

基于生成式 AI 的药物重定位研究

龚后武, 金敏

湖南大学信息科学与工程学院, 湖南 长沙 410082

摘要

针对当前药物重定位研究药物适应症预测数量固定、无法全面揭示药物潜在适应症的问题, 提出了生成式 AI 的药物重定位模型 GenDrugShifter。该模型由图注意力神经网络和 Transformer Decoder 模块组成, 能够进行端到端的药物重定位。该模型以 InChI 格式表示的药物分子结构为图注意力神经网络的输入, 使用自监督方法学习药物活性分子结构和药物适应症之间的潜在联系, 通过自回归的方法输出药物的适应症。西药重定位实验结果表明, GenDrugShifter 在预测性能上优于其他 4 种先进药物重定位方法。GenDrugShifter 能够更全面地揭示药物潜在的适应症, 具有优越性和可靠性。临床数据进一步证明了其在实际应用中的有效性。

关键词

生成式 AI; 药物重定位; 大语言模型; 人工智能; 药物发现

中图分类号: TP18, R9

文献标志码: A

doi: 10.11959/j.issn.2096-0271.2025030

Research on drug repositioning based on generative AI

GONG Houwu, JIN Min

College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China

Abstract

In view of the problem of fixed drug indication prediction quantity and inability to fully reveal potential indication of drug in current drug repositioning research, this article proposes a generative AI drug repositioning model GenDrugShifter, which consists of a graph attention neural network and a transformer decoder module. The model uses the drug molecular structure represented in the InChI format as input, learns the potential relationship between the drug active molecular structure and indication through self-supervision, and outputs the indication of the drug using autoregressive methods. The results of the Western medicine repositioning experiment show that GenDrugShifter is superior to the other four advanced drug repositioning methods in terms of predictive performance. The GenDrugShifter model can more comprehensively reveal the potential indication of drug, with superior performance and reliability. The validation of clinical data further proves its effectiveness in practical applications.

Key words

generative AI, drug repositioning, large language model, artificial intelligence, drug discovery

0 引言

药物重定位研究作为近年来药物研发领域的一个热点，正逐渐受到广泛关注^[1]。药物研发过程漫长且成本高昂，而药物重定位则避免了重新研发新药物的复杂过程，通过挖掘已有药物的潜在适应症，为药物研发开辟了新的路径^[2]。当前，药物重定位研究已经取得了一定的进展^[3]。通过大数据分析、生物信息学等方法，研究人员能够更深入地理解药物与疾病之间的复杂关系，进而发现药物新的适应症，但药物重定位研究仍面临诸多挑战^[4]。一方面，数据的稀疏性和噪声问题使药物与疾病之间的关联难以被准确识别^[5]；另一方面，现有的预测模型往往缺乏足够的泛化能力，难以应对复杂多变的生物系统^[6]。

生成式人工智能（AI）主要通过文本与数据挖掘，对输入的训练素材进行深度学习，形成“创作”逻辑，最终根据指令输出特定内容^[7]。ChatGPT是生成式人工智能技术浪潮的一部分，ChatGPT依赖于GPT（generative pre-trained transformer），GPT是深度学习模型，通过互联网上可用的数据进行训练，专门用于生成文本。GPT名字中之所以有Transformer，是因为GPT是OpenAI（OpenAI处于科技行业下一件大事件的最前沿）在谷歌的Transformer语言模型架构的基础上构建的。Transformer由Google于2017年提出^[7]，是一种基于自注意力机制的神经网络模型架构，用于处理从序列到序列的建模任务，在长距离依赖和并行技术方面具有显著优势，是目前在自然语言处理领域较先进的建模架构。

随着机器学习和深度学习技术的快速

发展，学者们尝试使用机器学习和深度学习技术来减少实验损失、降低风险。但是目前大部分药物重定位的研究直接采用多标签学习的分类模型，其面临的最大挑战是药物适应症的标签空间巨大（种类超过13 000种^[8-16]）。针对这一挑战，已有的典型解决方案为二元相关性方法，其将多标签问题转化为多个二分类问题，针对不同标签分别建模^[17-23]。这种方法忽视了标签之间可能存在的关联关系，如标签依赖性、共现和相关性；不良反应呕吐和腹泻存在一定的相关性，但多标签分类方法针对二者分别单独建模，没有捕捉和使用二者之间的关联关系^[24]；且多标签分类方法只能预测出某种药物是否具有已知标签中的某种药物适应症，导致其最多能预测出的药物适应症数量等于已知标签中的药物适应症数量，无法预测未知的药物适应症，预测范围受限^[25]。因此，本文提出了生成式AI的药物重定位模型GenDrugShifter，使用图注意力神经网络提取分子特征，采用Transformer Decoder模块以自回归方式生成药物适应症，使预测出的药物适应症的数量和范围不受限制。

1 相关工作

药物和疾病具有多种生物学属性特征，如药物分子化学结构、副作用、靶标等。不同的生物学特征在药物适应症预测中可能起到不同的作用，因此如何选取合适的生物学特征是当前药物重定位研究面临的挑战之一^[26]。药物重定位研究利用的数据特征主要包括药物的分子结构、靶标蛋白、药物副作用、药物-疾病关联等。这些数据特征可以从多个角度为药物重定位研究提供有用的信息和线索^[27]。其中，药物的

分子结构特征预测性强，具有相似化学结构的药物往往具有相似的药理作用和生物活性，因此可以通过分析药物的化学结构来预测其可能的新适应症。靶标蛋白的三维结构和活性位点等信息往往难以被获取，需要借助高级实验技术和计算方法。药物副作用的评估往往受到患者个体差异、用药剂量等因素的影响，具有一定的主观性。有些药物的副作用可能在临床试验中未能被完全发现，导致标签数据不完整^[28]。药物与疾病的关联数据往往包含大量的噪声和干扰信息，需要进行数据处理和分析。因此，药物化学结构特征在药物重定位研究中具有独特的优势。其预测性和基础性使药物化学结构特征成为药物重定位研究中的重要工具。同时，通过了解药物的化学结构，研究人员可以对药物进行结构改造和优化，为新药物的设计提供有力的支持。因此，本文将充分利用药物化学结构特征的优势开展药物重定位研究。

早期的药物重定位方法依赖于物理规则、专家知识，计算代价高并且准确率^[29]。目前基于计算的药物重定位方法可以粗略地分为4类：经典的机器学习模型、基于矩阵的方法、基于网络的方法和深度学习方法系^[30]。在药物重定位研究中，学者们通常将药物重定位模型视为多标签学习的分类模型来处理^[31]。传统的分类模型，如逻辑回归、支持向量机和随机森林等，虽然在一定程度上能够预测药物对特定疾病的潜在治疗作用，但其预测能力受限于数据的稀疏性和复杂性。为了克服这些挑战，学者们不断探索新的算法和模型，以提高预测的准确性和泛化能力^[32]。Cao等^[33]将药物-靶标相互作用预测转换为二元分类问题，利用生物和药物-靶标相互作用网络的信息训练随机森林模型。近年来，基于矩阵分解的方法，如奇异值分解和非

负矩阵分解，在药物-靶点相互作用预测中展现出了良好的性能。基于矩阵分解的方法通过分解药物和靶点的特征矩阵，揭示它们之间的潜在关联，为药物重定位提供了新的视角。Meng等^[34]提出了相似性约束概率矩阵分解方法，通过识别新的药物-病毒相互作用实现药物重定位。同时，基于复杂网络分析的方法，如蛋白质-蛋白质相互作用网络和药物-疾病网络，通过挖掘网络中的拓扑结构和路径信息，进一步丰富了药物重定位的方法。深度学习方法，特别是卷积神经网络、循环神经网络和图神经网络等，在药物重定位研究中更是被广泛应用。它们能够自动提取药物分子结构、靶点蛋白序列以及药物-靶点相互作用等复杂特征，实现高精度的预测和分类。Zhao等^[35]将深度神经网络与多种传统机器学习方法在药物重定位研究中进行了系统性对比，研究表明，深度神经网络优于支持向量机和随机森林等传统的机器学习算法。而多标签分类算法往往需要大量的标签用于模型的训练，且标签的获取成本往往较高，限制了模型预测范围。面对未知不良反应预测和获取标签成本高的挑战，本研究引入生成式AI，其成为推动药物重定位研究领域发展的新思路。

2 GenDrugShifter

GenDrugShifter由图注意力神经网络和Transformer Decoder模块组成，是一个端到端的药物重定位模型，如图1所示。其以InChI格式表示的药物化学结构为图注意力神经网络的输入，使用自监督方法学习药物活性分子结构和药物适应症之间的潜在联系，通过自回归的方法输出药物的适应症，对药物适应症的数量没有限制。

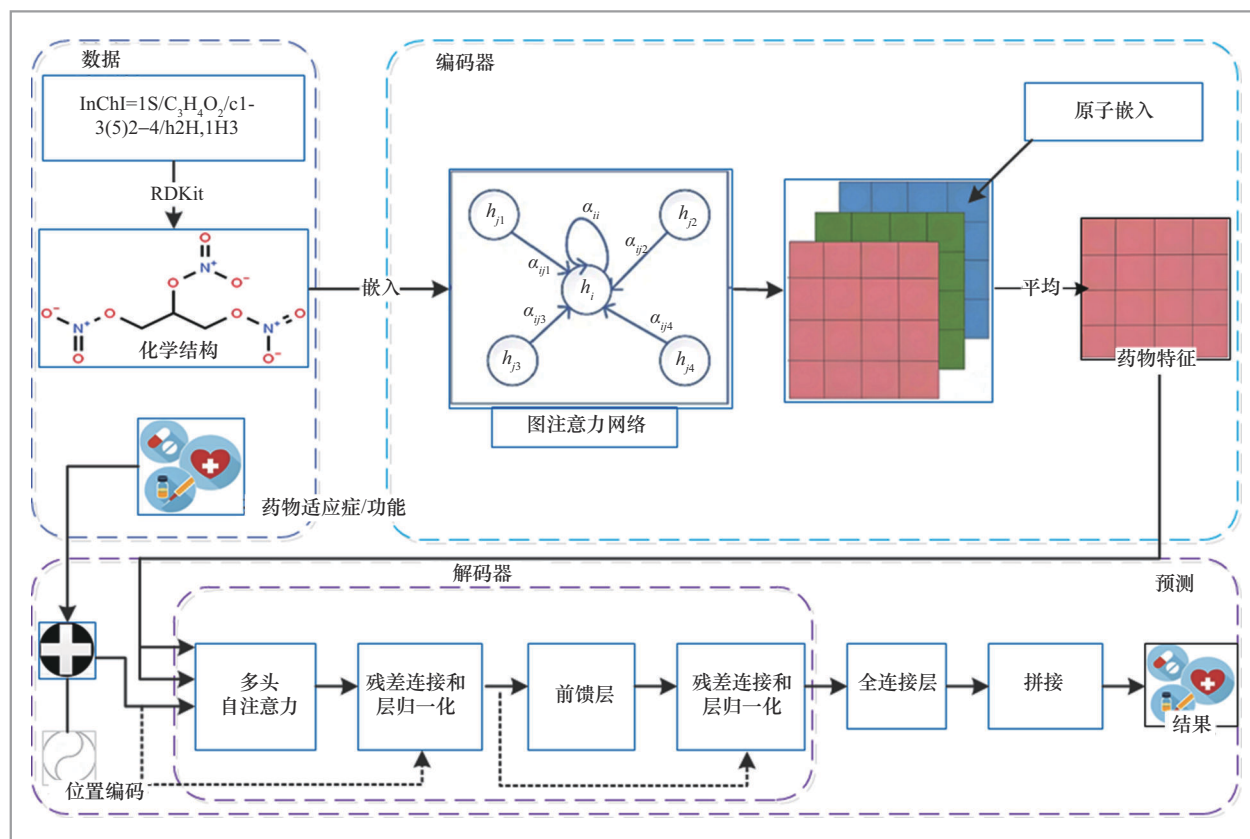


图1 GenDrugShifter模型

以InChI格式表示的药物活性分子结构在经过RDKit化学信息处理库转换为图数据结构后输入图注意力神经网络，图注意力神经网络将分子结构中的原子作为节点，将原子之间的化学键作为边，使用多个独立的注意力机制分开计算聚合原子特征，学习不同特征空间中的特征以增强其表达能力，将不同特征空间中的特征连接起来并取平均值作为最终的特征。在获得药物化学结构特征后，GenDrugShifter将其和药物适应症一起输入Transformer Decoder模块，使其可以通过自回归的方法不断地输出药物的适应症，从而实现预测药物潜在适应症的目的。本文最终实现了一个端到端的、根据药物活性分子结构预测药物潜在适应症的药物重定位生成模型。

2.1 图注意力网络

药物化学结构天生具有图的数据结构特征，因此本文使用能有效处理非欧氏空间数据特征的图神经网络来处理提取药物分子结构特征。图注意力网络（graph attention network, GAT）^[36]由图神经网络发展而来，采用注意力机制来计算节点间的相似性、聚合节点特征并放大关键节点的影响，能更好地处理复杂的图结构数据和捕获节点之间的不均匀关系。使用GAT处理药物化学结构时，将药物分子中的原子作为节点、原子之间的化学键作为边，可以在更新每个原子的特征时都考虑到其邻居原子的特征。在更新原子特征的过程中，GAT会计算该原子与所有邻居原

子之间的注意力权重, 并使用计算出来的注意力权重和对应的邻居原子特征来更新自身的特征, 使GAT可以很好地捕捉药物分子中原子间的关系。GAT注意力权重计算过程如图2(a)所示。

药物化学结构中原子特征的更新过程可用式(1)、式(2)、式(3)解释, 对于每个原子 h_i , GAT首先使用一个共享矩阵 W 对其特征做线性变换; 然后计算其与邻居原子 h_j 之间的注意力分数 e_{ij} , 计算出的注意力分数用于指示不同邻居原子 h_j 对当前原子 h_i 的重要性; 接着使用Softmax函数对注意力分数进行归一化, 获取注意力权重 α_{ij} , 使用该权重来计算聚合当前原子的所有邻居原子的特征, 并通过非线性变换函数 σ , 将聚合后的特征作为当前原子的新的特征 h'_i , 完成节点状态的更新。

$$e_{ij} = a(W\vec{h}_i, W\vec{h}_j) \quad (1)$$

$$\alpha_{ij} = \text{softmax}_j(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in N_i} \exp(e_{ik})} \quad (2)$$

$$\vec{h}'_i = \sigma \left(\sum_{j \in N_i} \alpha_{ij} W \vec{h}_j \right) \quad (3)$$

其中, a 表示计算两个节点相关度的函数, N 表示图中节点数。为了缩短图注意力神经网络的训练过程和提高模型的表达能力, 本文还使用了GAT的多头注意力机制, 如图2(b)所示。其中concat/avg表示将所有的注意力系数矩阵进行拼接后求平均值。GAT的多头注意力机制使用式(4)更新原子特征。

$$\vec{h}'_i = \sigma \left(\frac{1}{K} \sum_{k=1}^K \sum_{j \in N_i} \alpha^k_{ij} W^k \vec{h}_j \right) \quad (4)$$

其中, K 表示头数。

2.2 Transformer Decoder模块

在提取到药物分子的结构特征后, Transformer Decoder模块依据药物化学结构特征预测药物适应症。此过程通过Transformer Decoder模块的生成能力, 以自回归的方式不断地生成药物适应症。Transformer Decoder模块的结构如图3所示。

Transformer Decoder将每种药物对应的所有的适应症作为一个句子, 将每个

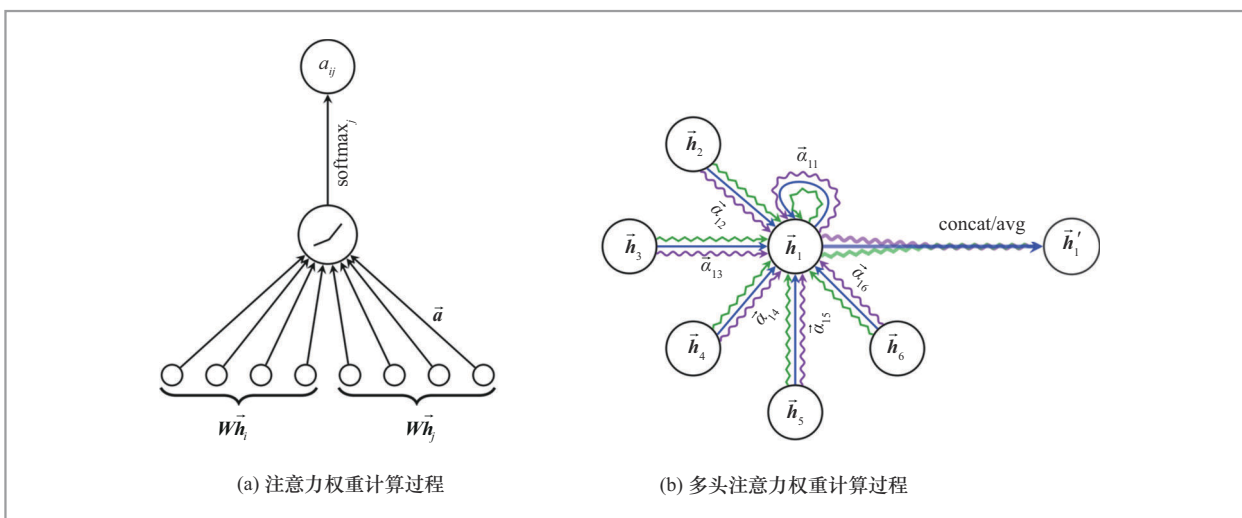


图2 GAT注意力权重计算过程

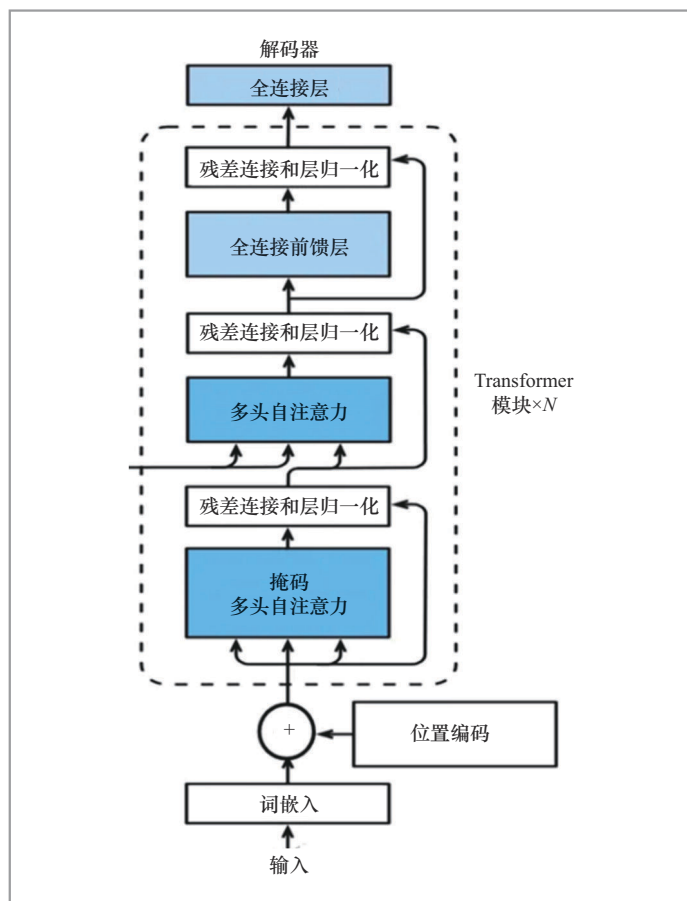


图3 Transformer Decoder模块结构

适应症作为句子中的一个单词。Transformer Decoder 模块通过自然语言处理中的嵌入技术将单词转化为数值向量之后，添加位置编码使 Transformer Decoder 模块可以感知到适应症之间的相对位置。Transformer Decoder 模块由多个 Decoder Layer 堆叠而成，每个 Decoder Layer 都包含两个多头注意力机制的计算过程。第一个多头注意力模块主要计算聚合药物适应症自身的特征，以增强表达能力；第二个多头注意力模块可以建立药物分子结构特征和药物适应症之间的联系，为适应症生成做准备。Transformer Decoder 模块基于单个注意力的计算过程，计算多个互相独立的注意力分数，并将计

算结果连接起来，使用一个前馈神经网络做非线性变换。单个注意力用特征向量 Q 、 K 、 V ，根据式 (5) 计算得出。

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

在第二个多头注意力模块中，Transformer Decoder 模块将图注意力神经网络提取出来的药物化学特征作为注意力计算过程中的 V 和 K 、药物适应症作为 Q ，通过多头注意力机制的聚合，建立起药物化学特征和药物适应症之间的潜在联系。

2.3 模型评估指标

本文在模型训练结束后，测试集中的数据将被用来验证模型的性能。为了量化分析和比较模型性能，本文选用了准确率、精确率、召回率以及 F1 分数 4 个指标来评估模型的表现。评估指标计算方法如下。

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (6)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (7)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (8)$$

$$\text{F1-score} = \frac{2TP}{2TP + FP + FN} \quad (9)$$

其中，TP 表示预测结果为存在关联的已知药物适应症，FP 表示预测结果为存在关联的未知药物适应症，FN 表示预测结果为不存在关联的已知药物适应症，TN 表示预测结果为不存在关联的未知药物适应症。

3 实验

3.1 西药重定位实验

(1) 数据集的构建

本文所用数据来自之前的相关研究中 Luo 的数据集^[34-37], 包括 593 种药物、313 种疾病、1 933 对药物-疾病关联关系。其中, 药物信息来自 DrugBank, 疾病信息来自 Online Mendelian Inheritance in Man (OMIM)。

补充数据源自 aBiofilm 库, 包括 1 720 种药物、72 种疾病、436 对药物-疾病关联关系。然后, 本文结合 PubChem 数据库补充药物的 InChI 特征。

(2) 参数设置

训练过程包括将训练集数据的输入深度学习模型, 使用反向传播算法结合优化器调整模型中的变量值优化损失函数, 使损失函数的值降到最低。GenDrugShifter 使用

Adam 作为优化器, 在训练过程中的每一步都会自动调整学习率的值。为了防止过拟合, GenDrugShifter 采用了 L2 正则化和神经元随机失活的技术^[38]。

同时 GenDrugShifter 采用了预热学习率的技术来稳定模型训练过程、加速训练、提高模型性能。在每个训练轮次结束后, 本实验使用验证集数据来测试模型的性能, 并在所有训练轮次结束后使用测试集数据获得模型在该折交叉验证数据划分下的性能度量数据。待全部训练完成后, 本实验将每一折获得的性能度量数据取平均值作为模型最终的性能度量数据。

GenDrugShifter 超参数设置见表 1, 调参过程如图 4 所示。

(3) 交叉验证

首先实验按照 9: 1 的划分比例, 将数据集随机划分为训练-验证集和测试集, 其中测试集中的数据量占总数据量的 10%。随后在训练-验证集上采用 5 折交叉验证的方法划分数据集, 本文将每一折交叉验证下在测试集上的性能度量数据取平

表1 超参数设置

名称	值	备注
gat_heads	2	GAT的头数
dropout	0.4	随机失活神经元比例
lr_min	1.5×10^{-5}	自调节学习率的最小值
lr_max	0.005	自调节学习率的最大值
epochs	50	训练轮次
warm_epochs	10	学习率预热轮数(在该轮数内,学习率由lr_min线性上升至lr_max)
batch_size	32	批次大小
nlayeres	2	Transformer Decoder Layer 数量
d_model	256	词向量维度
max_len	50	序列填充。药物分子中固定的原子数量和适应症中固定的适应症数量。当原子数量和适应症数量小于该值时会添加padding
K-fold	5	K折交叉验证, K=5

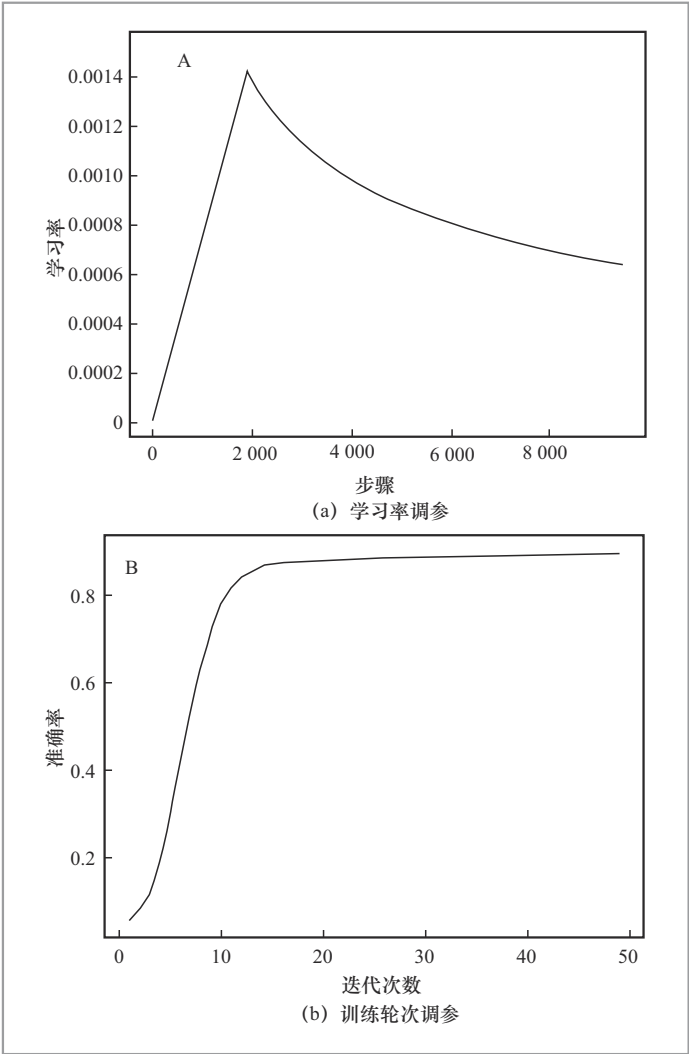


图4 GenDrugShifter调参过程

均值作为最终的性能度量数据。
为了直观地评估 GenDrugShifter 的性

能优势和局限性，本文将其与先前相关研究中采用的方法进行了详细比较。**表2**列出了 GenDrugShifter 以及选取的若干具有代表性的先进研究方法的关键性能指标，包括准确率、精确率、召回率、F1 分数，以全面衡量各个方法在药物重定位预测任务中的表现。

选取的具有代表性的先进药物重定位研究方法如下。

- deepDR：基于网络的深度学习方法^[39]利用多模态深度自动编码器从异构网络中学习药物的特性，使用变分自动编码器对所学的药物低维表征和临床报告中的药物-疾病对进行编码与解码操作，从而推断出药物潜在的适应症。
- HeTDR：基于异构网络和文本挖掘开发了一种新的药物重定位模型 HeTDR^[40]，该方法结合多网络的药物特征和生物学语料库中获取的疾病特征，揭示已批准药物和非适应症疾病之间的潜在关联。
- LAGCN：基于层注意图卷积网络开发了用于药物-疾病关联预测模型^[41]。该方法整合异构网络，使用注意机制将来自多个图卷积层的表示相结合，基于整合表示对未观察到的药物-疾病关联进行评分。
- DRHGCN：该方法利用异构网络拓扑信息，通过对网络进行图卷积操作，该方法设计了域间和域内特征提取模块来学习药物和疾病的嵌入，引入了层注意机制

表2 GenDrugShifter与其他模型的比较

模型	准确率	精确率	召回率	F1 分数
deepDR	0.9435	0.9230	0.7000	0.9080
HeTDR	0.9590	—	—	0.9010
LAGCN	0.9342	0.9491	0.3577	0.2918
DRHGCN	0.9450	—	—	0.8110
GenDrugShifter	0.9658	0.9505	0.9327	0.9410

组合来自多个图卷积层的嵌入，以提高模型预测性能^[42]。

（4）消融实验

在本研究的第一个消融实验 GenDrugShifter-1 中，本文修改了 GenDrugShifter 架构，将原始 GenDrugShifter 中的 Transformer Decoder 模块替换为双向长短期记忆（bidirectional long short-term memory, BiLSTM）模块，旨在深入探究不同序列生成结构对模型整体性能的影响。其核心目标是通过比较经过结构调整后的模型与 GenDrugShifter 的表现，评估 Transformer Decoder 和 BiLSTM 在生成式药物重定位研究中的相对优劣，以揭示两种序列生成机制在处理复杂医学数据和生成药物适应症的特性和局限性。

在本研究的第二个消融实验（GenDrugShifter-2）中，本文采用了一种结合图注意力神经网络与 Transformer Encoder 模块的方法来提取药物化学结构特征，旨在探索这两种先进深度学习技术的协同效应在捕获和提炼复杂药物化学结构信息中的潜力。本文首先利用图注意力神经网络对分子的图结构建模，该网络通

过其独特的注意力机制，根据节点的重要性动态地调整邻域信息的权重，从而有效地提取出分子的局部和全局特征。接着，本文将 GAT 提取的图特征输入 Transformer Encoder 模块中，该模块凭借其自注意力机制能够进一步捕捉分子特征之间的长距离依赖关系和上下文信息。通过该消融实验，本文验证了这种组合方法能够提高药物重定位模型的准确性和泛化能力。消融实验结果如图 5 所示。

由图 5 可知，GenDrugShifter 的准确率分别比 GenDrugShifter-1、GenDrugShifter-2 提高了 2.68% 和 6.38%。实验结果表明，GenDrugShifter 的精确率、召回率和 F1 值优于 GenDrugShifter-1、GenDrugShifter-2。这说明在药物重定位方面，Transformer Decoder 模块要优于 BiLSTM 模块（GenDrugShifter-1、GenDrugShifter-2）。使用 GAT 提取的特征表现最佳，可能是因为药物分子结构天然具有图结构数据特征，所以专门用来处理图非欧氏空间的图注意力神经网络能够较好地处理并提取药物分子结构的特征。使用 Transformer Decoder 结构作为生成

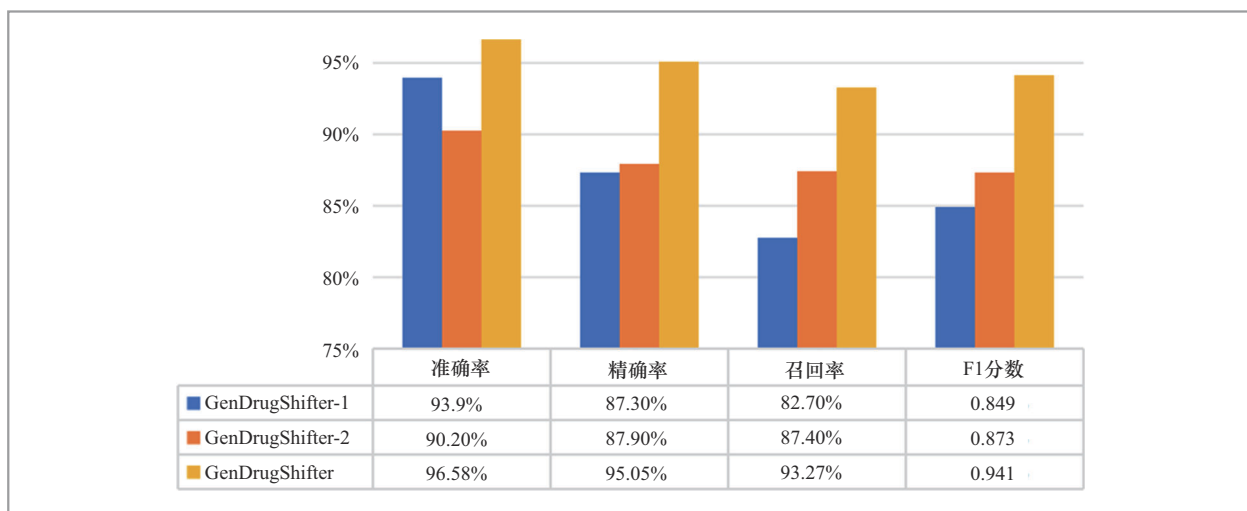


图5 GenDrugShifter消融实验结果

模块的性能要优于 BiLSTM 模块，可能是 Transformer Decoder 模块使用的注意力机制可以放大药物化学结构特征中关键部分的影响，而 BiLSTM 模块则只能同等看待药物化学结构特征中的所有子特征，无法赋予关键子特征更高的权重。

(5) 案例研究与药学意义

为了切实地评估 GenDrugShifter 的实际应用能力，本文随机选取了阿莫西林、阿司匹林、地西洋共 3 种药物，运用

GenDrugShifter 对这些药物的适应症预测。通过对预测结果进行分析，笔者发现 GenDrugShifter 成功地预测出一些药物的已有适应症。为进一步验证预测结果的准确性，笔者进一步查阅了 DrugBank、DrugCentral、Pubmed 等数据库中的文献证据，经确认，得到 GenDrugShifter 预测的药物适应症（部分），见表 3。

本文利用 GenDrugShifter 进行癌症、新型冠状病毒（COVID-19）感染（简称新冠肺炎）等疾病的药物重定位研究。由表 4 可知，治疗乳腺癌的候选药物中，氯法拉滨是一种治疗白血病的药物，醋酸甲地孕酮是一种治疗乳腺癌的药物，阿比特龙是一种治疗前列腺癌的药物。该预测结果可以在先前临床文献中得到证实^[29]。治疗新冠肺炎的候选药物中，利巴韦林在 2019 年 COVID-19 诊断和治疗计划中被推荐为临床实践药物，阿比多尔进入了临床试验阶段。福尔可定没有得到文献证实，关于预测药物的实际临床效果需要设计湿实验进一步验证。

表3 GenDrugShifter预测的药物适应症(部分)

药物	预测的适应症	证明内容
阿莫西林	胆汁淤积	PMID: 32198411
	动脉粥样硬化	PMID: 31419266
	肺炎	DrugBank
阿司匹林	腹泻	DrugBank
	肺炎	DrugBank
	膀胱瘤	PMID: 37773843
地西洋	胆汁淤积	DrugBank
	肝肿瘤	DrugBank
	食管鳞状细胞癌	DrugCentral

注:PMID指PubMed唯一标识码,PubMed Unique Identifier。

表4 疾病潜在药物发现(部分)

疾病	药物	证明内容
乳腺癌	氯法拉滨	PMID: 32908731
	雌莫司汀	PMID: 33535297
	醋酸甲地孕酮	PMID: 37924026
	柔红霉素	PMID: 38611925
	阿比特龙	PMID: 38160476
COVID-19	利巴韦林	PMID: 38652363
	曲马多	PMID: 33385878
	氯喹	PMID: 32251767
	阿比多尔	PMID: 38314855
	福尔可定	未验证

3.2 中药重定位模型“时珍”实验

当前，已有许多的中药重定位研究集中在利用现有工具和手工操作进行统计分析，但其与通过 AI 大规模挖掘中药与疾病（证候，表示中医的诊断结论）之间的关联联系还存在一定差距^[43]。对此，本文进行了中药重定位模型“时珍”实验。

(1) 数据集的构建与特征处理

数据集来源于 SymMap（symptom mapping）、Herb 两个中医药数据库。其中，SymMap 是一个中医药证候关联数据库^[44]，于 2018 年 8 月发表于 *Nucleic Acids Research* 杂志。该数据库提供了大量关于中药、化合物、中医症状和疾病等

的信息，SymMap 通过内部分子机制和外部症状映射用于药物筛选。SymMap 研究收录了 2015 版中国药典包含的 499 种中药，含 19 595 种化合物以及 1 717 种中医证候，集成了中药高通量实验和参考指南数据库 HERB^[45]（中文名“本草组鉴”）。该数据库整合了多个中药数据库，是目前最全面的中医药数据库。HERB 包含了中药的植物学特性、化学成分、药理学作用、临床应用等方面的信息，涵盖 7 263 种中药（如板蓝根、青蒿和人参）、49 258 种成分、12 933 个靶点、28 212 个疾病（如乳腺癌和 COVID-19）以及它们之间的关联关系数据。本文结合 PubChem 数据库补充药物的 InChI 特征。

中药的化学成分复杂，活性物质种类繁多，且作用机制复杂，使中药具有“多成分多靶点”的数据特点，对中药的药理作用进行深入理解变得困难。以中药“人参”为例，其主要的活性物质包括人参皂苷、人参多肽、人参酚及醌类化合物等。人参皂苷能抗菌、抗炎、促进伤口愈合；人参多肽能改善消化功能、增强食欲；人参酚及醌类化合物能抗氧化、延缓细胞衰老、抵御紫外线。中药的活性物质在治疗不同疾病时发挥作用，在中药重定位的研究中，需要明确中药的活性物质及其作用机制，以便将其应用于新的疾病治疗领域。针对中药多种活性物质的情况，本文所提模型将不同成分的适应症进行拼接汇总，形成最佳的预测结果。

- (2) 参数设置
- 按前文的参数设置方法设置 GenDrugShifter 超参数。
- (3) 交叉验证
- 为了直观地评估“时珍”的性能优势和局限性，本文将其与相关先进方法进行了详细的比较。具有代表性的中药药物重

表5 “时珍”与其他模型的比较

模型	准确率	精确率	召回率	F1 分数
AHNPL	84.68%	–	–	0.8311
CAHNPL	88.98%	–	–	0.8454
“时珍”	90.25%	87.90%	87.41%	0.8730

定位研究方法如下。

- AHNPL：结合属性信息和异质网络表示学习的中药药方重定位方法^[43]。
- CAHNPL：基于社群注意力机制增强的中药药方重定位方法^[43]。

(4) 消融实验

通过实施消融实验（见表5），本文验证了 GAT、Transformer Decoder 模块这种组合方法能够提高“时珍”的预测性能。

(5) 案例研究与中药学意义

为了进一步验证“时珍”的预测性能，本文随机选取了 3 种中药（板蓝根、青蒿、人参）作为案例研究，通过“时珍”对这些药物适应症进行预测。表6 是对 3 种中药重定位的预测结果。

为进一步验证预测结果的准确性，本文将 GenDrugShifter 用于寻找乳腺癌和

表6 中药重定位的适应症(部分)

药物	预测的适应症	证明内容
板蓝根	祛湿化痰	TCMBank
	益气平喘	TCMBank
	清热解毒	TCMBank
青蒿	清热祛湿	TCMBank
	温脾止泻	TCMBank
	消食导滞止痛	TCMBank
人参	益气养肺	TCMBank
	消湿开胃	TCMBank
	醒脾和胃	TCMBank

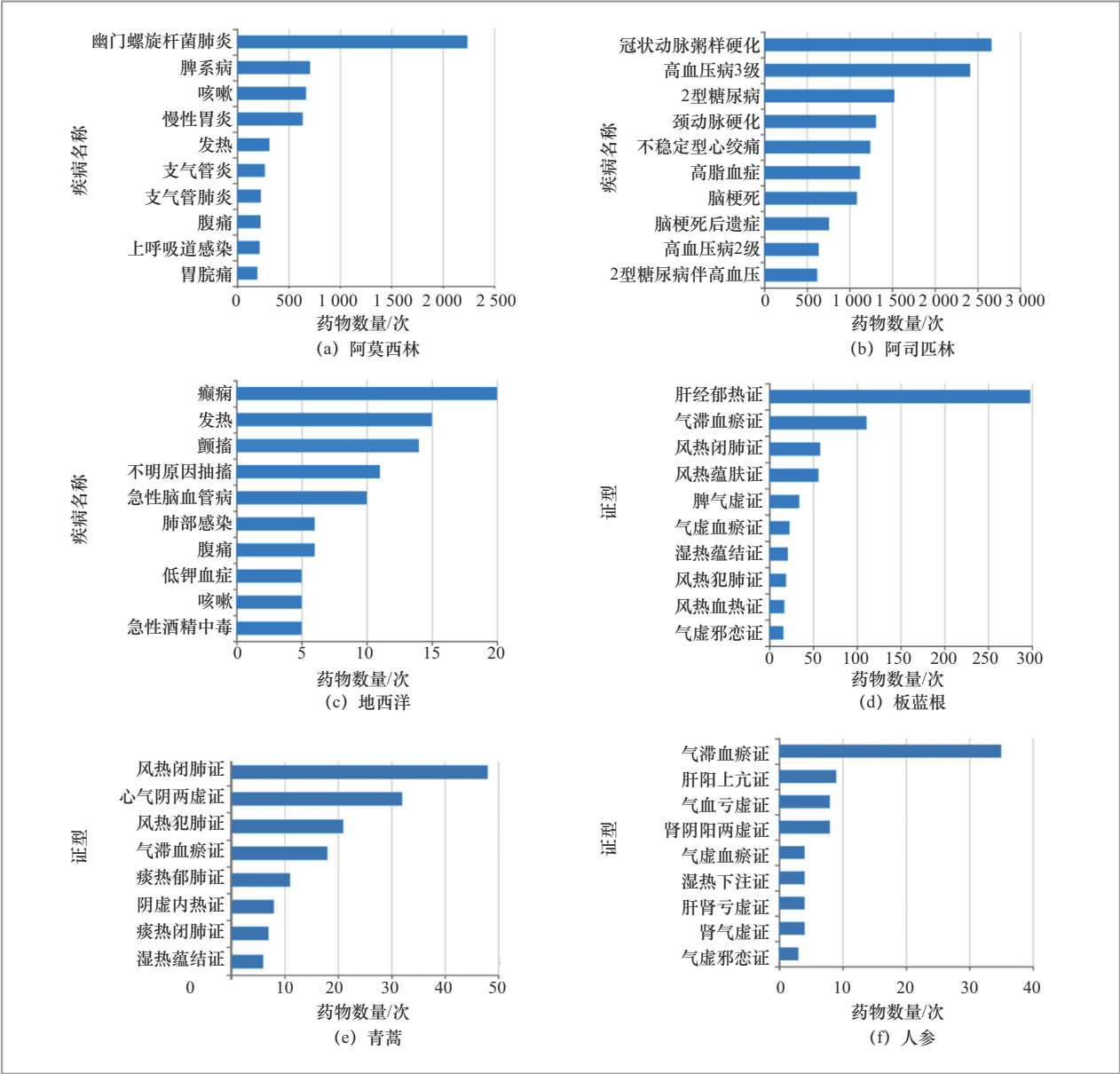


图6 临床数据验证结果

COVID-19 的治疗中药研究。结合图6所示的临床数据验证结果，本文进一步查阅了TCMBank等数据库中的相关记录，发现治疗乳腺癌的黄芪、女贞子、白花蛇舌草是黄芪解毒汤的重要组成；治疗COVID-19的金银花、麻黄、板蓝根、石膏、青蒿等，在清肺排毒汤、连花清瘟胶

囊等“三药三方”中得到临床实践。由此证明了“时珍”预测药物适应症的准确性，疾病潜在的中药发现见表7。

对于没有得到文献证实的预测结果，本文分析中药化合物分子，利用KEGG富集分析寻找药物代谢通路及其内涵机制，给湿实验分析做理论基础。本文利用“时

珍”进行罕见病药物挖掘，针对肌萎缩侧索硬化（amyotrophic lateral sclerosis, ALS），挖掘了肉苁蓉、猪腰子、绵萆薢、菟丝子、杜仲等中药。图7通过疾病候选靶标及中药相关基因的功能富集分析，发现细胞内受体信号通路、腺苷酸环化酶调节G蛋白偶联受体信号通路可能是这5味中药治疗ALS的潜在作用机制，可为湿设计提供参考。其中， $\text{Neglog10}(P_{\text{value}})$ 表示 $-\log_{10}(P_{\text{value}})$ 。 $\log_{10}(P_{\text{value}})$ 越大， P 值越小，表示该通路越显著。

表7 疾病潜在中药发现(部分)

疾病	药物	证据
乳腺癌	黄芪	黄芪解毒汤
	白芍	柴胡白芍乳癌汤
	女贞子	黄芪解毒汤
	茯苓	人参养荣汤
	白花蛇舌草	黄芪解毒汤
COVID-19	金银花	连花清瘟胶囊
	麻黄	清肺排毒汤
	板蓝根	连花清瘟胶囊
	石膏	化湿败毒方
	青蒿	宣肺败毒方

4 临床数据验证与分析

本文使用湖南中医药大学第一附属医

院的数据库，对本文中的案例研究进行了临床数据验证，统计了近3年案例药物的

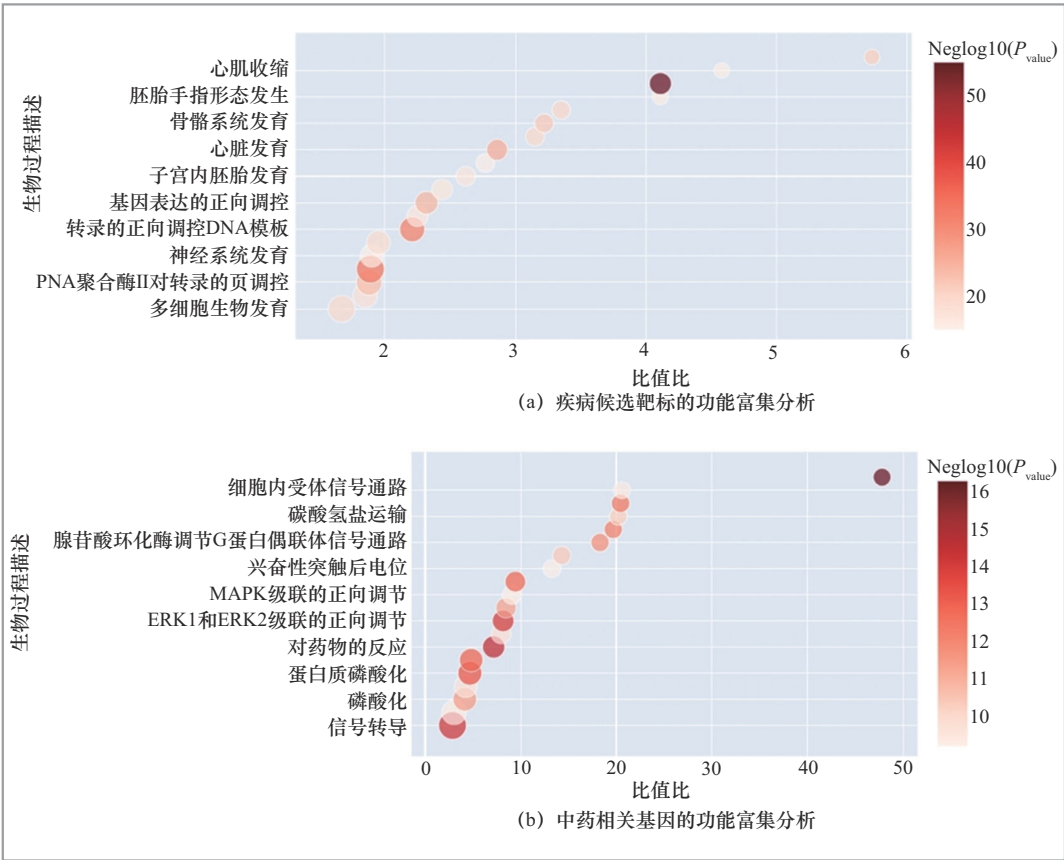


图7 罕见病药物发现功能富集分析

常用治疗疾病证。对 GenDrugShifter 重定位的药物适应症进行了临床应用验证。由图 6 可知，GenDrugShifter 发现了一些药物在临床上的适应症，如阿莫西林治疗肺炎、地西泮治疗癫痫、板蓝根用于祛湿化痰、青蒿用于去湿热等，验证了药物重定位模型的部分预测结果。

GenDrugShifter 重定位预测的药物适应症在 DrugBank 中得到验证后，在医院既往真实诊疗数据得到了进一步验证，如阿莫西林治疗肺炎、地西泮治疗癫痫。阿莫西林是一种广谱半合成抗生素，能抑制细菌细胞壁的合成，具有高效的广谱抗菌作用。其治病机理主要是通过杀灭细菌间接缓解炎症。在临床上，阿莫西林主要用于治疗由细菌引起的多种感染性疾病，包括但不限于呼吸道感染，如咽炎、急性支气管炎和肺炎等。地西泮是一种苯二氮䓬类精神药物，临床主要用于治疗焦虑、失眠、肌肉痉挛及部分癫痫。地西泮具有中枢神经系统抑制作用，可以增强大脑皮层的抑制作用，降低神经元的兴奋性，降低癫痫发作的可能性，此外，地西泮具有增强机体抗惊厥的作用，使肌肉神经放松，减少癫痫发作时的肌肉抽搐和痉挛。以上，为 GenDrugShifter 的发现提供了一定的合理性论据。

同时，“时珍”预测的中药适应症在 TCMBank 中得到验证后，在医院既往真实诊疗数据得到了进一步验证。如板蓝根用于祛湿化痰、青蒿用于去湿热。板蓝根在传统中医中被视为清热解毒的良药，常用于治疗风热感冒、咽喉肿痛、丹毒和痈肿疮毒等病症。板蓝根中的靛蓝、靛玉红等活性成分能有效抑制炎症反应，减轻炎症引起的痰液分泌过多。此外，板蓝根中的挥发油能刺激呼吸道黏膜，促进痰液的排出，实现“祛湿化痰”的功效。青蒿，古名“蒿”，表示“治疗疟疾之草”，始载

于《神农本草经》，具有广泛的药用价值。青蒿具有清热燥湿的功效，主要与其含有的挥发油和黄酮类化合物等活性成分有关。这些化合物具有抑制病原体生长、抗菌、抗病毒等作用，能够有效地降低体内的炎症反应，并对湿热病症有一定的消炎作用。此外，人参具有显著的补气虚效果，其补气虚的活性成分主要是人参皂苷，不同种类的人参皂苷具有不同的药理作用。例如，人参皂苷 Rb1 具有增强胆碱系统的功能，增加乙酰胆碱的合成和释放，从而改善记忆力；人参皂苷 Rb2 对 DNA 和 RNA 的合成有促进作用，还具有抗氧化、清除体内自由基、改善心肌缺血/再灌注损伤等作用；人参皂苷 Rc 具有抑制癌细胞的功能。以上，为“时珍”的发现提供了一定的合理性论据。

临床数据治疗疾病中，阿司匹林常用于治疗脑梗死，而 GenDrugShifter 没有预测到这一发现。阿司匹林是一种常用的抗血小板聚集药物，具有抗血小板聚集的作用，这种特性使其能明显减少周围动脉内形成阻塞性血栓的情况。其提示人们需要对 GenDrugShifter 进行迭代和优化，以及 与生物学家和医学专家合作，共同研究模型的不足、改进药物重定位模；需要通过交流和合作，获取更多的专业知识和数据资源，并整合到模型中，尝试使用不同的模型结构或算法提高模型的预测能力。

综上所述，临床数据进一步验证了实验的结果。GenDrugShifter 发现的药物适应症可以在临床上高频率疾病/证诊断中观察到，如阿莫西林治疗肺炎、地西泮治疗癫痫、板蓝根用于祛湿化痰、青蒿用于去湿热等。同时，需要对模型进一步迭代与优化，提升预测性能。

GenDrugShifter 为临床治疗提供了新的潜在解决方案，其不仅挖掘了多种已被

验证的疾病适应症和疾病的治疗药物，还揭示了多种之前未被发现的疾病潜在治疗药物。GenDrugShifter 预测了一种名为 Pholcldine 的 COVID-19 潜在治疗药物。尽管这一发现尚未在现有文献中得到证实，这种药物可能为抗击 COVID-19 提供新的治疗策略，对于全球公共卫生事业具有重要意义。此外，GenDrugShifter 在罕见病治疗领域也取得了突破。针对罕见且难以治疗的 ALS，GenDrugShifter 发现了一系列中药作为潜在治疗药物，包括肉苁蓉、猪腰子、绵萆薢、菟丝子和杜仲等。这些中药在现有文献中并未被证实为治疗 ALS 的药物，但 GenDrugShifter 的预测结果为药物重定位研究提供了新的方向，说明了 GenDrugShifter 为抗击疾病、拯救生命提供了新的潜在的治疗策略和选择。

5 结束语

本文针对药物重定位领域中的预测局限性问题，提出了基于生成式 AI 的药物重定位模型 GenDrugShifter。通过结合图注意力神经网络和 Transformer Decoder 模块，模型能够全面且系统地预测药物的潜在适应症，突破了传统方法的限制。实验结果表明，GenDrugShifter 在多个性能指标上优于现有的先进方法，且在实际案例预测中得到了临床文献的支持。此外，本文还基于 GenDrugShifter 设计了中药重定位模型“时珍”，并同样取得了显著的预测效果。未来，笔者将继续深入研究生成式 AI 在药物重定位领域的应用，探索整合多模态生物学数据（如靶点通路、临床表型）以提升预测全面性；结合实验验证与医学专家知识，构建可解释的 AI-临床协同决策系统，以推动药物研发领域的创新和发展。

参考文献：

- [1] HUANG K X, CHANDAK P, WANG Q W, et al. A foundation model for clinician-centered drug repurposing[J]. Nature Medicine, 2024, 30(12): 3601-3613.
- [2] PERDOMO-QUINTEIRO P, BELMONTE-HERN N A. Knowledge graphs for drug repurposing: a review of databases and methods[J]. Briefings in Bioinformatics, 2024, 25(6): 461.
- [3] JIN B Z, SHUN L Y, WEN C, et al. Signature Search Polestar: a comprehensive drug repurposing method evaluation assistant for customized oncogenic signature[J]. Bioinformatics, 2024(9): 9.
- [4] YU M, LI W M, YU Y R, et al. Deep learning large-scale drug discovery and repurposing[J]. Nature Computational Science, 2024, 4(8): 600-614.
- [5] 鱼亮, 赵晋. 基于组织特异性和直接邻居相似度方法预测疾病-药物关系[J]. 中国科学: 信息科学, 2019, 49(9): 1175-1185.
YU L, ZHAO J. Prediction of disease-drug relationships based on tissue specificity and direct neighbor similarity[J]. Scientia Sinica (Informationis), 2019, 49(9): 1175-1185.
- [6] 卢艳峰, 杨思瀚, 莫鸿仪, 等. 基于知识图谱嵌入的阿尔茨海默病药物重定位研究[J]. 中国药科大学学报, 2023, 54(3): 344-354.
LU Y F, YANG S H, MO H Y, et al. Drug repurposing for Alzheimer's disease using knowledge graph embedding[J]. Journal of China Pharmaceutical University, 2023, 54(3): 344-354.
- [7] LI Q, LI L, LI Y. Developing ChatGPT for biology and medicine: a complete review of biomedical question answering[J]. Biophysics Reports, 2024, 10(3): 152-171.
- [8] SÉGALAT L. Invertebrate animal models of diseases as screening tools in drug discovery[J]. ACS Chemical Biology,

- 2007, 2(4): 231–236.
- [9] RAVASSARD P, HAZHOUS Y, PECHBERTY S, et al. A genetically engineered human pancreatic β cell line exhibiting glucose-inducible insulin secretion[J]. Journal of Clinical Investigation, 2011, 121(9): 3589–3597.
- [10] HANN M M, LEACH A R, HARPER G. Molecular complexity and its impact on the probability of finding leads for drug discovery[J]. Journal of Chemical Information and Computer Sciences, 2001, 41(3): 856–864.
- [11] SAINI B. Validated HPLC–UV method for quantification of mirtazapine in human whole blood[C]//Proceedings of the Indo-US Symposium on “Frontiers in Medicinal Chemistry and Drug Discovery”. [S.l.:s.n.], 2011.
- [12] GAO Z X, LI L Y, MA S Y, et al. Examining the potential of ChatGPT on biomedical information retrieval: fact-checking drug–disease associations[J]. Annals of Biomedical Engineering, 2024, 52(8): 1919–1927.
- [13] BRAN A M, COX S, SCHILTER O, et al. Augmenting large language models with chemistry tools[J]. Nature Machine Intelligence, 2024, 6(5): 525–535.
- [14] WANG Q, DOWNEY D, JI H, et al. Learning to generate novel scientific directions with contextualized literature-based discovery[EB]. arXiv preprint, 2023, arXiv: 2305.14259.
- [15] LUO J Y, QIAN C, WANG X C, et al. pADR: towards personalized adverse drug reaction prediction by modeling multi-sourced data[C]//Proceedings of the ACM International Conference on Information & Knowledge Management ACM International Conference on Information and Knowledge Management. [S.l.: s.n.], 2023, 4724–4730.
- [16] JOSHI P, MASILAMANI V, MUKHERJEE A. A knowledge graph embedding based approach to predict the adverse drug reactions using a deep neural network[J]. Journal of Biomedical Informatics, 2022, 132: 104122.
- [17] LEE C Y, CHEN Y P. Descriptive prediction of drug side-effects using a hybrid deep learning model[J]. International Journal of Intelligent Systems, 2021, 36(6): 2491–2510.
- [18] NAFEA A A, IBRAHIM M S, MUKHLIF A A, et al. An ensemble model for detection of adverse drug reactions[J]. Aro-the Scientific Journal of Koya University, 2024, 12(1): 41–47.
- [19] PUKKILA-WORLEY R, HOLSON E, WAGNER F, et al. Antifungal drug discovery through the study of invertebrate model hosts[J]. Current Medicinal Chemistry, 2009, 16(13): 1588–1595.
- [20] LUBLIN A L, LINK C D. Alzheimer’s disease drug discovery: in vivo screening using *Caenorhabditis elegans* as a model for β -amyloid peptide-induced toxicity[J]. Drug Discovery Today Technologies, 2013, 10(1): e115–e119.
- [21] BORDOGNA A, PANDINI A, BONATI L. Predicting the accuracy of protein–ligand docking on homology models[J]. Journal of Computational Chemistry, 2011, 32(1): 81–98.
- [22] CLAUS B L, UNDERWOOD D J. Discovery informatics: its evolving role in drug discovery[J]. Drug Discovery Today, 2002, 7(18): 957–966.
- [23] LI X M, WANG J W, AN Z L, et al. Optically imageable metastatic model of human breast cancer[J]. Clinical & Experimental Metastasis, 2002, 19(4): 347–350.
- [24] BERGSTRÖM R. The role of the pharmaceutical industry in meeting the public health threat of antibacterial resis-

- tance[J]. Drug Resistance Updates, 2011, 14(2): 77–78.
- [25] FITZGERALD G A. Opinion: anticipating change in drug development: the emerging era of translational medicine and therapeutics[J]. Nature Reviews Drug Discovery, 2005, 4(10): 815–818.
- [26] NG M S Y, KAUR G, FRANCIS R S, et al. Drug repurposing for glomerular diseases: an underutilized resource[J]. Nature Reviews Nephrology, 2024, 20(11): 707–721.
- [27] LIU I T T, KESSELHEIM A S. Drug repurposing via the best pharmaceuticals for children act[J]. JAMA Pediatrics, 2024, 178(9): 855–856.
- [28] LEINONEN H, ZHANG J Y, OCCELLI L M, et al. A combination treatment based on drug repurposing demonstrates mutation-agnostic efficacy in pre-clinical retinopathy models[J]. Nature Communications, 2024, 15(1): 5943.
- [29] GHARIZADEH A, ABBASI K, GHAR-EYAZI A, et al. HGTDR: Advancing drug repurposing with heterogeneous graph transformers[J]. Bioinformatics, 2024, 40(7): btac349.
- [30] CHAKRABORTY A, GHOSH R, SOUMYA MOHAPATRA S, et al. Repurposing of antimycobacterium drugs for COVID-19 treatment by targeting SARS CoV-2 main protease: an in-silico perspective[J]. Gene, 2024, 922: 148553.
- [31] ABU-ZAHRA T, GRIMM S E, SCHOLTE M, et al. How health technology assessment can help to address challenges in drug repurposing: a conceptual framework[J]. Drug Discovery Today, 2024, 29(6): 104008.
- [32] MORAL-TURÓN C, ASENCIO-CORTÉS G, RODRIGUEZ-DIAZ F, et al. ASACO: Automatic and Serial Analysis of CO-expression to discover gene modifiers with potential use in drug repurposing[J]. Briefings in Functional Genomics, 2024, 23(4): 484–494.
- [33] CAO D S, ZHANG L X, TAN G S, et al. Computational prediction of drug target interactions using chemical, biological, and network features[J]. Molecular Informatics, 2014, 33(10): 669–681.
- [34] MENG Y J, LU C C, JIN M, et al. A weighted bilinear neural collaborative filtering approach for drug repositioning[J]. Briefings in Bioinformatics, 2022, 23(2): bbab581.
- [35] ZHAO T Y, HU Y, VALSDOTTIR L R, et al. Identifying drug-target interactions based on graph convolutional network and deep neural network[J]. Briefings in Bioinformatics, 2021, 22(2): 2141–2150.
- [36] ZHANG G, GAO Z J, YAN C K, et al. KGANSynergy: knowledge graph attention network for drug synergy prediction[J]. Briefings in Bioinformatics, 2023, 24(3): bbad167.
- [37] LUO H M, WANG J X, LI M, et al. Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm[J]. Bioinformatics, 2016, 32(17): 2664–2671.
- [38] LIANG H Y, CHEN L, ZHAO X, et al. Prediction of drug side effects with a refined negative sample selection strategy[J]. Computational and Mathematical Methods in Medicine, 2020, 2020(1): 1573543.
- [39] ZENG X X, ZHU S Y, LIU X R, et al. deepDR: a network-based deep learning approach to in silico drug repositioning[J]. Bioinformatics, 2019, 35(24): 5191–5198.
- [40] JIN S T, NIU Z M, JIANG C Z, et al. HeTDR: Drug repositioning based on heterogeneous networks and text mining[J].

Patterns, 2021, 2(8): 100307.

[41] YU Z X, HUANG F, ZHAO X H, et al. Predicting drug - disease associations through layer attention graph convolutional network[J]. Briefings in Bioinformatics, 2021, 22(4): bbab243.

[42] CAI L J, LU C C, XU J L, et al. Drug repositioning based on the heterogeneous information fusion graph convolutional network[J]. Briefings in Bioinformatics, 2021, 22(6): bbab319.

[43] 朱润. 基于异质网络表示学习的中药药方重定位方法研究与实现 [D]. 南京: 东南大学, 2022.

ZHU R. Research and implementation of TCM prescription repositioning method based on heterogeneous network representation learning[D]. Nanjing: Southeast University, 2022.

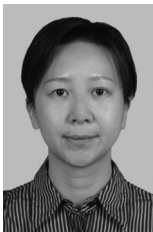
[44] WU Y, ZHANG F L, YANG K, et al. SymMap: an integrative database of traditional Chinese medicine enhanced by symptom mapping[J]. Nucleic Acids Research, 2019, 47(D1): D1110-D1117.

[45] FANG S S, DONG L, LIU L, et al. HERB: a high-throughput experiment- and reference-guided database of traditional Chinese medicine[J]. Nucleic Acids Research, 2021, 49(D1): D1197-D1206.

作者简介



龚后武（1988- ），男，湖南大学信息科学与工程学院博士生，主要研究方向为生物医学大数据与人工智能。



金敏（1973- ），女，湖南大学信息科学与工程学院教授、博士生导师，主要研究方向为大数据与数据科学、机器学习与深度学习。

收稿日期: 2024-10-17

通信作者: 金敏, jinmin@hnu.edu.cn

基金项目: 湖南省自然科学基金项目(No.2025JJ50343) ;国家重点研发计划项目(No.2023YFC3503404)

Foundation Items: The Natural Sciences Foundation of Hunan Province (No. 2025JJ50343); The National Key Research and Development Program of China (No.2023YFC3503404)