

· 药咖论坛 ·

基于变分自编码器神经网络的脑卒中药物重定位研究

李瑞丰¹, 张议月¹, 罗秀菊², 彭军^{1*}

(1. 中南大学湘雅药学院药理系, 湖南 长沙 410013; 2. 中南大学湘雅三医院检验系, 湖南 长沙 410013)

[摘要] 目的 构建基于变分自编码器的神经网络模型, 预测不同药物作用机制的相似性, 支持高效筛选脑卒中候选治疗药物, 提高药物研发效率。**方法** 设计双编码器-解码器对比学习神经网络模型, 整合药物作用下的单细胞转录组测序 (RNA-seq) 数据, 构建药物扰动基因表达数据训练集用于模型训练; 将药物表型效应特征映射至双曲潜在空间; 基于潜在空间中药物特征表示的距离进行药物重定位, 筛选出候选药物; 通过生物信息学分析对筛选结果进行验证。**结果** 构建的双曲对比学习变分自编码器 (hyperbolic contrastive variational autoencoder, HCVAE) 模型, 在聚类效果上显著优于标准变分自编码器 (variational autoencoder, VAE) 模型, 能有效提取药物表型特征; 通过模型推断筛选出赛度替尼等与依达拉奉具有相似效应的候选药物。进一步分析表明, 赛度替尼的表型效应与其对细胞凋亡过程中受体相互作用蛋白激酶 1、caspase 等关键蛋白的调控相关。**结论** HCVAE 模型能有效提取“药物-疾病”表型特征, 筛选出与依达拉奉效应具有高度相似性的缺血性脑卒中候选治疗药物赛度替尼, 该模型可作为脑卒中药物筛选的高效平台, 同时为药物重定位研究提供新思路。

[关键词] 变分自编码器; 对比学习; 赛度替尼; 依达拉奉; 脑卒中

[中图分类号] R91; TP183

[文献标志码] A

[文章编号] 1001-5094 (2026) 06-0496-12

DOI: 10.20053/j.issn1001-5094.202510190795

Drugs Repositioning for Ischemic Stroke Based on Variational Autoencoder Neural Network

LI Ruifeng¹, ZHANG Yiyue¹, LUO Xiuju², PENG Jun¹

(1. Department of Pharmacology, Xiangya School of Pharmaceutical Science, Central South University, Changsha 410013, China;

2. Department of Laboratory Medicine, the Third Xiangya Hospital, Central South University, Changsha 410013, China)

[Abstract] Objective This study aimed to predict the similarity in the mechanisms of action for different drugs, facilitate the efficient screening of candidate therapeutics for ischemic stroke, and enhance drug development efficiency through a modified variational autoencoder-based neural network. **Methods** A dual encoder-decoder variational autoencoder (VAE), termed hyperbolic contrastive VAE (HCVAE), was developed to extract features from training datasets comprising integrated drug-perturbed single-cell RNA sequencing (RNA-seq) gene expression profiles. Phenotypic effect features of drugs were mapped into a hyperbolic latent space. Drug repositioning was performed based on the distances between drug feature representations in the latent space to identify candidate compounds. Biological effects were preliminarily validated through bioinformatics analyses. **Results** The constructed HCVAE model demonstrated superior performance compared to standard VAEs in effectively extracting drug phenotypic features. Cerdulatinib was identified as a candidate therapeutic exhibiting effects comparable to edaravone in ischemic stroke models. Further analysis indicated that cerdulatinib's phenotypic effects are associated with the regulation of key proteins RIPK1 and caspases within the cellular apoptosis pathway. **Conclusion** The HCVAE model effectively extracts “drug-disease” phenotypic features and can serve as a robust platform for candidate drug screening, providing novel insights for drug repositioning research.

[Key words] variational autoencoder; contrastive learning; cerdulatinib; edaravone; ischemic stroke

缺血性脑卒中是因脑部供血不足引发缺氧和能量代谢紊乱^[1-2], 最终导致神经细胞死亡的一类疾病。缺血性脑卒中涉及多种细胞死亡方式相互交集, 形成复杂的调控网络^[3-8], 致使治疗靶点难以精准识别, 因此药物研发周期长、成功率较低。

近年来, 计算生物学、网络药理学和人工智能技术的迅速发展极大地推动了药物研发领域的革新^[9], 单细胞测序技术的出现为解析疾病状态等医

学研究带来了革命性突破^[10-12]。深度学习技术凭借处理高维稀疏数据和捕捉复杂非线性关系的优势, 已逐渐成为单细胞分析的高效工具^[13-14], 可对差异表达基因和致病通路进行深入分析挖掘。因此, 在基于药物-药物相似性进行药物重定位研发领域中, 传统依赖疾病与药物基因表达匹配^[15-19]的研究模式, 正逐步转为集成化的深度学习模型策略, 通过训练人工神经网络, 构建药物-靶点作用模型 (drug targets interaction, DTI), 从而高效筛选候选药物^[20]。

变分自编码器 (variational autoencoder, VAE) 是一类重要的深度学习生成模型^[21], 自开发以来便在表征学习与生成建模领域受到广泛关注^[22]。VAE 在提取高维数据的低维潜在表征方面具有强大能

接受日期: 2025-12-02

基金项目: 国家自然科学基金资助项目 (No. 82173815, No. 81872873)

***通信作者:** 彭军, 教授, 博士生导师

研究方向: 心脑血管疾病发生机制、心脏药物毒性机制及药物防治

E-mail: junpeng@csu.edu.cn

力^[23], 尤其在生物医学数据挖掘领域展现出显著的应用潜力^[24-26]。

本研究基于已上市药物的成熟药理特性, 整合 VAE 神经网络等深度学习方法提取药物在潜在空间的表示相似性信息, 并识别脑卒中复杂疾病机制中的关键基因变化特征, 高效筛选与已验证有效药物高度相似的候选分子。本研究构建了一种用于脑卒中治疗药物筛选的平台模型, 并提出了基于效应模拟相似性进行药物重定位的新思路。

1 材料与方法

1.1 模型训练数据集构建

通过美国国家生物技术信息中心 (national center for biotechnology information, NCBI) 在线数据库检索获取脑卒中相关单细胞 RNA-seq 数据 (GSE180601、GSE174450、GSE183546、GSE207034、GSE268634、GSE151997、GSE141570), 提取其中药物扰动基因表达数据, 与 L1000 数据集^[27] (GSE70138, 包含约 50 个人类细胞系在不同浓度下对约 20 000 种化合物的基因表达谱) 融合形成新的药物扰动数据集。该数据集每个样本检测约 1 000 个特征基因, 共涵盖 3 924 种不同小分子化合物及空白载体处理、药物扰动的 314 个独特细胞系。该数据集中所有细胞系 (包括非脑细胞的细胞系) 均参与模型训练。

将数据集分为扰动组 (treat, 细胞经过药物扰动处理) 与对照组 (control, 细胞未经药物扰动或采用溶剂、空白载体共孵育处理, 或药物扰动 0 时刻时细胞进行测序)。每个观测样本至少包括 landmark 基因 (参照 L1000 数据集)、细胞类型标签、扰动药物名称或 SMILES 标签等。采用 R 软件 (RStudio Server 2023.09.0 Build 463) 进行基因表达数据处理、转换。所有样本的表达数据均进行标准化处理。

1.2 模型框架构建与训练

1.2.1 编码器 编码器 (encoder) 由 2 个并行的全连接线性层 (multilayer perceptron, MLP, 又称多层感知机) 组成, 其中一个编码器用于提取共享特征 (细胞与空白载体、溶剂共孵育的对照组), 另一个编码器用于提取药物处理、扰动后的特异性特征 (细胞与药物共孵育的实验组)。

$\text{inv_exp_map}(x, loc) =$

$$\frac{\text{lorentzian_distance}(x, \mu)}{\sqrt{\max(-\text{lorentzian_product}(x, \mu), 1 + 10^{-6}) + 1} \cdot \sqrt{\max(-\text{lorentzian_product}(x, \mu), 1 + 10^{-6}) - 1}} \cdot (x - \max(-\text{lorentzian_product}(x, \mu), 1 + 10^{-6}) \cdot \mu) \quad (\text{公式3})$$

通过 2 个编码器将数据分别投影到 2 个不同的 32 维潜在分布的参数上, 分别代表对照组特征与实验组特征。每一个编码器均基于标准变分自编码器设计, 其功能是将输入数据 X 映射到潜在空间的参数分布, 其变分推断中对后验分布的高斯近似可表示为公式 1:

$$q_{\phi}(Z|X) = N(Z; \mu, \sigma^2 I) \quad (\text{公式1})$$

其中, X 是观测数据; Z 是数据的潜在表示, 通常为低维向量; $N(Z; \mu, \sigma^2 I)$ 表示多元正态分布, 其中 Z 是隐变量; μ 和 σ 是分布的均值和标准差, 标准差通常取对数: $\log(\sigma^2)$; I 是单位矩阵。

用 MLP 提取输入数据的特征。MLP 是一种全连接的前馈神经网络 (feed forward neural network, FNN), 由多层神经元 (输入层、隐藏层、输出层) 组成使用非线性激活函数。

1.2.2 解码器 解码器 (decoder) 根据潜在变量 z , 生成重构数据的条件分布, $P_{\theta}(Z|X)$ 表示近似真实情况的后验分布。

解码器采用与编码器完全对称的网络结构反向构建, 并与 2 个编码器连接; 输出层根据数据类型 $N(Z; \mu, \sigma^2 I)$ 选为高斯分布。

1.2.3 潜在空间 标准 VAE 模型的潜在空间通常设计为标准高斯分布, 本研究针对 RNA-seq 数据高维、稀疏性特点进行建模, 将潜在空间嵌入双曲空间。高维双曲空间的正态分布概率密度函数 (PDF) 表示为公式 2:

$$\text{Log}(x) = \log p_{\text{normal}}(v) - C(\text{dim}, d(x, loc)) \quad (\text{公式2})$$

x 是超双曲空间中的一个点 [RNA-seq 数据中一个样本 (如细胞的 RNA 表达谱) 的双曲嵌入向量在超双曲空间中的坐标]。loc 是超双曲空间中的位置 [均值 μ , 表示类别中心 (如细胞类型、分化进展阶段的均值点), 在双曲空间中代表该类型细胞的“层次锚点”]。

$v = \text{inv_exp_map}(x, loc)$ 为点 loc 处的切向量, 其中 inv_exp_map 代表逆指数映射 (inverse exponential map)。在双曲空间中, 指数映射 (exponential map) 和逆指数映射是将切空间与流形点之间进行映射的数学工具。通过计算点 loc 到点 x 的逆指数映射得到切向量 v , 表示为公式 3:

其中, $\text{lorentzian_product}$ (洛伦兹内积) 是计算 2 个向量 x 和 μ 在超双曲空间中的点积, 其定义为公式 4:

$$\langle x, y \rangle = -x_0 y_0 + \sum x_i y_i \quad (\text{公式 4})$$

其中, $x = (x_0, x_1, \dots, x_n)$, $y = (y_0, y_1, \dots, y_n)$ 。

$\text{lorentzian_distance}$ (洛伦兹距离) 是计算 2 个向量 x 和 μ 在超双曲空间中的距离, 其定义为公式 5:

$$d(x, y) = \ln(-\langle x, y \rangle + \sqrt{\langle x, y \rangle^2 - 1}) \quad (\text{公式 5})$$

$p_{\text{normal}}(v)$ 是均值为 0, 标准差为 scale (控制数据的离散程度) 的正态分布在向量 v 上的概率密度函数, 忽略 v 的第一个元素。 P 的概率密度函数符合正态分布函数定义为公式 6:

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (\text{公式 6})$$

其中, μ 是均值, σ 是标准差。

$C(\text{dim}, d(x, \text{loc}))$ 是一个修正项, 用于考虑超双曲空间的曲率, 其中 dim 是超双曲空间的维度, $d(x, \text{loc})$ 是 x 和 loc 之间的超双曲距离。

采用的重参数化的计算方法可表示为公式 7:

$$t > z = \mu + \sigma \odot \varepsilon, \varepsilon \sim N(0, I) \quad (\text{公式 7})$$

其中, $\odot$ 表示逐元素乘法。

1.2.4 损失函数 本研究构建模型的损失来自以下 5 个部分。

重构损失包括 2 个部分。

Loss 1: 处理组 (药物扰动) 数据经过解码器的输出与原始输入数据分布的相近程度, 定义为 $\log_likelihood_xt$ 。

Loss 2: 对照组 (载体、溶剂处理组) 数据经过解码器的输出与原始输入数据分布的相近程度, 定义为 $\log_likelihood_xc$ 。

KL 散度 (KL_divergence , D_{KL}) 损失包括 3 个部分。

散度损失为编码器输出的潜在空间分布与先验分布之间的差异程度, 在本研究中先验分布为高维双曲空间的正态分布, 则该损失项可定义为公式 8:

$$D_{\text{KL}}(P(x|\theta^*)||P(x|\theta)) = E_{x \sim P(x|\theta^*)} \left[\log \frac{P(x|\theta^*)}{P(x|\theta)} \right] = E_{x \sim P(x|\theta^*)} [\log P(x|\theta^*)] - E_{x \sim P(x|\theta^*)} [\log P(x|\theta)] \quad (\text{公式 8})$$

其中, θ^* 是本研究的真实数据分布的参数, θ 是本研究由训练数据得到的估计参数。

Loss 3: 处理组 (药物扰动) 数据经过编码器 E1 输出的潜在空间分布与先验分布之间的差异程度, 定义为 KL_tz 。

Loss 4: 对照组 (载体、溶剂处理组) 数据经过编码器 E1 输出的潜在空间分布与先验分布之间的差异程度, 定义为 KL_cz 。

Loss 5: 处理组 (药物扰动) 数据经过编码器 E2 输出的潜在空间分布与先验分布之间的差异程度, 定义为 KL_ts 。

最终的模型训练的目标函数定义为公式 9:

$$\text{Loss} = (\log_likelihood_xt + \log_likelihood_xc) - (\text{KL_tz} + \text{KL_cz} + \text{KL_ts}) \quad (\text{公式 9})$$

1.2.5 模型训练 将对照组数据、扰动组数据输入网络, 训练自编码器神经网络参数, 并对模型的损失函数值及隐空间参数进行可视化。采用定义好的损失函数, 通过优化算法 (如随机梯度下降、Adam 优化器) 来训练模型。在训练过程中, 同时优化编码器和解码器的参数。模型经过迭代训练 (初始 10 000 次, $\text{epochs} = 10\,000$), 通过最大化变分推断下界 (ELBO) 最小化损失函数 [$\text{Loss} = -(\log_likelihood - \text{KL_divergence})$] 直至收敛停止训练。

标准 VAE 模型训练包括以下 3 个阶段。

1) 编码阶段 输入 x (扰动组) 通过多层全连接网络 (含 ELU 激活和 BatchNorm) 逐步压缩至 32 维潜空间表示: $h_{\text{enc}} = \text{ELU}[\text{BatchNorm}(W_1 \cdot x + b_1)]$, 其中 W_1 , b_1 为编码器权重与偏置。

2) 潜空间参数计算 从 h_{enc} 推导潜变量分布的均值 μ 和方差 σ^2 : $\mu = W_\mu \cdot h_{\text{enc}} + b_\mu$, $\sigma^2 = \text{Softplus}(W_\sigma \cdot h_{\text{enc}} + b_\sigma)$, 其中 Softplus 函数 $\text{Softplus}(x) = \log(1 + e^x)$ 确保 $\sigma^2 > 0$ 。

3) 解码阶段 将 μ 与 σ^2 拼接为 4 维输入 $z_{\text{input}} = [\mu; \sigma^2]$, 通过对称的解码网络重建输出:

$\hat{x} = W_{\text{out}} \cdot \text{ELU}[\text{BatchNorm}(W_2 \cdot z_{\text{input}} + b_2)]$, 其中 W_2 , b_2 为解码器权重与偏置, 输出层线性映射至 978 维。

双曲对比学习变分自编码器 (hyperbolic contrastive variational autoencoder, HCVAE) 模型训练过程与标准 VAE 模型训练过程类似, 主要区别体现在编码阶段。在 HCVAE 模型中, 输入数据包括 xt (扰动组) 和 xc (对照组)。扰动组数据 xt 通过 2 个不同的编码器进行编码, 而对照组数据 xc 仅通过一个编码器进行编码。之后, HCVAE 模型的训练过程与标准 VAE 模型相同, 继续进行数据重建和隐空间正则化, 以实现模型目标函数的收敛。

本研究分别训练标准 VAE 模型和 HCVAE 模型。2 种模型均采用 Adam 随机优化算法进行训练, 学习率设置为 0.000 5, mb_size 为 256 (即小批量样本数, 该参数会影响显存占用、训练速度、梯度噪声及最终泛化性能)。使用 Python 的 Pytorch 库进

行模型构建及训练。

1.3 模型性能评估

本研究采用外部聚类指标[标准化互信息(normalized mutual information, NMI)、调整兰德指数(adjusted rand index, ARI)]和内部聚类指标[平均轮廓宽度(average silhouette width, ASW)、Calinski-Harabasz 指数(Calinski-Harabasz index, CH)、Davies-Bouldin 指数(Davies-Bouldin index, DB)], 评估 VAE、HVAE 模型的性能。

1.3.1 标准化互信息 NMI 常用于度量 2 个聚类结果的相近程度(通常用于对比聚类结果和真实标签的相似程度), 其取值范围为[0, 1], 数值越高表示 2 个聚类结果越相似。NMI 可表示为公式 10:

$$NMI(Y; C) = \frac{2 \times I(Y; C)}{H(Y) + H(C)} \quad (\text{公式 10})$$

其中, Y 代表数据类别的真实值; C 代表聚类的结果; $H(X)$ 代表交叉熵, 可表示为公式 11:

$$H(X) = \sum_{i=1}^{|x|} P(i) \cdot \log P(i) \quad (\text{公式 11})$$

$I(Y; C)$ 代表互信息, 可表示为公式 12:

$$I(Y; C) = H(Y) - H(Y|C) \quad (\text{公式 12})$$

1.3.2 调整兰德指数 ARI 以 C 表示真实类别划分, K 表示聚类算法得到的聚类结果。定义 a 为在真实划分 C 中属于同一类, 且在聚类结果 K 中也被划分为同一簇的实例对数量。定义 b 为在真实划分 C 中属于不同类, 且在聚类结果 K 中也被划分为不同簇的实例对数量。在此基础上, ARI 的定义为公式 13:

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}}{\frac{1}{2} \left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2} \right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}} \quad (\text{公式 13})$$

ARI 的取值范围为[-1, 1], 其数值越大表示聚类结果与真实类别划分的一致性越好, 即 2 种划分的重叠程度越高。需注意的是, 使用 ARI 作为评估指标的前提是数据本身具有类别标记。

1.3.3 聚类评价指标 通过轮廓系数对聚类效果进行评价。轮廓系数[silhouette score, 又称平均轮廓宽度(ASW)]通过比较类内距离与最近类间距离来衡量聚类的质量。轮廓系数值越高, 聚类结果越好, 其定义为公式 14:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (\text{公式 14})$$

1.3.4 Calinski-Harabasz 指数 CH 指数通过计算簇间散度矩阵(即簇间的离散度)和簇内散度矩阵(即簇内的紧密度)之比来评估聚类质量。该指数是

比较聚类间和聚类内协方差, 其数值越大表示聚类的效果越好, 聚类结果越有效。CH 指数的定义为公式 15:

$$CH = \frac{tr(B_k)}{tr(W_k)} \times \frac{n-k}{k-1} \quad (\text{公式 15})$$

其中, $tr(B_k)$ 为簇间散度矩阵 B_k 的迹, 即矩阵对角线元素的和, 代表簇间的分离度。 $tr(W_k)$ 为簇内散度矩阵 W_k 的迹, 代表簇内的紧密度。 n 为数据集的样本数量, k 为簇的数量。

1.3.5 Davies-Bouldin 指数 DB 指数用于评估子群(或聚类)的紧密度和分离度。簇与簇之间的相似度越高, DB 指数越大, 表明簇与簇之间距离越小, 聚类结果越差; 反之, DB 指数越小, 聚类结果越理想。DB 指数的定义为公式 16:

$$DB = \frac{1}{c} \times \sum_{i=1}^c \max_{i \neq j} \left(\frac{S_i + S_j}{M_{ij}} \right) \quad (\text{公式 16})$$

S_i 表示第 i 个簇中所有样本点到簇中心距离的平均值, 又称簇内直径; M_{ij} 表示第 i 个簇与第 j 个簇之间的距离(即 2 个簇中心之间的距离)。基于上述参数, 簇 i 与簇 j 之间的相似度 R_{ij} 定义为公式 17:

$$R_{ij} = \frac{S_i + S_j}{M_{ij}} \quad (\text{公式 17})$$

DB 指数可表示为公式 18:

$$DB = \frac{1}{c} \times \sum_{i=1}^c \max_{i \neq j} (R_{ij}) \quad (\text{公式 18})$$

其中, c 表示总簇数。对于第 i 个簇, 其相似度 $R_i = \max\{R_{i1}, R_{i2}, \dots, R_{i(c-1)}\}$ (其中 c 为聚类的总簇数); DB 指数为所有簇相似度的平均值。DB 指数的取值范围为[0, +∞], 其数值越小表示聚类结果越好。

1.3.6 Pearson 相关指标 本研究使用 Pearson 相关矩阵中上三角元素的绝对值之和作为线性相关指标, 其定义为公式 19:

$$S = \sum_{i=1}^n \sum_{j=i+1}^n |R_{ij}| \quad (\text{公式 19})$$

该指标的取值范围为[0, 1], 其数值越接近 1, 表示特征间线性相关性程度越高。

1.4 模型参数优化

根据损失函数值(loss function)的收敛情况及性能评估指标, 对模型的批次大小(batch_size)、学习率、编码器结构维度等超参数进行调优。使用 Pytorch、seaborn、matplotlib 等 Python 开源深度学习库进行模型的构建。

1.5 药物效应相似性推断

1.5.1 药物在潜空间定位 将 RNA-seq 数据(1 000 个 landmark 关键基因的表达向量)输入预训练模型, 经过编码器处理后, 输出潜在空间均值 μ (潜

在空间为双曲空间, 维度为3维)。随后, 通过维度变换, 将 μ 降维为2维, 得到该数据点在坐标系中的 x 和 y 坐标, 进一步计算药物在潜空间定位, 其可表示为公式20:

$$e' = \left(\frac{e_1}{1 + e_0}, \frac{e_2}{1 + e_0} \right) \quad (\text{公式20})$$

$$d(p_1, p_2) = \log \left(\frac{\sqrt{(nx - my)^2 + (mx + ny - 1)^2} + \sqrt{(m - x)^2 + (n - y)^2}}{\sqrt{(nx - my)^2 + (mx + ny - 1)^2} - \sqrt{(m - x)^2 + (n - y)^2}} \right) \quad (\text{公式21})$$

其中, 两个点的坐标分别为 $p_1 = (x, y)$ 和 $p_2 = (m, n)$ 。按照距离从小到大, 将被测药物进行排序, 距离值越小, 表明二者的药效越相似。

1.6 生物信息学分析

1.6.1 药物作用张量分析 本研究基于NCBI GEO数据库的多项数据集 (GSE180601、GSE174450、GSE183546、GSE141570、GSE151997), 涉及 selinexor、MI-773、DGMI、edaravone、cerdulatinib 等多种药物, 构建了药物在 SH-SY5Y 细胞系获得的基因表达张量。

化合物 selinexor (KPT-330) 是一种口服生物有效的选择性染色体区域维护蛋白1 (chromosomal maintenance 1, CRM1) 抑制剂。化合物 MI-773 (SAR405838) 是一种口服生物有效的双微体同源基因2蛋白 (mousedouble minute 2, MDM2) 拮抗剂。化合物 DGMI 是一种新型银杏叶提取物制剂 (银杏二萜内酯葡胺注射液), 其通过调控炎症和凋亡信号通路减轻低氧诱导的细胞损伤。化合物依达拉奉 (edaravone) 在脑卒中治疗中具有显著疗效, 尤其与右莰醇联合使用时, 能有效保护脑细胞, 减轻神经损伤。此外, GSK-J4 作为一种有效的 H3K27me3 组蛋白赖氨酸脱甲基酶抑制剂, 在癌症研究特别是视网膜母细胞瘤的治疗中展现出应用前景。

如图1A所示, 本研究构建的张量为3维数组 (如 $m \times n \times p$), 即药物 $\times$ 基因表达 $\times$ 试验条件; 如图1B所示, 实验条件以细胞是否经药物处理分为: 处理组 (treat)、未处理组 (baseline)。张量中每个元素 $D[i, j, k]$ 表示药物 i 通过处理方式 j 对 SH-SY5Y 细胞系处理后基因 k 的表达量。通过对药物维度进行张量分解, 提取药物潜在特征。经过张量匹配 (药物 $\times$ 药物), 分析药物与药物之间的相似性关系。

采用非负矩阵因子分解 (non-negative matrix factorization, NMF) 的扩展形式进行特征提取和分析。该方法假设数据中的每个元素均为非负, 且可在某些情况下更有效地捕捉到更加底层的特征。NMF 分解可表示为公式22:

其中, e_0 、 e_1 、 e_2 是模型输出的潜在空间 μ 的3个维度的变量数据。

1.5.2 药物相似性计算 以各个药物在隐空间的映射结果的相互距离表征药效相似程度, 双曲空间采用 Poincaré 盘模型, 在 Poincaré 盘模型中两点之间的双曲距离定义为公式21:

$$V(n \times m) = W(n \times k) \times H(k \times m) \quad (\text{公式22})$$

其中, $V(n \times m)$ 表示 R^n 空间中的 m 个数据按列排列; $W(n \times k)$ 表示 R^n 空间中的 k 个 n 基向量按列排列; $H(k \times m)$ 表示矩阵 H 的每一列向量为矩阵 V 对应列向量投影到 W 上的坐标。

该方法通过优化矩阵分解的目标函数进行计算, 可表示为公式23:

$$\min(\|V - V'\|^n) = \min\left(\sum_{ij} (V_{ij} - V'_{ij})^2\right) \quad (\text{公式23})$$

采用如下公式24、公式25对矩阵 W 和矩阵 H 进行迭代:

$$H_{ij}^{k+1} = H_{ij}^k \frac{\left((W^k)^T \cdot V \right)_{ij}}{\left((W^k)^T \cdot W^k \cdot H^k \right)_{ij}} \quad (\text{公式24})$$

$$W_{ij}^{k+1} = W_{ij}^k \frac{\left(V \cdot (W^k)^T \right)_{ij}}{\left(W^k \cdot H^k \cdot (H^k)^T \right)_{ij}} \quad (\text{公式25})$$

设定迭代数为 k 。采用 Python 语言的 Pytorch 库构建, 使用 TensorLy 库进行张量分解计算。

1.6.2 基因集富分析 通过多样性可视化集成数据库 (<https://david.ncifcrf.gov/home.jsp>, DAVID 版本 6.8) 进行基因本体论 (Gene Ontology, GO) 和京都基因与基因组百科全书 (Kyoto Encyclopedia of Genes and Genomes, KEGG) 通路富集分析, 进一步解析药物作用的生物相关信号通路。

2 结果

2.1 HCVAE 模型构建与训练

2.1.1 模型构建 本研究设计构建2种VAE模型: 标准VAE模型与HCVAE模型。标准VAE模型使用了四层编码器网络 (256、128、64、32) 和四层解码器网络 (32、64、128、256)。HCVAE模型使用了2个四层编码器网络 (256、128、64、32) 和1个四层解码器网络 (32、64、128、256)。为了便于可视化, 随机层的维数 z 通常设置为2, 与其他

潜在空间进行比较时保持一致性。

2.1.2 模型训练 VAE 和 HCVAE 这 2 个模型在数据集上均进行循环迭代, 在整个训练过程中, 模型的目标函数均实现收敛。标准 VAE 模型经 500 次循环 (13.5 万轮 epochs) 迭代训练后, 在训练集上, VAE 的目标函数值收敛至 1330.8416 ($\log_likelihood$: $-1\,322.523\,3$, KL 散度: 8.318 228), 表明模型在隐空间正则化与数据重建间取得平衡。模型得到有效训练, 且能够在给定数据集上进行良好的拟合。HCVAE

模型经 500 次循环 (13.5 万 epochs) 迭代训练后, 损失函数 [$Loss = -(\log_likelihood - KL_divergence)$] 逐渐收敛并趋于稳定。在训练集上, HCVAE 的目标函数值收敛至 2 273.901 ($\log_likelihood$: $-2\,253.3608$, KL 散度: 20.539 978), 表明模型在隐空间正则化与数据重建间取得平衡。模型训练后, 药物作用下的细胞基因表达特征 (扰动后的基因表达数据) 在潜空间中的表示分布逐渐聚类, 模型能够有效捕捉到药物对细胞基因表达特征 (见图 2)。

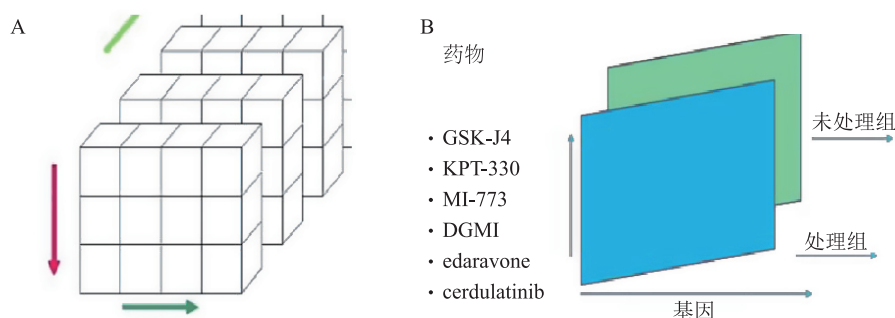


图1 药物-细胞-基因表达张量数据结构

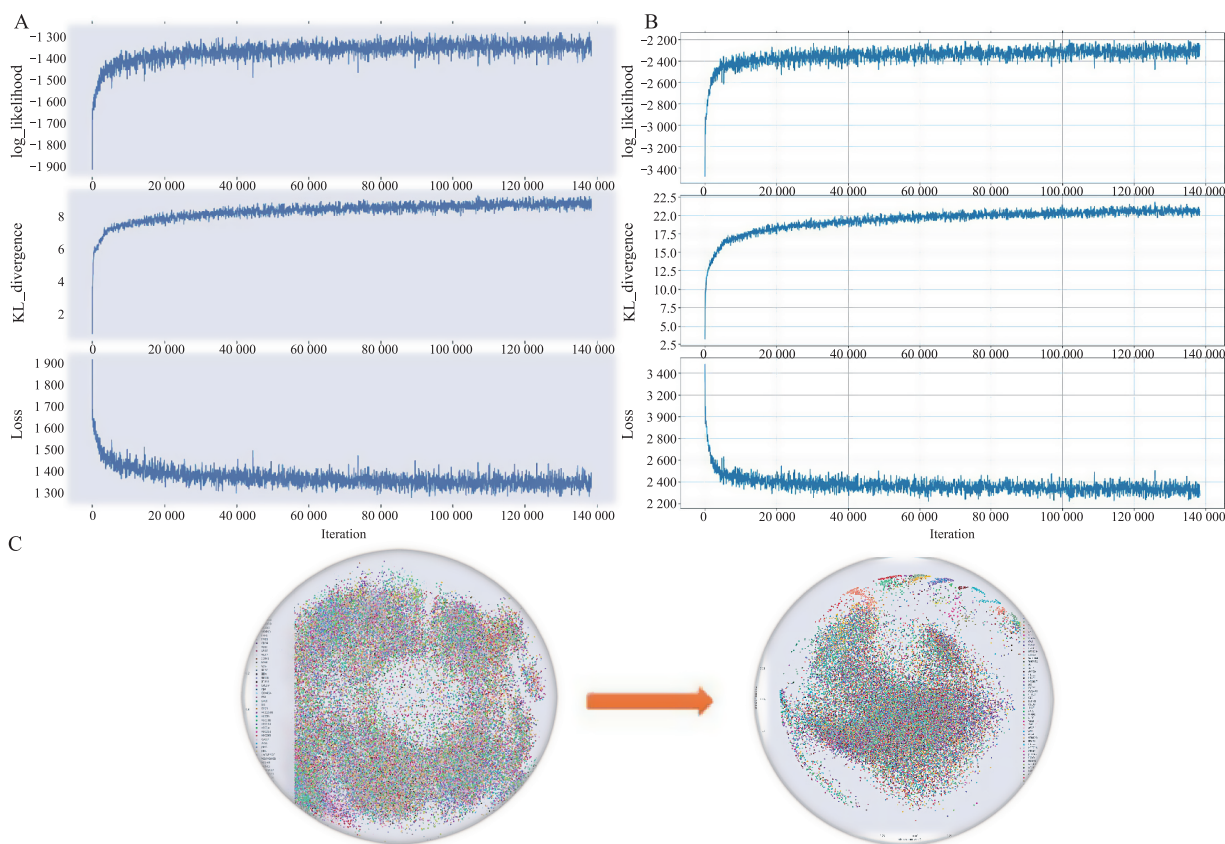


图2 模型训练过程目标函数变化及潜在空间聚类可视化

2.2 HCVAE 模型性能评估

本研究使用内部指标 (ASW、CH、DB) 和外部指标 (NMI、ARI), 对基于药物扰动后细胞基因表达数据集训练所得模型的推断性能进行系统评估。

评估内容包括 VAE 和 HCVAE 模型在训练过程中及训练后聚类性能差异。训练结果显示, 所有聚类指标 [ARI、NMI、ASW、CH、DB、Pearson Correlation (PC)] 在模型迭代后均有效收敛, 表明

模型能够稳定地优化聚类性能(见图3)。

(ARI、NMI、ASW、CH、DB、PC)在模型迭代后

HCVAE 模型训练结果显示,所有聚类指标 均有效收敛(见图4)。

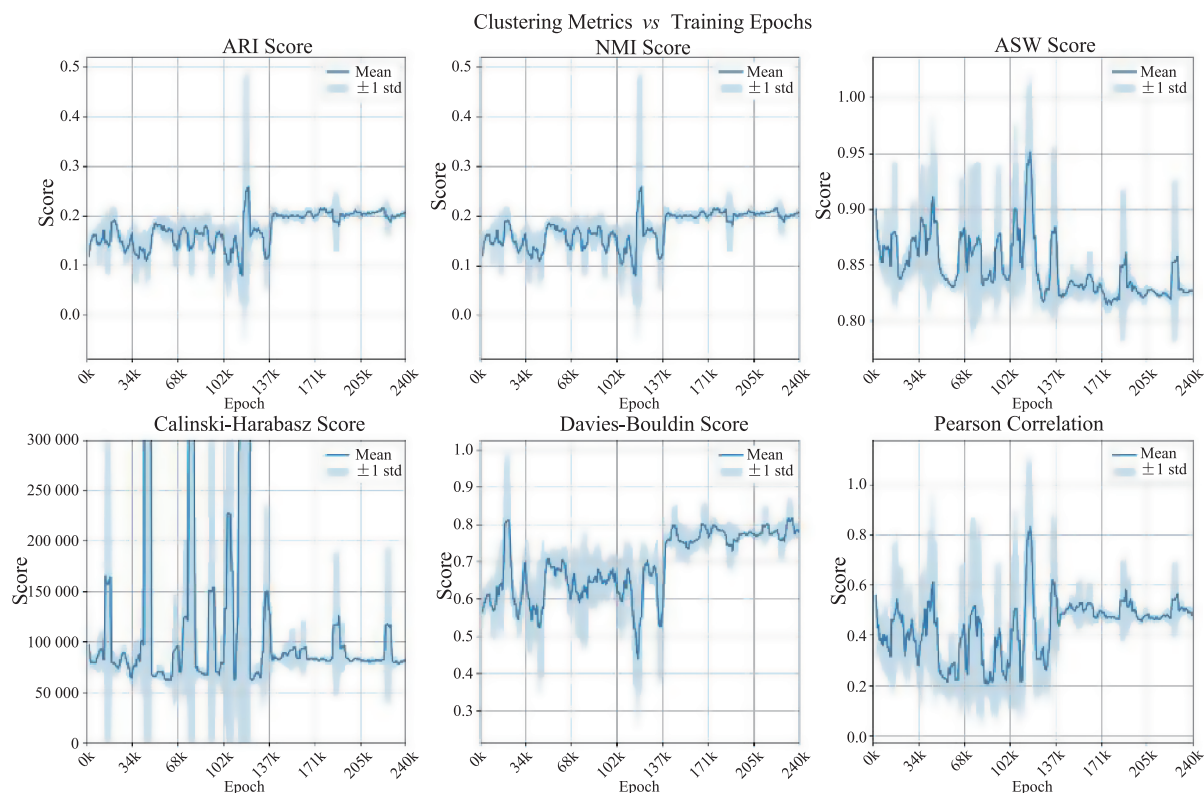


图3 标准VAE架构模型训练过程的性能变化

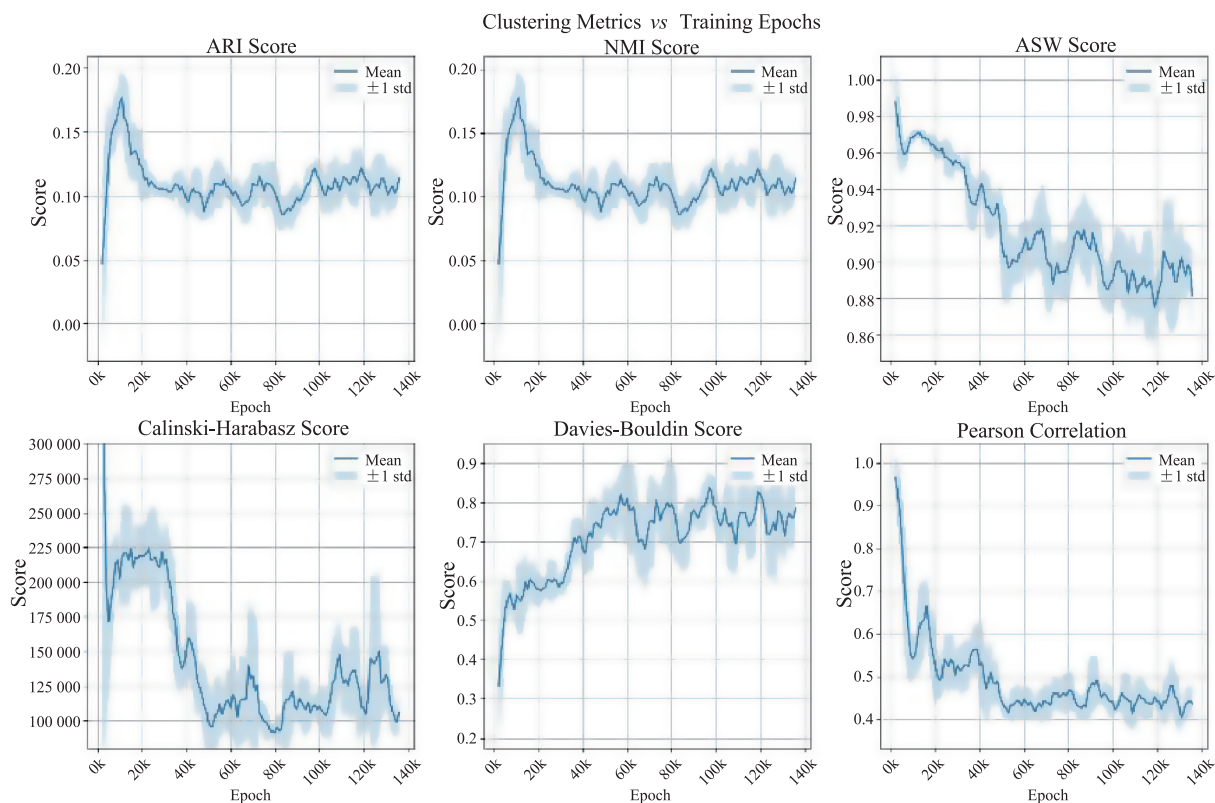


图4 HCVAE架构模型训练过程的性能变化

本研究计算了多次训练 ($n=6$) 后各模型的外部 (NMI、ARI)、内部 (ASW、CH、DB) 聚类指标和潜在变量相关性 (PC)。对于满足正态性假设的指标 (ARI、NMI、ASW、CH 和 PC), 采用配对 t 检验评估不同模型间的统计显著性差异。对于 DB 指标, 首先进行正态性检验 (Shapiro-Wilk 检验), 标准 VAE 的 DB 指标正态性检验结果 $p=0.91$, HCVAE 的 DB 指标正态性检验结果 $p=0.61$, 均大于 0.05, 表明数据服从近似正态分布。随后进行方差齐性检验

(Levene 检验) 结果显示, $p=0.12$, 表明方差齐性假设成立。因此, DB 指标同样采用 t 检验进行比较。

标准 VAE 模型各聚类指标为 ARI: 0.213 ± 0.009 , NMI: 0.214 ± 0.009 , ASW: 0.821 ± 0.004 , CH: $78\,661.512 \pm 2\,682.700$, DB: 0.785 ± 0.009 , PC: 0.466 ± 0.013 。HCVAE 模型各聚类指标为 ARI: 0.112 ± 0.013 , NMI: 0.113 ± 0.013 , ASW: 0.881 ± 0.025 , CH: $104\,985.774 \pm 14\,553.550$, DB: 0.762 ± 0.042 , PC: 0.417 ± 0.021 (见表 1)。

表 1 标准 VAE 和 HCVAE 模型性能评价指标统计检验结果

指标	VAE ($\bar{x} \pm s$)	HCVAE ($\bar{x} \pm s$)	统计量	P	显著性水平
ARI	0.213 ± 0.009	0.112 ± 0.013	36.123	< 0.001	***
NMI	0.214 ± 0.009	0.113 ± 0.013	36.162	< 0.001	***
ASW	0.821 ± 0.004	0.881 ± 0.025	-12.772	< 0.001	***
CH	$78\,661.512 \pm 2\,682.700$	$104\,985.774 \pm 14\,553.550$	-9.743	< 0.001	***
DB	0.785 ± 0.009	0.762 ± 0.042	2.943	0.005	**
PC	0.466 ± 0.013	0.417 ± 0.021	10.932	< 0.001	***

ARI、NMI、ASW、CH、DB 和 PC 等指标均满足正态性假设, 采用配对 t 检验评估统计显著性; HCVAE 模型与 VAE 模型比较, *** $P < 0.001$, ** $P < 0.005$, * $P < 0.01$

由于训练数据集为无标签数据集 (即药物、细胞类型等标签数据未输入模型进行训练), 本研究在评估模型的训练效果时, 更加关注聚类评价的内部指标分析。其中, ASW 值大于 0.80 表示样本大多位于正确簇内, 且簇间分离明显; CH 值大于 1 000 则表明簇间离散度远高于簇内离散度, 聚类结构清晰; DB 值小于 0.8 则表明簇内离散度与簇间分离度

的比值较小, 聚类效果较好; PC 值小于 0.45 则表明特征间存在微弱的正相关性 (见图 5)。

比较 2 个模型的主要指标显示: 相较于 VAE 模型, HCVAE 模型在训练数据集的各个聚类指标 (ASW、CH、DB 和 PC) 上均表现出显著提升。这说明 HCVAE 模型在聚类效果上明显优于 VAE, 模型特征提取性能得到进一步增强。

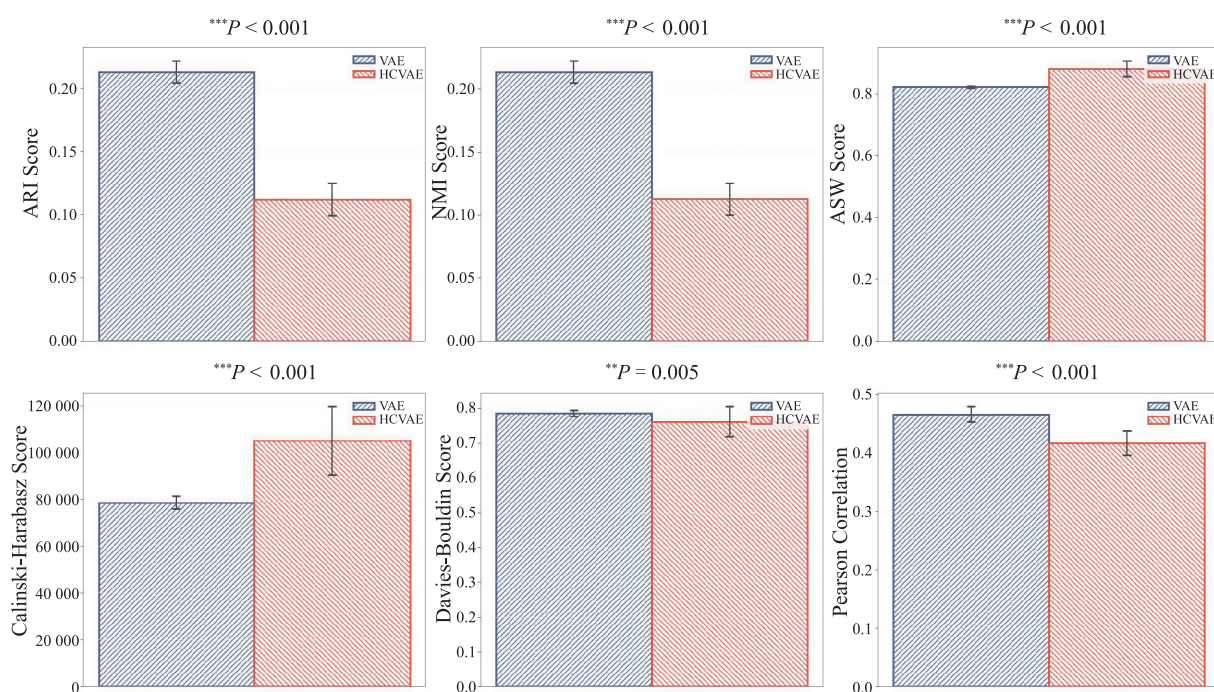


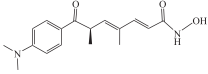
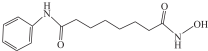
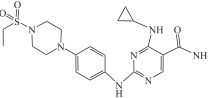
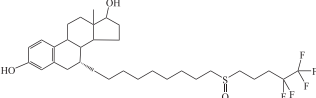
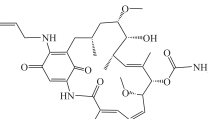
图 5 经过训练后标准 VAE 和 HCVAE 模型的性能比较

2.3 HCVAE 模型推断药物相似性

根据前述空间定位计算, 首先将阳性对照药物(依达拉奉)的细胞扰动基因表达数据输入模型编码器进行推断, 获得其作用特征在隐空间的位置表示。随后, 计算数据集中各药物对诱导多能干细胞分化的运动神经元(iPSCs-derived motor neurons)细胞扰动的基因表达在隐空间的位置表示与依达拉

奉的细胞扰动的基因表达在潜在空间位置的距离(distance scores, d -scores), 并按照距离从小到大, 将被测药物进行排序。 d -scores 越小, 表明药物的药效与依达拉奉的作用越相似。结果显示, 赛度替尼、曲古抑菌素 A、伏立诺他等药物与依达拉奉的 d -scores 接近, 提示上述药物与依达拉奉具有相似的药学效应(见表 2)。

表 2 经模型推断与依达拉奉作用相似的化合物

化合物	结构式	生物活性剂	相似性 (d -scores)
曲古抑菌素 (trichostatin)		HDAC 抑制剂 ^[28-29]	0.081
伏立诺他 (vorinostat)		HDAC 抑制剂 ^[30-31]	0.101
赛度替尼 (cerdulatinib)		SYK/JAK、caspase-3 ^[32]	0.172
氟维司群 (Faslodex)		雌激素受体拮抗剂	0.788
坦螺旋霉素 (tanespimycin)		HSP90- α 抑制剂	1.056

HDAC: 组蛋白去乙酰化酶; SYK: 酪氨酸激酶; JAK: Janus 激酶; HSP90: 热休克蛋白 90

2.4 生物信息学分析

张量分解算法分析结果显示, 赛度替尼与依达拉奉在药物效应上存在较高相关性(在共同因子 4 上具有相似权重)。因此进一步推断, 赛度替尼与依达拉奉具有类似的作用效应, 该结论与 HCVAE 模型推断结果吻合。

根据张量分解获取的特征基因, 按照其权重排序依次为 *JAK1*、*IRF9*、*TRAF2*、*STAT1*、*IFNAR2*、*IFNAR1*、*RIPK1*、*FADD* 等, 其中多个基因与细胞凋亡过程相关(见图 6)。受体相互作用丝氨酸/苏氨酸激酶 1 (receptor interacting serine/threonine kinase 1, RIPK1) 是一种衔接蛋白和丝氨酸/苏氨酸激酶, 在细胞凋亡、自噬及炎症反应中起重要作用。在肿瘤坏死因子(tumor necrosis factor, TNF)受体信号传导通路中, RIPK1 能够激活下游的 caspase 依赖性和非依赖性通路。Fas 相关死亡结构域蛋白(Fas-associated death domain protein, FADD)是细胞凋亡过程中的关键衔接蛋白, 能够与死亡受体(如 Fas)相互作用, 并招募 pro-caspase 形成死亡诱导信号复合物(death-inducing signaling complex, DISC), 进

而触发细胞凋亡。

GO 和 KEGG 通路富集分析表明, 赛度替尼的作用靶点与炎症反应、细胞凋亡、酶活性及缺氧相关信号通路密切相关, 这与既往研究中 caspase-3 在缺血性脑卒中中的作用报道相符^[33]。结合张量分析结果进一步验证, 赛度替尼的表型效应与其对细胞凋亡过程中的 RIPK1、caspase 关键蛋白的调控密切相关(见图 7)。

3 讨论

本研究通过整合 VAE、对比学习与双曲潜在空间特征映射, 构建 HCVAE 模型, 有效提取“药物-疾病”作用表型效应的高维特征, 为脑卒中药物重定位提供新策略。以依达拉奉为阳性对照药物, 根据 HCVAE 模型推断, 筛选多个治疗缺血性脑卒中的候选药物。进一步通过生物信息学分析发现, 赛度替尼可通过调控细胞凋亡通路中的关键蛋白发挥治疗缺血性脑卒中的作用, 为阐明赛度替尼多靶点作用机制提供了重要的分子基础与理论支持。

单细胞 RNA-seq 数据分析的核心挑战在于其隐

含的层次结构(如细胞分化、疾病进展)。双曲空间具有指数膨胀特性,其空间距离与层次深度呈正相关,从而适用于匹配分化进展等生物学过程。针对 RNA-seq 数据特点,本研究在建模过程中将潜在空间嵌入双曲空间,并通过洛伦兹距离将生物学层次

转化为概率衰减过程。相较于仅反映“表达谱相似性”的欧氏距离(平直几何),洛伦兹距离在信息承载方面具有优势,不仅能量化生物学层次深度,还能有效区分“表达谱相似但处于不同分化阶段”的细胞类型。

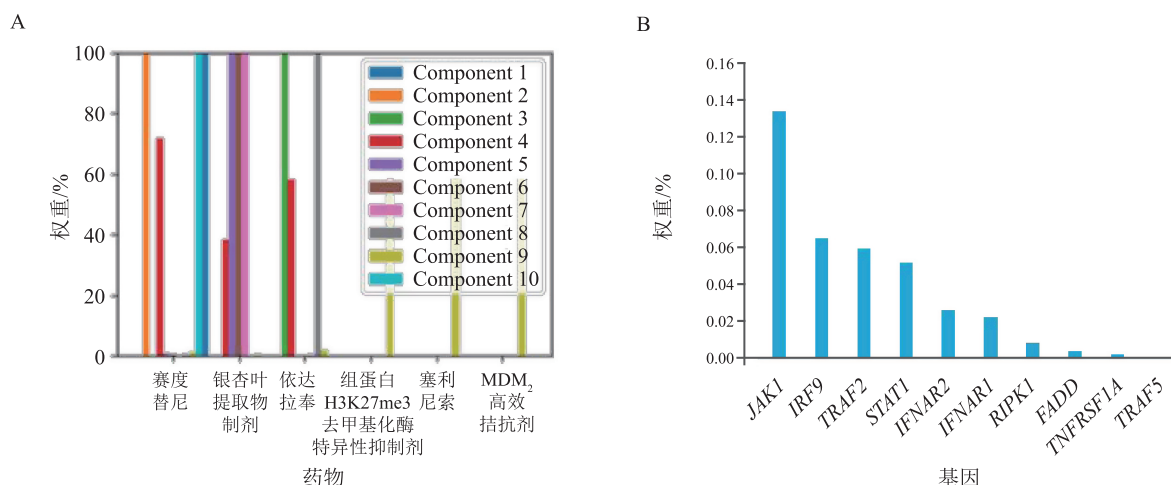


图6 张量分解分析药物作用相关性

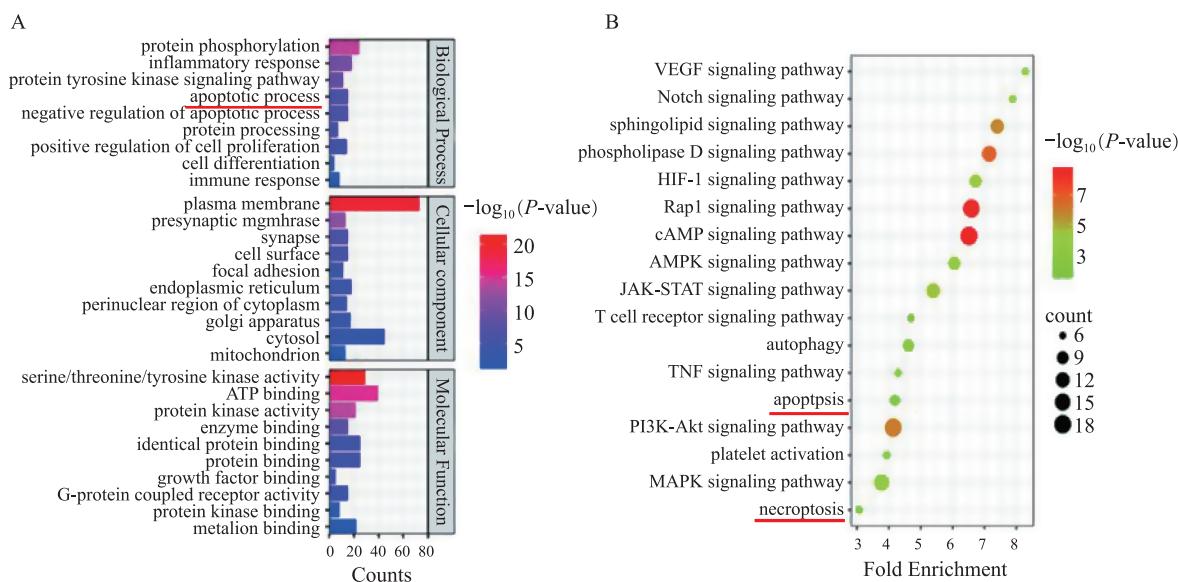


图7 赛度替尼潜在作用靶点基因富集分析

本研究开发的 HCVAE 模型,是一种融合 VAE 与对比学习机制的深度学习模型,旨在从高维单细胞转录组数据中学习具有判别性的低维表征。该模型通过显式约束潜在表示空间的结构,使生物学意义相近的样本在表示空间中彼此靠近,而异质样本则相互远离,实现了良好的类别区分能力。

通过同时输入处理组与对照组数据,并以对照组作为基线,HCVAE 模型能够有效识别由药物扰动诱导的特异性基因表达变化,显著提升了对技术

噪声与批次效应的鲁棒性,增强了潜在空间的结构可解释性,从而在复杂生物医学数据中表现出优异的泛化能力。本研究在 HCVAE 模型中引入了双曲几何空间作为潜在表示空间,进一步提升对生物学系统中层次化、序列化结构的表征能力。与欧几里得空间相比,双曲空间在单位半径内可容纳的节点数量随半径呈指数级增长,具备更优异的信息承载能力,特别适用于刻画细胞类型分化、药物应答通路等具有自然层次结构的生物过程。通过将潜在表

示映射至双曲空间,使模型能够更精细地捕捉细胞状态之间的多层次关系,从而为解析药物作用效应提供更富解释性的特征表示。相较于标准 VAE 模型,HCVAE 模型在聚类效果上明显提升,模型对特征的辨识能力进一步增强。

本研究未采用 GAN 人工神经网络进行建模,主要基于以下 3 个方面的原因:1) RNA-seq 数据的高维性和高稀疏性结构(包含大量的“0”)在 GAN 训练过程中易导致模式崩溃问题,在高维稀疏空间中,GAN 生成器往往集中在少数模式,导致模型输出多样性不足,无法覆盖真实数据分布的全部模式^[34];2) GAN 等深度学习模型训练需数万至百万级样本(如 ImageNet 的 140 万图像),而 RNA-seq 数据的样本量通常较小(多为数百至数千个细胞)。由于基因数量庞大而样本量有限,训练过程中易出现过拟合现象^[35]。3) scRNA-seq 数据的多样性要求高度定制化的解决方案,而现有 GAN 模型通常结构复杂、计算资源消耗大,限制了其应用^[36]。

本研究也未采用 graphVAE 人工神经网络进行建模。尽管 graphVAE 专为图数据设计,可高效捕获节点与边的依赖关系,但使用该模型需先构建图结构(如基因调控网络)。不同于图数据,标准 RNA-seq 数据是高维、稀疏、非负连续型矩阵(基因 $\times$ 样本),不适用于 graphVAE 的图数据结构。若强行将 RNA-seq 数据适用于图神经网络结构(如基于 PPI 网络),需先定义节点与边的对应关系,但 RNA-seq 数据本身并不提供图信息。

本研究选用 VAE 人工神经网络模型,可有效克服 GAN 等其他生成模型存在的问题,且与目标建模数据(RNA-seq)类型更匹配。VAE 模型的优势包括以下 4 个方面:1)天然适用于非负数据,解码器用激活函数计算输出结果 ≥ 0 ;2)降维与聚类效果好,能学习低维潜在空间,有效区分细胞类型;3)数据增强实用性高,VAE 解码器生成的样本保留基因表达的统计特性(如负二项分布),可直接用

于分类器训练,无需构建大规模训练集,更适用于训练样本量小的场景;4)稀疏数据处理能力突出, RNA-seq 数据通常包含大量“0”值,数据高度稀疏,VAE 模型使用变分正则化能够抑制过拟合。因此,在标准 RNA-seq 数据(基因表达矩阵)分析中,VAE 模型是兼具科学性和合理性的优选方案。

相较于 CMap (Connectivity Map) 的“基因表达谱反转”策略,本研究将药物重定位从“特征匹配”升级为“机制可解释的表型相似性深度学习”,通过最小化药物-疾病表型在潜在空间中的距离,直接优化重定位目标。此外,CMap 依赖预定义数据集(如 L1000),其泛化能力有限。本研究通过自建模型整合 GSE180601 等单细胞数据,方法更灵活且特异性强。

基于 HCVAE 模型,本研究选取缺血性脑卒中临床治疗药物依达拉奉为阳性对照药物,对大规模药物扰动转录组数据进行药效相似性计算,成功筛选出与依达拉奉具有高度药效相似性的候选药物即赛度替尼。进一步通过药物作用张量分析、生物学信息学分析等方法,验证了赛度替尼与依达拉奉表型效应的相似性,两者均与缺血性脑卒中中细胞死亡调控机制高度相关,提示赛度替尼具备跨适应证治疗缺血性脑卒中的潜力,极大提升了预测结果的可靠性及生物学合理性。值得注意的是,卒中(脑卒中)的病理过程是多细胞协同作用的结果,仅用依达拉奉在单一神经元细胞系(如 SH-SY5Y)中进行测试,可能无法全面反映卒中复杂的病理机制,存在增加候选药物临床转化失败的风险。未来研究需构建多细胞(神经元细胞+胶质细胞+内皮细胞)协同模型,以进一步阐明卒中药物筛选的生物学基准。

综上,HCVAE 模型不仅显著提升了药物表征学习的性能,更为缺血性脑卒中治疗药物重定位提供了高效、稳健且可解释的计算框架,有望作为平台工具筛选出更多具有潜在治疗优势的候选药物,为缺血性脑卒中相关药物的高效筛选提供新策略。

【参考文献】

- [1] Eltzschig H K, Eckle T. Ischemia and reperfusion: from mechanism to translation[J]. *Nat Med*, 2011, 17(11): 1391-1401.
- [2] Qin C, Yang S, Chu Y H, et al. Signaling pathways involved in ischemic stroke: molecular mechanisms and therapeutic interventions[J]. *Signal Transduct Target Ther*, 2022, 7 (1): 215.
- [3] Tian J, Guo S, Chen H, et al. Combination of emricasan with ponatinib synergistically reduces ischemia/reperfusion injury in rat brain through simultaneous prevention of apoptosis and necroptosis[J]. *Transl Stroke Res*, 2018, 9(4): 382-392.
- [4] Peng Z M, Zhang Y Y, Wei D, et al. MALT1 promotes necroptosis in stroke rat brain via targeting the A20/RIPK3 pathway[J]. *Arch Biochem Biophys*, 2023, 735: 109502.
- [5] Zhang Y Y, Liu W N, Li Y Q, et al. Ligustroflavone reduces necroptosis in rat brain after ischemic stroke through targeting RIPK1/RIPK3/MLKL pathway[J]. *Naunyn Schmiedeberg's Arch Pharmacol*, 2019, 392(9): 1085-1095.
- [6] Tian J, Zhang Y Y, Peng Y W, et al. Polymyxin B reduces brain injury in ischemic stroke rat through a mechanism involving targeting ESCRT-III machinery and RIPK1/RIPK3/MLKL pathway[J]. *J Cardiovasc Transl Res*, 2022, 15(5): 1129-1142.
- [7] Davaanyam D, Lee H, Seol S I, et al. HMGB1 induces hepcidin upregulation in astrocytes and causes an acute iron

- surge and subsequent ferroptosis in the postischemic brain[J]. *Exp Mol Med*, 2023, 55(11): 2402–2416.
- [8] Peng J, He J, Lin L, *et al.* Neural stem cell extracellular vesicles carrying YBX1 inhibited neuronal pyroptosis through increasing m6A-modified GPR30 stability and expression in ischemic stroke[J]. *Transl Stroke Res*, 2025, 16(2): 262–279.
- [9] Duan Q, Reid S P, Clark N R, *et al.* L1000CDS2: LINCS L1000 characteristic direction signatures search engine[J]. *NPJ Syst Biol Appl*, 2016, 2: 16015.
- [10] Butler A, Hoffman P, Smibert P, *et al.* Integrating single-cell transcriptomic data across different conditions, technologies, and species[J]. *Nat Biotechnol*, 2018, 36(5): 411–420.
- [11] Kiselev V Y, Andrews T S, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data[J]. *Nat Rev Genet*, 2019, 20(5): 273–282.
- [12] Svensson V, Vento-Tormo R, Teichmann S A. Exponential scaling of single-cell RNA-seq in the past decade[J]. *Nat Protoc*, 2018, 13(4): 599–604.
- [13] Eraslan G, Simon L M, Mircea M, *et al.* Single-cell RNA-seq denoising using a deep count autoencoder[J]. *Nat Commun*, 2019, 10: 390.
- [14] Lopez R, Regier J, Cole M B, *et al.* Deep generative modeling for single-cell transcriptomics[J]. *Nat Methods*, 2018, 15(12): 1053–1058.
- [15] Kwon O S, Lee H, Kong H J, *et al.* Connectivity map-based drug repositioning of bortezomib to reverse the metastatic effect of GALNT14 in lung cancer[J]. *Oncogene*, 2020, 39(23): 4567–4580.
- [16] Keenan A B, Wojciechowski M L, Wang Z, *et al.* Connectivity mapping: methods and applications[J]. *Annu Rev Biomed Data Sci*, 2019, 2: 69–92.
- [17] Zhao Y, Chen X, Chen J, *et al.* Decoding connectivity map-based drug repurposing for oncotherapy[J]. *Brief Bioinform*, 2023, 24(3): bbad142.
- [18] Subramanian A, Narayan R, Corsello S M, *et al.* A next generation connectivity map: L1000 platform and the first 1, 000, 000 profiles[J]. *Cell*, 2017, 171(6): 1437–1452.e17.
- [19] Lin K, Li L, Dai Y, *et al.* A comprehensive evaluation of connectivity methods for L1000 data[J]. *Brief Bioinform*, 2020, 21(6): 2194–2205.
- [20] Peng J J, Zhang Y Y, Li R F, *et al.* Hybrid approach for drug-target interaction predictions in ischemic stroke models[J]. *Artif Intell Med*, 2025, 161: 103067.
- [21] Kingma D P, Welling M. Auto-encoding variational bayes[EB/OL]. (2022-11-10) [2025-09-07]. <https://arxiv.org/abs/1312.6114>.
- [22] Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives[J]. *IEEE Trans Pattern Anal Mach Intell*, 2013, 35(8): 1798–1828.
- [23] Weinberger E, Lin C, Lee S I. Isolating salient variations of interest in single-cell transcriptomic data with contrastiveVI[J]. *Nat Methods*, 2023, 20(9): 1336–1345.
- [24] Yang M, Yang Y, Xie C, *et al.* Contrastive learning enables rapid mapping to multimodal single-cell atlas of multimillion scale[J]. *Nat Mach Intell*, 2022, 4(8): 696–709.
- [25] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[EB/OL]. (2015-05-09) [2025-09-07]. <https://arxiv.org/abs/1503.02531>.
- [26] van den Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding[EB/OL]. (2019-01-22) [2025-09-07]. <https://arxiv.org/abs/1807.03748>.
- [27] Duan Q, Reid S, Clark N, *et al.* L1000CDS2: LINCS L1000 characteristic direction signatures search engine[EB/OL]. (2025-03-31) [2025-09-07]. <http://amp.pharm.mssm.edu/static/hdfs/l1000cds2/cpcd-gse70138.bson>.
- [28] Schnekenburger M, Dicato M, Diederich M F. Anticancer potential of naturally occurring immunoepigenetic modulators: a promising avenue?[J]. *Cancer*, 2019, 125(10): 1612–1628.
- [29] Wang C, Huang M, Lin Y, *et al.* ENO₂⁻ derived phosphoenolpyruvate functions as an endogenous inhibitor of HDAC1 and confers resistance to antiangiogenic therapy[J]. *Nat Metab*, 2023, 5(10): 1765–1786.
- [30] Cozzolino F, Iacobucci I, Monaco V, *et al.* Lysines acetylome and methylome profiling of H3 and H4 histones in trichostatin A: treated stem cells[J]. *Int J Mol Sci*, 2021, 22(4): 2063.
- [31] Lu Z P, Ju Z L, Shi G Y, *et al.* Histone deacetylase inhibitor trichostatin A reduces anti-DNA autoantibody production and represses IgH gene transcription[J]. *Biochem Biophys Res Commun*, 2005, 330(1): 204–209.
- [32] Ma J, Xing W, Coffey G, *et al.* Cerdulatinib, a novel dual SYK/JAK kinase inhibitor, has broad anti-tumor activity in both ABC and GCB types of diffuse large B cell lymphoma[J]. *Oncotarget*, 2015, 6(41): 43881–43896.
- [33] Porter A G, Jänicke R U. Emerging roles of caspase-3 in apoptosis[J]. *Cell Death Differ*, 1999, 6(2): 99–104.
- [34] Ai X, Smith M C, Feltus F A. Generative adversarial networks applied to gene expression analysis: an interdisciplinary perspective[J]. *Comput Syst Oncol*, 2023, 3(3): e1050.
- [35] Kim J, Seok J. ctGAN: combined transformation of gene expression and survival data with generative adversarial network[J]. *Brief Bioinform*, 2024, 25(4): bbae325.
- [36] Lee M. Recent advances in generative adversarial networks for gene expression data: a comprehensive review[J]. *Mathematics*, 2023, 11(14): 3055.



【专家介绍】彭军：中南大学二级教授，“升华学者计划”特聘教授，博士生导师。中国药理学学会理事、国际动脉硬化学会中国分会理事、中国药理学学会心血管专委会常务委员、中国药理学学会中药与天然药物专委会常务委员、湖南省药理学学会副理事长兼秘书长等；《中国动脉硬化杂志》常务编委、《中国药理学与毒理学杂志》《中药药理与临床》和 *CNS Neurol Disord-DR* 等期刊编委。主要从事心脑血管疾病发生机制、心脏药物毒性机制及药物防治研究。已在 *Circulation*, *PNAS*, *Circ Res*, *Free Radical Biol Med* 等国际权威和重要期刊发表SCI收录科研论文190余篇；获授权发明专利11项，已成功转化1项；主持国家自然科学基金重大项目（1项）、面上项目（5项），美国心脏学会博士后基金（1项），全国优博专项基金（1项）等10余项；获得全国优秀博士学位论文奖、湖南省青年科技奖、中华医学科技二等奖、湖南省科学技术进步二等奖、湖南省自然科学二等奖、湖南省医学科技一等奖等10余项奖励；入选2022—2024年度全球2%顶尖科学家榜单；获2024年第四季度威立Wiley中国高贡献作者奖。