic embeddings[J]. Knowledge-Based Systems, 2024,295: 111797.

[19] 朱喜珍, 张齐齐, 赵中英. 生成式图自监督学习综述[J]. 集成技术, 2025,14(04): 71-86.

[20] ZHENG Y, JIA C, YU J, et al. Deep embedded clustering with distribution consistency preservation for attributed networks[J]. Pattern Recognition, 2023,139: 109469.

[21] 陈洁, 刘斌斌, 赵姝, 等. 基于模块度感知图自编码器的重叠社区发现模型[J]. 清华大学学报(自然科学版), 2024,64(08): 1319-1329.

[22] 奉国和, 陈丽霞, 邓伟伟, 等. 基于重叠社区的新兴技术识别[J]. 图书馆论坛, 2024,44(09): 48-59.

[23] CHEN Z, LIU Y, LYU M, et al. Classifications of triple-negative breast cancer: insights and current therapeutic approaches[J]. Cell & Bioscience, 2025,15(1): 13.

[24] WANG X, LI X, DONG T, et al. Global biomarker trends in triple-negative breast cancer research: a bibliometric analysis[J]. Int J Surg, 2024,110(12): 7962-7983.

[25] ROUSSEL-SIMONIN C, FERNANDEZ-MARTINEZ A, POSTEL-VINAY S, et al. Exploring KAT6 as a therapeutic target in breast cancer: epigenetic approaches for precision medicine[J]. NPJ Breast Cancer, 2025,11(1): 146.

[26] RETAMERO J A, GULTURK E, BOZKURT A, et al. Artificial Intelligence Helps Pathologists Increase Diagnostic Accuracy and Efficiency in the Detection of Breast Cancer Lymph Node Metastases[J]. Am J Surg Pathol, 2024,48(7): 846-854.

[27] CHANG L, LIU J, ZHU J, et al. Advancing precision medicine: the transformative role of artificial intelligence in immunogenomics, radiomics, and pathomics for biomarker discovery and immunotherapy optimization[J]. Cancer Biol Med, 2025,22(1): 33-47.

10.12201/bmr.202605.00070V1