第 41 卷 第 3 期
2026 年 6 月

天津科技大学学报
Journal of Tianjin University of Science and Technology

Vol. 41 No. 3
Jun. 2026

DOI:10.13364/j.issn.1672-6510.20240222
网络首发日期：2025–04–30；网络首发地址：http://link.cnki.net/urlid/12.1355.n.20250430.1116.002

基于迁移学习框架的单细胞药物反应预测

李玉田，葛凤雅，王　林
（天津科技大学人工智能学院，天津 300457）

摘　要：肿瘤组织中存在多种类型的细胞，这些细胞在形态、功能和生物学特性上存在差异，例如不同细胞类型对同一药物的敏感性不同，而这也是导致癌症治疗复发率高和效力低的重要原因。单细胞 RNA 测序技术是针对这一细胞类型异质性的重要研究工具，本文提出一种基于变分自动编码器的迁移学习框架，用于单细胞药物反应预测。该框架整合了混合测序的细胞系药物基因组数据和单细胞 RNA 测序数据，使用最大均值差异和最优传输，在低维潜在空间中对细胞系和单细胞进行特征对齐，从而在单细胞水平上准确预测药物反应。将模型与最新的单细胞药物反应预测方法进行基准比较，在 8 个数据集上进行训练和验证，模型取得了较高的预测性能。消融实验和鲁棒性实验验证了模型各模块的有效性以及模型的稳定性。本研究提供了新的迁移学习思路，为个性化和精确的癌症治疗提供一种新策略，有助于精准医学的发展。

关键词：迁移学习；特征对齐；单细胞；药物反应预测
中图分类号：TP3-05　　**文献标志码**：A　　**文章编号**：1672-6510(2026)03-0018-09

Single-Cell Drug Response Prediction Based on Transfer Learning Framework

LI Yutian，GE Fengya，WANG Lin
(College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin 300457, China)

Abstract: Tumor tissues contain diverse cell types that differ in morphology, function and biological characteristics. Notably, different cell types have different sensitivity to the same drug, which is also an important reason for the high recurrence rate and low efficacy of cancer treatment. Single-cell RNA sequencing technology is an important research tool for this cellular heterogeneity. This article proposes a transfer learning framework based on variational auto-encoder for single-cell drug response prediction. The framework integrates bulk-sequenced cell line pharmacogenomic data with single-cell RNA sequencing data, and adopts maximum mean discrepancy and optimal transport to align cell lines and single cells in a low-dimensional latent space, thereby enabling accurate prediction of drug response at the single-cell level. The proposed model was benchmarked against state-of-the-art methods for single-cell drug response prediction and validated on eight data sets, achieving superior predictive performance. Ablation and robustness experiments further verified the effectiveness of each module of the model and the overall stability of the model. This study provides a novel transfer learning approach, offering a new strategy for personalized and precise cancer treatment and contributing to the development of precision medicine.

Key words: transfer learning; feature alignment; single cell; drug response prediction

引文格式：
李玉田，葛凤雅，王林. 基于迁移学习框架的单细胞药物反应预测[J]. 天津科技大学学报，2026，41(3)：18–26.
LI Y T, GE F Y, WANG L. Single-cell drug response prediction based on transfer learning framework[J]. Journal of Tianjin university of science and technology, 2026, 41(3): 18–26.

收稿日期：2024–11–06；**修回日期**：2025–01–12
基金项目：天津市企业科技特派员项目(20YDTPJC00560)；天津科技大学校级大学生创新创业训练计划资助项目(202410057133)
作者简介：李玉田(2000—)，男，河南人，硕士研究生；**通信作者**：王　林，副教授，linwang@tust.edu.cn

精准医学在揭示癌症基因组复杂性方面取得了显著进展[1]。随着基因测序技术的快速发展，对癌症的治疗研究越来越全面，混合细胞系测序技术可以帮助研究人员在组织水平上研究癌症对药物的反应，这有力促进了精准医疗的发展。然而，由于混合细胞系测序数据中不同细胞状态之间的异质性，癌症药物治疗常常面临复发率高和效力低的问题。肿瘤内异质性是癌症治疗失败的重要原因[2]，这种异质性导致个别细胞对药物的反应不同，最终导致残留少量癌细胞并引起癌症复发[3]。近年来，单细胞 RNA 测序技术得到了快速发展。单细胞 RNA 测序技术可以在细胞水平上研究癌细胞异质性，使研究者能够以更高的分辨率分析肿瘤内异质性[4]。但是，由于目前单细胞 RNA 测序技术的成本较高以及数据量较少，使用单细胞 RNA 测序技术支撑肿瘤药物研发仍较为受限。体外药物筛选研究已经为各种癌症细胞系产生了药物反应数据[5-6]，积累了大规模的细胞系混合测序数据。因此，将大规模的细胞系混合 RNA 测序数据与单细胞 RNA 测序数据整合起来，可以充分利用两者的优势。

迁移学习是将一个数据域（源域）中获得的知识转移到另一个数据域（目标域），以改善目标域的预测性能。在此背景下，迁移学习可以用于将细胞系混合 RNA 测序数据中的药物反应信息映射到单细胞 RNA 测序数据，从而实现单细胞水平的药物反应预测。一些研究方法已经展示了迁移学习在单细胞药物反应预测研究中的潜力[1,7]。本研究提出基于迁移学习的单细胞药物反应预测框架（single-cell drug response prediction based on transfer learning，SCTL），该框架基于最优传输和最大均值差异，在细胞系测序数据和单细胞测序数据之间构建迁移桥梁，以期在单细胞水平上实现准确的药物反应预测。

1　数据介绍

共收集整理 8 个数据集。使用来自公开数据库（Genomics of Drug Sensitivity in Cancer，GDSC）[8]和（Cancer Cell Line Encyclopedia，CCLE）[9]的混合细胞系测序数据作为源域数据。GDSC 数据库的在线访问途径为 https://www.cancerrxgene.org/。CCLE 数据库的在线访问途径为 https://sites.broadinstitute.org/ccle。数据 1—5 的源域数据收集了 1 280 个细胞系，其中有 804 个细胞系来自 GDSC 数据库，剩下的来自 CCLE 数据库。数据 6—8 的源域数据收集了 1 001 个细胞系，全部来自 GDSC 数据库。目标域数据使用了来自口腔鳞状细胞癌、肺癌、前列腺癌、急性髓性白血病、头颈部鳞状细胞癌的单细胞测序数据，数据 1—5 目标域数据的 GEO 登记号在表 2 中，数据 6—8 的目标域数据来自 Zheng 等[7]的研究。详细的源域数据见表 1，目标域数据见表 2。

表 1　细胞系混合 RNA 测序数据（源域数据）

Tab. 1　Bulk RNA-seq data of cell lines（source data）

序号	药物	耐药细胞系个数	敏感细胞系个数	基因数
数据 1	Cisplatin	791	489	12 966
数据 2	Gefitinib	799	481	12 499
数据 3	Docetaxel	788	492	13 476
数据 4	I.BET.762	724	556	9 566
数据 5	Bortezomib	770	510	12 715
数据 6	Gefitinib	714	115	10 610
数据 7	Vorinostat	774	60	10 610
数据 8	AR-42	811	80	10 610

表 2　单细胞 RNA 测序数据（目标域数据）

Tab. 2　Single-cell RNA-seq data（target data）

序号	药物	耐药细胞个数	敏感细胞个数	基因数	来源
数据 1	Cisplatin	187	361	12 966	GSE117872
数据 2	Gefitinib	470	37	12 499	GSE112274
数据 3	Docetaxel	162	162	13 476	GSE140440
数据 4	I.BET.762	685	734	9 566	GSE110894
数据 5	Bortezomib	4 565	2 875	12 715	GSE114461
数据 6	Gefitinib	33	33	10 610	JHU006
数据 7	Vorinostat	33	33	10 610	JHU006
数据 8	AR-42	33	33	10 610	JHU006

使用 Scanpy 包[10]对单细胞 RNA 测序数据进行预处理，过滤掉检测到的基因少于 200 个的细胞，以及仅在 3 个以下细胞中检测到的基因。仅保留在细胞系和单细胞中共有的基因。最后，使用 MinMaxScaler 函数将源域和目标域数据归一化。

2　计算方法

2.1　迁移学习框架的构建

针对细胞系混合 RNA 测序数据和单细胞 RNA 测序数据之间的差异，构建深度学习网络提取跨域数据的特征。通过特征对齐的计算方法，实现从细胞系水平到单细胞水平的药物反应预测。模型框架如图 1 所示。

对源域数据和目标域数据构建 1 个变分自动编码器[11]，包括用于数据特征提取的编码器和相同层

数的解码器。解码器用于将特征重构回原数据。通过模型的训练迭代，编码器会将源域数据和目标域数据映射到共享的潜在空间中，解码器会在每一轮训练中从潜在空间中重构数据，规范潜在空间数据特征。训练结束后，可以获取在同一潜在空间的源域和目标域数据的特征。

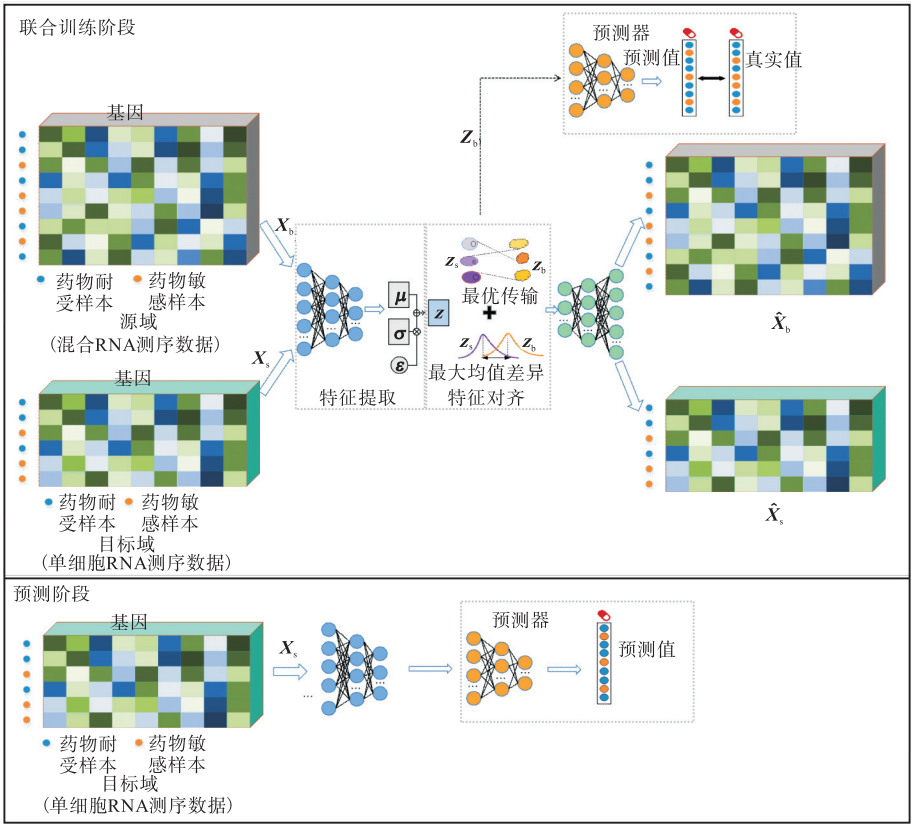


图 1　SCTL 模型框架图

Fig. 1　Framework of SCTL model

在特征对齐的计算方法上，采用多种手段弥合跨域数据的特征，包括最大均值差异[12]的计算和最优传输[11]的方法。

针对单细胞药物反应预测，构建用于药物反应预测二分类任务的多层感知机，其中输出神经元 1 表示敏感，0 表示耐药。在训练药物反应预测器的过程中，使用自动编码器提取的细胞系特征和真实的细胞系药物反应信息训练药物反应预测器。当模型训练结束后，将编码器提取的单细胞潜在空间特征输入药物反应预测器，获得模型对单细胞的药物反应预测值。

2.2　跨域特征的提取

来自源域的细胞系数据是包含 m 个细胞系的混合 RNA 测序数据 $\boldsymbol{X}_{\mathrm{b}}=\left\{\boldsymbol{x}_{\mathrm{b}}^{i}\right\}_{i=1}^{m}$，来自目标域的单细胞数据是包含 n 个单细胞的 RNA 测序数据 $\boldsymbol{X}_{\mathrm{s}}=\left\{\boldsymbol{x}_{\mathrm{s}}^{i}\right\}_{i=1}^{n}$。将两个数据 $\boldsymbol{X}_{\mathrm{b}}$ 和 $\boldsymbol{X}_{\mathrm{s}}$ 纵向拼接作为联合输入 $\boldsymbol{X}=[\boldsymbol{X}_{\mathrm{b}},\boldsymbol{X}_{\mathrm{s}}]$，利用变分自动编码器将其映射到共享潜在空间，学习嵌入特征。构建的变分自动编码器由一个跨域共享的编码器 $q_{\phi}(\boldsymbol{z}|\boldsymbol{x})$，和一个共享的解码器 $p_{\theta}(\boldsymbol{x}|\boldsymbol{z})$ 组成，其中 ϕ 和 θ 分别是编码器和解码器的权重参数。对于数据 $\boldsymbol{x}$，编码器输出均值向量 $\boldsymbol{\mu}_{\boldsymbol{z}|\boldsymbol{x}}$ 以及标准差向量 $\boldsymbol{\sigma}_{\boldsymbol{z}|\boldsymbol{x}}$，潜在向量 $\boldsymbol{z}$ 通过重参数化计算得到，即 $\boldsymbol{z}=\boldsymbol{\mu}_{\boldsymbol{z}|\boldsymbol{x}}+\boldsymbol{\sigma}_{\boldsymbol{z}|\boldsymbol{x}}\odot v$，其中 $\odot$ 表示元素逐个相乘，v 从 $N(0,I)$ 中采样。通过这种方式，可以分别得到细胞系 $\boldsymbol{x}_{\mathrm{b}}^{i}$（$i=1,\cdots,m$）和单细胞 $\boldsymbol{x}_{\mathrm{s}}^{j}$（$j=1,\cdots,n$）的 H 维潜在向量 $\boldsymbol{z}_{\mathrm{b}}^{i}$ 和 $\boldsymbol{z}_{\mathrm{s}}^{j}$。

对于重构部分，基于潜在向量 $\boldsymbol{z}_{\mathrm{b}}^{i}$ 和 $\boldsymbol{z}_{\mathrm{s}}^{j}$ 作为输入，解码器会输出源域重构数据 $\hat{\boldsymbol{x}}_{\mathrm{b}}^{i}$ 以及目标域重构数据 $\hat{\boldsymbol{x}}_{\mathrm{s}}^{j}$。重构的损失函数为

$$L_{\mathrm{rec}}=L_{\mathrm{rec\_b}}+L_{\mathrm{rec\_s}}=\frac{1}{B_{\mathrm{b}}}\sum_{i=1}^{B_{\mathrm{b}}}\mathrm{MSE}(\boldsymbol{x}_{\mathrm{b}}^{i},\hat{\boldsymbol{x}}_{\mathrm{b}}^{i})+\frac{1}{B_{\mathrm{s}}}\sum_{j=1}^{B_{\mathrm{s}}}\mathrm{MSE}(\boldsymbol{x}_{\mathrm{s}}^{j},\hat{\boldsymbol{x}}_{\mathrm{s}}^{j})=\frac{1}{B_{\mathrm{b}}g}\sum_{i=1}^{B_{\mathrm{b}}}\sum_{k=1}^{g}(\boldsymbol{x}_{\mathrm{b}}^{i_k}-\hat{\boldsymbol{x}}_{\mathrm{b}}^{i_k})^2+\frac{1}{B_{\mathrm{s}}g}\sum_{j=1}^{B_{\mathrm{s}}}\sum_{k=1}^{g}(\boldsymbol{x}_{\mathrm{s}}^{j_k}-\hat{\boldsymbol{x}}_{\mathrm{s}}^{j_k})^2 \tag{1}$$

其中：B_b 和 B_s 分别表示数据集 $\boldsymbol{X}_b$ 和 $\boldsymbol{X}_s$ 的批量大小，MSE() 表示均方误差，g 为基因个数。此外，还考虑了 H 维潜在向量 $\boldsymbol{z}$ 的后验分布与多元高斯分布之间的 Kullback-Leibler（KL）散度，表示为

$$D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}\mid\boldsymbol{x})\,\|\,N(0,I))=-\frac{1}{2}\sum_{k=1}^{H}(1+\log((\boldsymbol{\sigma}_{\boldsymbol{z}\mid\boldsymbol{x}}^{k})^2)-(\boldsymbol{\mu}_{\boldsymbol{z}\mid\boldsymbol{x}}^{k})^2-(\boldsymbol{\sigma}_{\boldsymbol{z}\mid\boldsymbol{x}}^{k})^2) \tag{2}$$

最后，结合重构损失和 KL 散度，定义变分自动编码器损失函数为

$$\begin{aligned}L_{\mathrm{vae}}=\lambda_1 L_{\mathrm{rec}}+\lambda_2 L_{\mathrm{KL}}=&\frac{1}{B_b}\sum_{i=1}^{B_b}\mathrm{MSE}(\boldsymbol{x}_b^i,\hat{\boldsymbol{x}}_b^i)+\\&\frac{1}{B_s}\sum_{j=1}^{B_s}\mathrm{MSE}(\boldsymbol{x}_s^j,\hat{\boldsymbol{x}}_s^j)+\frac{1}{B_b+B_s}\cdot\\&\sum_{i=1}^{B_b}D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{z}_b^i\mid\boldsymbol{x}_b^i)\,\|\,N(0,I)\right)+\frac{1}{B_b+B_s}\cdot\\&\sum_{j=1}^{B_s}D_{\mathrm{KL}}\left(q_\theta(\boldsymbol{z}_s^j\mid\boldsymbol{x}_s^j)\,\|\,N(0,I)\right)\end{aligned} \tag{3}$$

其中：$\lambda_1=1$，$\lambda_2=1$。

2.3　特征对齐的计算

对模型提取的源域低维特征 $\boldsymbol{z}_b$ 和目标域低维特征 $\boldsymbol{z}_s$，采用最大均值差异和最优传输的计算方法，在潜在空间上，缩小两种数据域特征分布的距离，实现两种数据域的对齐。

最大均值差异（maximum mean discrepancy，MMD）是一种基于核的统计检验方法，用于判断两个给定分布是否相同[12]。MMD 是通过比较两个分布在一个再生核希尔伯特空间中的嵌入完成的。在本研究中，对于输入的细胞系 RNA 测序数据 $\boldsymbol{X}_b$ 和单细胞 RNA 测序数据 $\boldsymbol{X}_s$，数据经过编码器映射到潜在空间时，引入最大均值差异估算源域嵌入 $\boldsymbol{z}_b$ 和目标域嵌入 $\boldsymbol{z}_s$ 之间的相似度。损失函数定义为

$$L_{\mathrm{mmd}}=\left\|\frac{1}{m}\sum_{i=1}^{m}\phi(\boldsymbol{x}_b^i)-\frac{1}{n}\sum_{j=1}^{n}\phi(\boldsymbol{x}_s^j)\right\|_H \tag{4}$$

其中：$\phi(.)$ 表示将数据映射到再生核希尔伯特空间（RKHS），RKHS 范数 $\|.\|_H$ 用于测量两个数据向量之间的距离。

最优传输是一种求解两个分布之间最小距离的优化算法[13]，其核心思想是将一个概率分布转化为另一个概率分布，使转化后的分布与目标分布之间的距离最小。最优传输算法在机器学习领域受到了越来越多的关注，特别是 Wasserstein-GAN[14]的出现，引起了研究者对最优传输的关注。在本研究中，引入最优传输作为正则化项，调整两种数据域的嵌入。对于嵌入 $\boldsymbol{z}_b^i$ 和 $\boldsymbol{z}_s^j$，其后验分布分别为高斯分布 $N(\boldsymbol{\mu}_{\boldsymbol{z}_b^i\mid\boldsymbol{x}_b^i},\boldsymbol{\sigma}^2_{\boldsymbol{z}_b^i\mid\boldsymbol{x}_b^i}I)$ 和 $N(\boldsymbol{\mu}_{\boldsymbol{z}_s^j\mid\boldsymbol{x}_s^j},\boldsymbol{\sigma}^2_{\boldsymbol{z}_s^j\mid\boldsymbol{x}_s^j}I)$，嵌入之间的传输成本为

$$C_{ij}=\left\|\boldsymbol{\mu}_{\boldsymbol{z}_b^i\mid\boldsymbol{x}_b^i}-\boldsymbol{\mu}_{\boldsymbol{z}_s^j\mid\boldsymbol{x}_s^j}\right\|^2+\left\|\boldsymbol{\sigma}_{\boldsymbol{z}_b^i\mid\boldsymbol{x}_b^i}-\boldsymbol{\sigma}_{\boldsymbol{z}_s^j\mid\boldsymbol{x}_s^j}\right\|^2 \tag{5}$$

批量大小为 B_b 和 B_s 的最优传输方案 $\boldsymbol{T}^*$ 为

$$\begin{aligned}\boldsymbol{T}^*=&\mathrm{argmin}_{T\in\mathbb{R}_+^{B_b\times B_s}}<\boldsymbol{C},\boldsymbol{T}>+\lambda_3 H(\boldsymbol{T})+\lambda_4 D_{\mathrm{KL}}\cdot\\&\left(\boldsymbol{T}\mathbf{1}_{B_s}\,\|\,\frac{1}{B_b}\mathbf{1}_{B_b}\right)+\lambda_4 D_{\mathrm{KL}}\left(\boldsymbol{T}^{\mathrm{T}}\mathbf{1}_{B_b}\,\|\,\frac{1}{B_s}\mathbf{1}_{B_s}\right)\end{aligned} \tag{6}$$

其中：$\lambda_3=0.1$ 和 $\lambda_4=1$ 为平衡参数；$\boldsymbol{T}$ 为传输方案；$<\boldsymbol{C},\boldsymbol{T}>=\sum_{i,j}C_{ij}T_{ij}$ 表示传输成本；$H(\boldsymbol{T})=\sum_{i,j}T_{ij}\log T_{ij}$ 为熵正则化项，可以平衡传输成本和分布均匀性之间的关系；$D_{\mathrm{KL}}\left(\boldsymbol{T}\mathbf{1}_{B_s}\,\|\,\frac{1}{B_b}\mathbf{1}_{B_b}\right)$ 和 $D_{\mathrm{KL}}\left(\boldsymbol{T}^{\mathrm{T}}\mathbf{1}_{B_b}\,\|\,\frac{1}{B_s}\mathbf{1}_{B_s}\right)$ 两个 KL 散度项分别表示当源域和目标域的经验分布为均匀分布时的软边缘约束，$\mathbf{1}_{B_s}$ 和 $\mathbf{1}_{B_b}$ 表示元素全为 1 的向量，其维度分别为 B_s 和 B_b。式(6)表示一个严格的凸优化问题，可以通过采用一种高效的不精确近端点方法[15]计算最优传输方案 $\boldsymbol{T}^*$，其迭代公式为

$$\alpha^{(l+1)}=\left(\frac{\frac{1}{B_b}\mathbf{1}_{B_b}}{\boldsymbol{G}\beta^{(l)}}\right)^{\frac{\lambda_4}{\lambda_4+\lambda_3}},\beta^{(l+1)}=\left(\frac{\frac{1}{B_s}\mathbf{1}_{B_s}}{\boldsymbol{G}^{\mathrm{T}}\alpha^{(l+1)}}\right)^{\frac{\lambda_4}{\lambda_4+\lambda_3}} \tag{7}$$

在初始化时，$\beta^{(0)}=\frac{1}{B_s}\mathbf{1}_{B_s}$，其中 $\boldsymbol{G}=[G_{ij}]$ 且 $G_{ij}=T_{ij}^{(l)}\mathrm{e}^{-\frac{c_{ij}}{\lambda_4}}$。可以从最后一次近端点迭代中获得 α 和 β 的最终值。最优传输方案的计算就转换成了 $\boldsymbol{T}_{ij}^*=\alpha_i G_{ij}\beta_j$。因此，最优传输损失为

$$L_{\mathrm{ot}}=\sum_{i,j}C_{ij}\times\boldsymbol{T}_{ij}^* \tag{8}$$

2.4　药物反应预测器的构建

针对来自细胞系的药物反应数据 $\left\{\boldsymbol{y}_b^i\right\}_{i=1}^m$（即 m 个细胞系对给定药物的二值化反应信息，其中 0 表示药物耐受，1 表示药物敏感）以及来自变分自动编码器提取的细胞系特征 $\boldsymbol{z}_b^i$（$i=1,\cdots,m$），构建了深度为 5 层的多层感知机，用作药物反应预测这一二分类任务。对于训练数据 $\left\{\boldsymbol{z}_b^i,\boldsymbol{y}_b^i\right\}_{i=1}^m$，药物反应预测器的损失函数定义为交叉熵损失函数 BCE()，表达式为

$$\mathrm{BCE}(\boldsymbol{y}_b^i,\hat{\boldsymbol{y}}_b^i)=-\hat{\boldsymbol{y}}_b^i\log(\hat{\boldsymbol{y}}_b^i)-(1-\boldsymbol{y}_b^i)\log(1-\hat{\boldsymbol{y}}_b^i) \tag{9}$$

其中：$\boldsymbol{y}_b^i$ 为细胞系对于给定药物的真实二值化反应值，$\hat{\boldsymbol{y}}_b^i$ 为预测器输出的预测值。对于批量大小为 B_b 的细胞系，药物反应预测器的损失函数为

$$L_{\text{pred}} = \sum_{i=1}^{B_b} \text{BCE}\left(\boldsymbol{y}_b^i, \hat{\boldsymbol{y}}_b^i\right) \tag{10}$$

最后，SCTL 模型总的损失函数为

$$L_{\text{all}} = \rho_1 L_{\text{vae}} + \rho_2 L_{\text{ot}} + L_{\text{mmd}} + \rho_3 L_{\text{pred}} \tag{11}$$

其中：ρ_1、ρ_2 和 ρ_3 是平衡参数。ρ_1 可选为 0.1、0.5、1、2，ρ_2 可选为 0.5、1、10，ρ_3 可选为 0.5、1、5、15。每个数据集可以通过调整平衡参数，使模型的性能处在最佳状态。

当模型训练完成后，将经过跨域特征提取和特征对齐后的单细胞嵌入 $\boldsymbol{z}_s$ 输入药物反应预测器，得到单细胞对于给定药物的预测反应信息。

2.5　模型评价指标

在本研究中，单细胞药物反应预测是一种二分类预测任务。二分类预测问题通常采用 ROC（receiver operating characteristic）曲线下面积（AUC），以及 PR（precision-recall）曲线下面积（AUPR）作为评价指标。其中，ROC 曲线描绘了在不同阈值下模型预测的真阳性率（TPR）和假阳性率（FPR）；AUC 的值范围在 0 ~ 1 之间，值越大表示模型性能越好；AUPR 表示模型预测在所有阈值下的精确率与召回率的表现，值越高表示模型性能越好。

在模型的训练过程中，参考了 SCAD 工作[7]中关于环境随机种子的设置。SCAD 选取了种子（seed）为 42、50、60、70、80 的环境随机种子，在此基础上，又增加一个种子（seed）为 124 的环境随机种子，共 6 个随机种子，用于数据批次的划分，然后计算了模型在这 6 个随机种子下的平均性能。

2.6　模型的可解释性

模型在训练过程中，首先通过最优传输的计算，可以获得当 m 个细胞系和 n 个单细胞都是均匀分布时，最小化传输费用的最优传输矩阵，即 $\boldsymbol{T}^*_{m\times n}$，其中行表示细胞系，列表示细胞，矩阵值表示细胞系分布和单细胞分布间的数据传输量，即细胞系和单细胞间的映射匹配关系。接着，对模型取得较好性能的数据集展开关键基因的分析。对模型预测的细胞药物响应值二值化，将细胞划分为预测敏感细胞簇和预测耐药细胞簇。最后，进行 Wilcoxon 秩和检验，选择 Bonferroni 校正后的 p 小于 0.05、对数倍数变化大于 0.7 的基因作为高表达的基因。

3　实　验

3.1　性能与结果

在 8 个数据集上进行模型测试，并与当前最新的单细胞癌症药物敏感性预测方法进行比较，这些方法包括 scDEAL、SCAD、Beyondcell[16]和 DREEP[17]。

scDEAL 模型通过对细胞系混合 RNA 测序数据和单细胞 RNA 测序数据建模，构建两个去噪自编码器和一个药物反应预测器，用最大均值差异对齐细胞系和单细胞的表征，从而实现将细胞系的真实药物反应信息迁移学习到单细胞水平上。

SCAD 模型采用对抗生成网络 GAN[18]框架，其构建了一个特征提取器，用于学习细胞系和单细胞的表征；一个药物反应预测器，用于学习对药物敏感性推断有用的信息特征；一个域判别器，使特征提取器能够从两个域（细胞系和单细胞）中学习域不变特征。

Beyondcell 方法先从细胞系的基因表达数据中识别出药物反应的上调和下调基因集，然后基于单细胞基因表达数据，计算单细胞在这些基因集的 BCS（Beyondcell Score）得分，该得分可以用来对单细胞依据药物敏感反应的强弱进行排序。

DREEP 方法利用皮尔逊相关系数，将混合测序基因表达与药物基因组筛选的药物反应数据联系起来，构建了药物敏感性基因组图谱（GPDS）的排序列表，GPDS 排序列表中药物的耐药性预测生物标志物位于顶部，敏感性预测生物标志物位于底部。为了在单细胞水平上预测药物反应，该方法基于单细胞 RNA 测序数据，使用 GF-ICF（gene frequency-inverse cell frequency）归一化获取每个细胞的最相关基因集，然后计算该基因集在 GPDS 排序列表中的富集得分，用来衡量每个细胞的药物敏感性。

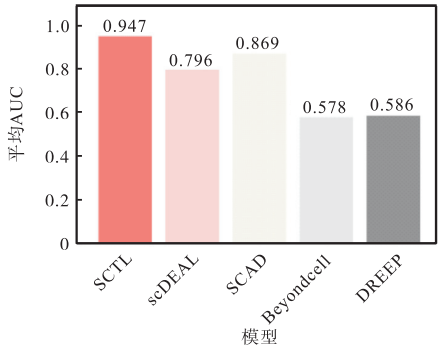
Beyondcell 和 DREEP 都是基于药物敏感性的基因标志物的方法。表 3 为模型在 8 个数据集的结果。SCTL 在 4 个数据集上取得了最优的预测性能，在其他 4 个数据集上的预测性能仅略低于最优预测性能。

与 4 种基线方法相比，SCTL 在 8 个数据集上的平均 AUC 为 0.947，平均 AUPR 为 0.933（图 2）。对平均 AUC 而言，SCTL 比 SCAD 提高了 9.21%，比 scDEAL 提高了 19.22%。对平均 AUPR 而言，SCTL 比 SCAD 提高了 14.5%，比 scDEAL 提高了 17.8%。另外，深度学习模型 SCTL、scDEAL、SCAD 比 Beyondcell 和 DREEP 这两种基于药物敏感性的基因标志物的方法的性能都好，这也反映出深度学习在提取深层次有效特征方面的能力。

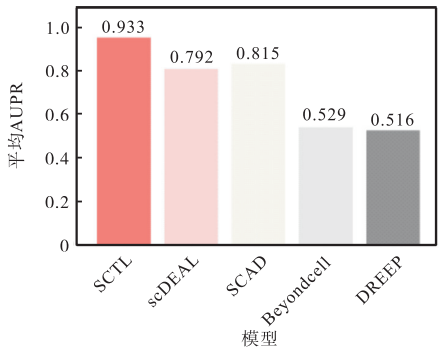
表 3　模型在 8 个数据集上的性能结果

Tab. 3　Performance results of models on eight datasets

序号	AUC					AUPR				
	SCTL	scDEAL	SCAD	Beyondcell	DREEP	SCTL	scDEAL	SCAD	Beyondcell	DREEP
数据 1	0.928	0.823	0.809	0.386	0.497	0.966	0.901	0.885	0.595	0.657
数据 2	0.996	0.880	0.936	0.698	0.590	0.961	0.850	0.498	0.141	0.088
数据 3	0.916	0.950	0.651	0.488	0.500	0.861	0.913	0.749	0.500	0.500
数据 4	0.944	0.947	0.918	0.269	0.521	0.931	0.965	0.856	0.382	0.528
数据 5	0.916	0.648	0.762	0.344	0.500	0.859	0.650	0.633	0.330	0.398
数据 6	0.964	0.904	0.967	0.977	0.788	0.966	0.890	0.973	0.982	0.788
数据 7	0.952	0.613	0.942	0.731	0.667	0.967	0.552	0.958	0.642	0.600
数据 8	0.962	0.600	0.968	0.733	0.621	0.954	0.616	0.970	0.663	0.569



(a) 平均 AUC



(b) 平均 AUPR

图 2　5 种方法的平均 AUC 和平均 AUPR

Fig. 2　Average AUC and average AUPR results of five methods

3.2　消融实验

为了评估模型的各个功能模块对预测性能的影响，在 8 个数据集上进行详细的消融实验分析。

针对特征对齐的计算环节，分别移除最优传输的特征对齐和最大均值差异的特征对齐计算，构建 2 个对比模型。保持每个数据集的最优参数组合不变，然后重新计算了 2 个对比模型在所有数据集上的 AUC 和 AUPR，结果见表 4。当把最大均值差异的特征对齐计算模块移除时，与原先模型相比，其平均 AUC 和平均 AUPR 均有所下降。当移除最优传输特征对齐计算时，模型的预测性能出现了明显的下降，与原先模型相比，平均 AUC 和平均 AUPR 均有所下降。这表明在特征对齐的计算里，最优传输计算发挥了重要作用。

表 4　特征对齐的消融实验结果

Tab. 4　Ablation experiment results for feature alignment

序号	SCTL		SCTL_1		SCTL_2	
	AUC	AUPR	AUC	AUPR	AUC	AUPR
数据 1	0.928	0.966	0.846	0.911	0.635	0.742
数据 2	0.996	0.961	0.886	0.828	0.828	0.328
数据 3	0.916	0.861	0.913	0.860	0.708	0.737
数据 4	0.944	0.931	0.932	0.926	0.517	0.516
数据 5	0.916	0.859	0.847	0.841	0.459	0.372
数据 6	0.964	0.966	0.943	0.942	0.640	0.674
数据 7	0.952	0.966	0.895	0.913	0.845	0.826
数据 8	0.962	0.967	0.889	0.896	0.617	0.620
均值	0.947	0.935	0.894	0.890	0.656	0.602
标准差	0.028 1	0.046 6	0.035 6	0.041 7	0.135 4	0.180 8

注：SCTL_1 表示去除最大均值差异，SCTL_2 表示去除最优传输。

针对药物反应预测器的模型结构，对多层感知机隐藏层的层数进行消融分析。构建 6 个网络深度各不相同的药物反应预测器，分别是具有 2 层、3 层、4 层、5 层和 6 层隐藏层层数的神经网络。设置每层隐藏层的节点数为 128 或 64，激活函数为 ReLU。按照表 5 设置药物反应预测器的网络结构，实验结果见表 6。初始预测器网络层数为 2，随着预测器隐藏层层数的增加，AUC 和 AUPR 得分增加；在隐藏层层数增加到 5 时，模型的整体性能表现最好；当多层感知机深度继续增加到 6 层时，模型在大部分数据集上的性能指标出现了明显下降。

表 5　药物反应预测器的网络结构

Tab. 5　Network architecture of drug response predictor

网络	网络层及节点数	激活函数
predictor1	128,64	ReLU
predictor2	128,128,64	ReLU
predictor3	128,128,128,64	ReLU
predictor4	128,128,128,128,64	ReLU
predictor5	128,128,128,128,128,64	ReLU

在跨域特征提取部分，对细胞系混合测序数据有 3 种数据采样方式，分别为不采样、加权随机采样、合成少数类过采样(synthetic minority oversampling technique，SMOTE)。加权随机采样是根据数据中正负样本数的比例，对占比小的数据，提高其被采样的概率。SMOTE 是通过在特征空间中合成新的少数类样本平衡数据集，从而提高分类模型对少数类的识别能力[19]。在 8 个数据集上，分别展示了 3 种数据采样方式对结果的影响，见表 7。

表 6　不同药物反应预测器的性能对比结果

Tab. 6　Performance comparison results for different drug response predictors

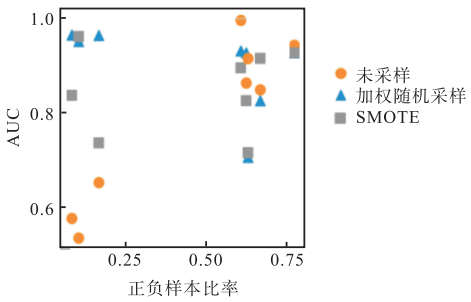
序号	AUC					AUPR				
	predictor1	predictor2	predictor3	predictor4	predictor5	predictor1	predictor2	predictor3	predictor4	predictor5
数据 1	0.711	0.797	0.786	0.928	0.907	0.875	0.908	0.882	0.966	0.958
数据 2	0.966	0.977	0.989	0.996	0.917	0.936	0.940	0.941	0.961	0.619
数据 3	0.878	0.889	0.891	0.916	0.862	0.817	0.831	0.852	0.861	0.820
数据 4	0.935	0.937	0.943	0.944	0.938	0.921	0.918	0.931	0.931	0.917
数据 5	0.805	0.832	0.801	0.916	0.908	0.821	0.771	0.816	0.859	0.894
数据 6	0.879	0.907	0.895	0.964	0.884	0.915	0.911	0.919	0.966	0.898
数据 7	0.906	0.919	0.924	0.965	0.776	0.925	0.932	0.939	0.967	0.852
数据 8	0.919	0.921	0.937	0.962	0.896	0.912	0.914	0.926	0.954	0.881

表 7　3种采样方式的性能对比结果

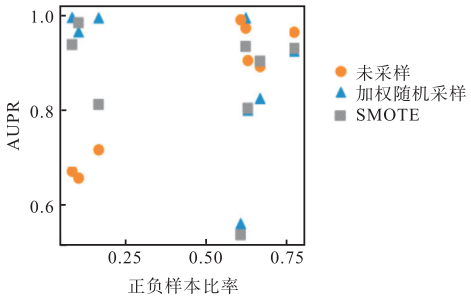
Tab. 7　Performance comparison results for three sampling methods

序号	不采样		加权随机采样		SMOTE	
	AUC	AUPR	AUC	AUPR	AUC	AUPR
数据 1	0.865	0.941	0.928	0.966	0.828	0.895
数据 2	0.996	0.961	0.932	0.458	0.897	0.430
数据 3	0.916	0.861	0.710	0.738	0.719	0.743
数据 4	0.944	0.931	0.927	0.884	0.928	0.891
数据 5	0.851	0.846	0.828	0.767	0.916	0.859
数据 6	0.657	0.640	0.964	0.966	0.740	0.752
数据 7	0.582	0.587	0.965	0.967	0.839	0.900
数据 8	0.541	0.571	0.952	0.932	0.962	0.954

此外，还对 8 个数据集的源域(混合 RNA 测序数据)的正负样本比率进行了计算，数据 1—数据 8 源域正负样本比率为 0.618、0.602、0.624、0.768、0.662、0.161、0.078、0.099。对正负样本比率和 3 种采样方式的 AUC、AUPR 进行可视化分析，结果如图 3 所示。当源域数据正负样本处于基本平衡时，不进行采样就可以获得较好的性能；而当正负样本数差别很大时，如数据 6、7、8，使用采样的方法可以很好地提高模型的性能。



(a) AUC



(b) AUPR

图 3　正负样本比率和采样方法的关系

Fig. 3　Relationship between positive and negative sample ratio and sampling method

3.3　鲁棒性实验

为了测定模型预测结果的稳定性，进行鲁棒性实验。随机的环境种子会影响数据集批次的划分。固定模型在每个数据集上训练的最优参数，然后在随机的环境种子下重复实验 20 次，结果如图 4 所示。模型在 8 个数据集上的波动都很小，AUC 和 AUPR 的平均标准差分别为 0.050 4 和 0.048 2，这表明模型具有很好的稳定性。

分别对每个单细胞 RNA 测序数据进行 80% 的随机采样，然后对采样得到的单细胞数据进行模型训练预测，并重复实验 20 次得到性能结果。模型每一次实验都会在不一样的目标域数据上进行跨域特征对齐和药物反应预测，结果如图 5 所示。模型在 8 个单细胞 RNA 测序数据上的结果均保持了良好的稳定性，这说明模型对不同的目标域数据均进行了有效的特征学习和特征对齐。

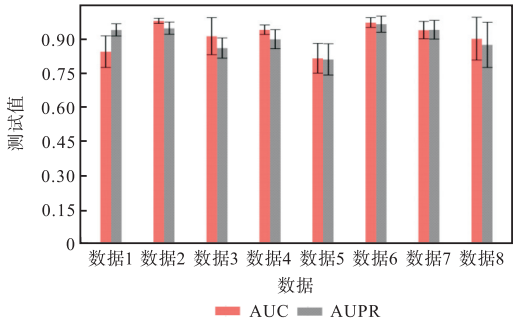


图 4　不固定随机种子重复 20 次的鲁棒性实验

Fig. 4　Robustness experiment repeated 20 times with unfixed random seed

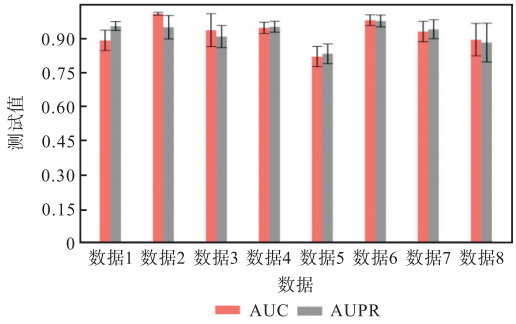


图 5　随机采样 80% 单细胞 RNA 测序数据重复 20 次的鲁棒性实验

Fig. 5　Robustness experiment repeated 20 times with 80% random sampling of single-cell RNA-seq data

3.4　模型可解释性的案例分析

以数据 6 为例，这是关于头颈部鳞状细胞癌在吉非替尼药物作用下的数据。该单细胞数据来自头颈部鳞状细胞癌模型细胞系[20-21]，由两簇细胞组成：JHU006_Gefitinib_Sen 和 JHU006_Gefitinib_Res。其中，JHU006_Gefitinib_Sen 细胞簇对吉非替尼敏感，JHU006_Gefitinib_Res 细胞簇对吉非替尼耐药。

模型训练完成后输出细胞系和单细胞间映射匹配的最优传输矩阵 $\boldsymbol{T}^*$。使用热图可视化细胞系和单细胞间的最优传输矩阵值（图 6），为了方便展示，这里分别随机选择了 90 个敏感细胞系以及 90 个耐药细胞系。在热图中，可以观察到对吉非替尼敏感的细胞与对吉非替尼敏感的细胞系之间的匹配关系更紧密，对吉非替尼耐药的细胞与对吉非替尼耐药的细胞系之间的联系也更强。这表明在特征提取和特征对齐的环节，模型获得了细胞系和单细胞间关于药物反应分层的准确匹配关系，从而模型可以获得准确的单细胞药物反应预测结果。

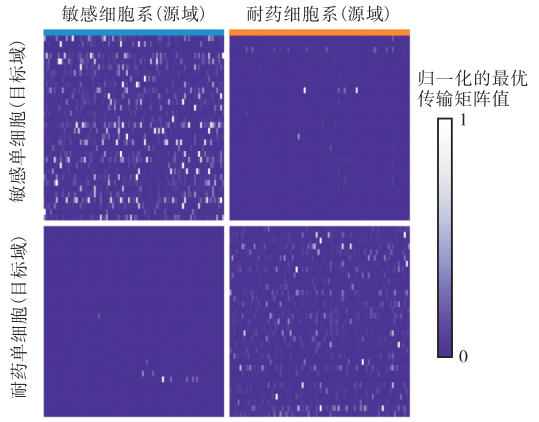


图 6　数据 6 中细胞系和单细胞间最优传输矩阵热图

Fig. 6　Heatmap of optimal matrix between source domain and target domain for data 6

为分析在模型输出药物响应预测值的过程中起到关键作用的基因的影响，对模型输出的药物响应预测连续值二值化，阈值选取这组预测连续值的中位数，将单细胞划分为预测敏感细胞簇（值为 1）和预测耐药细胞簇（值为 0），成功预测分类了 90.9% 细胞的药物响应状态。对两簇细胞采用 Wilcoxon 秩和检验，在两簇细胞上分别筛选了 Top 10 个敏感高表达基因和 Top 10 个耐药高表达基因，并在预测敏感细胞簇和预测耐药细胞簇上对这些基因的表达值用热图可视化，以及在真实敏感和真实耐药细胞簇上对这些基因的表达值可视化，如图 7 所示。

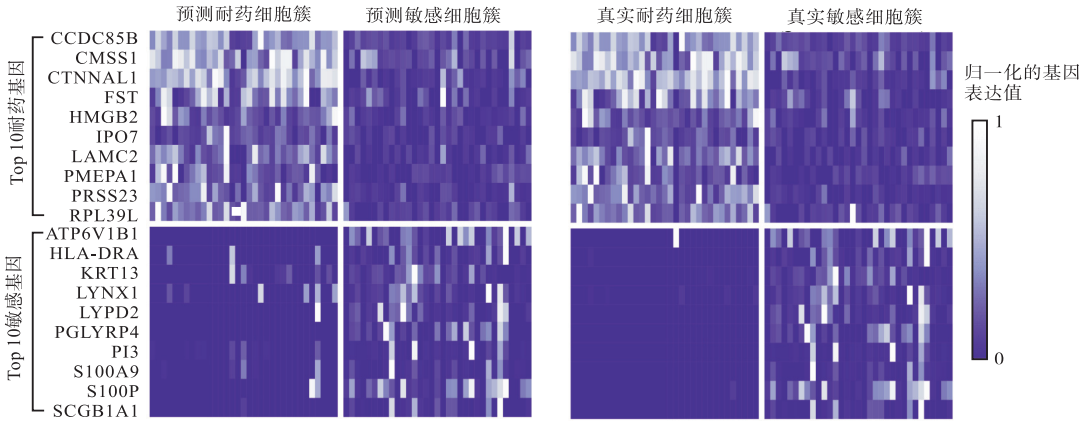


图 7　数据 6 中 Top 10 高表达基因在预测细胞簇和真实细胞簇上的基因表达热图

Fig. 7　Heatmap of gene expression for the Top 10 highly expressed genes in data 6 on predicted cell clusters and actual cell clusters

筛选得到的基因在模型预测的细胞簇上有明显的表达差异，也在真实的细胞簇上有明显的表达差异。这些高表达基因在模型预测细胞药物响应状态过程中，可被认为起到了关键作用。总体而言，本文模型在预测任务上取得优异性能的同时，还在预测的细胞簇上筛选到高表达基因，这些基因作为癌细胞的药物敏感标志物和耐药标志物，对后续的治疗分析起到了潜在支持。

4　结　语

本研究提出一种用于单细胞抗癌药物反应预测的迁移学习模型 SCTL，其利用现有的大规模细胞系混合 RNA 测序数据，将细胞系的药物基因组信息迁移到单细胞 RNA 测序数据上。与基准方法对比后发现，SCTL 具有较高的预测性能。消融实验的结果展示了不同模块对于模型准确预测的有效性。鲁棒性实验显示了模型预测结果在不同的环境随机种子以及单细胞数据随机采样等情况下具有稳定性。

针对单细胞药物响应预测研究领域，本研究提出使用迁移学习计算方法，有效地弥合了单细胞测序数据和细胞系测序数据之间的差异。然而，本文方法也存在着局限性。一个主要挑战是如何在不损失特征信息的情况下，实现源域与目标域数据特征的对齐和信息的跨域迁移。本文方法由于使用共享编码器学习源域和目标域特征，这一过程可能会损失不同数据域的特异性信息，继而在特征对齐的环节影响特征信息的迁移。虽然通过调整超参数可以使模型在不同数据集上表现出较好的性能，但模型的泛化能力仍有待提高。为了应对这些问题，计划在接下来的研究中引入注意力机制，以增强模型对源域和目标域数据特征的学习能力，提高模型的泛化性和预测性能。

本文中用到的单细胞数据主要集中在临床前模型上，缺乏临床水平的患者验证。随着临床单细胞组学数据的日益丰富，在单细胞水平上预测药物反应的 SCTL 模型可以更深入地揭示细胞间的异质性和功能差异，为个性化和精准的癌症治疗提供新的策略。

参考文献：

[1] CHEN J Y，WANG X Y，MA A J，et al. Deep transfer learning of cancer drug responses by integrating bulk and single-cell RNA-seq data[J]. Nature communications，2022，13(1)：6494.

[2] CARREIRA S，ROMANEL A，GOODALL J，et al. Tumor clone dynamics in lethal prostate cancer[J]. Science translational medicine，2014，6(254)：254ra125.

[3] WONG C H，SIAH K W，LO A W. Estimation of clinical trial success rates and related parameters[J]. Biostatistics，2019，20(2)：273–286.

[4] RAMBOW F，ROGIERS A，MARIN-BEJAR O，et al. Toward minimal residual disease-directed therapy in melanoma[J]. Cell，2018，174(4)：843–855.

[5] VERJANS E T，DOIJEN J，LUYTEN W，et al. Three-dimensional cell culture models for anticancer drug screening：worth the effort?[J] Journal of cellular physiology，2018，233(4)：2993–3003.

[6] SCHIRLE M，JENKINS J L. Identifying compound efficacy targets in phenotypic drug discovery[J]. Drug discovery today，2016，21(1)：82–89.

[7] ZHENG Z T，CHEN J Y，CHEN X J，et al. Enabling single-cell drug response annotations from bulk RNA-seq using SCAD[J]. Advanced science，2023，10(11)：e2204113.

[8] YANG W J，SOARES J，GRENINGER P，et al. Genomics of drug sensitivity in cancer (GDSC)：a resource for therapeutic biomarker discovery in cancer cells[J]. Nucleic acids research，2013，41(D1)：955–961.

[9] BARRETINA J，CAPONIGRO G，STRANSKY N，et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity[J]. Nature，2012，483(7391)：603–607.

[10] WOLF F A，ANGERER P，THEIS F J. SCANPY：large-scale single-cell gene expression data analysis[J]. Genome biology，2018，19(1)：15.

[11] CAO K，GONG Q，HONG Y G，et al. A unified computational framework for single-cell data integration with optimal transport[J]. Nature communications，2022，13(1)：7419.

[12] ZHANG W，WU D R. Discriminative joint probability maximum mean discrepancy (DJP-MMD) for domain adaptation[C]//IEEE. 2020 International Joint Conference on Neural Networks (IJCNN). New York：IEEE，2020：1–8.

[13] CUTURI M. Sinkhorn distances：lightspeed computation of optimal transport[J]. Advances in neural information processing systems，2013，26：15966283.

[14] ARJOVSKY M，CHINTALA S，BOTTOU L. Wasser-

（下转第 80 页）

113646.

[7] FENG J L, JIANG W S, LI D L, et al. Characteristics of tide-surge interaction and its roles in the distribution of surge residuals along the coast of China[J]. Journal of oceanography, 2019, 75: 225–234.

[8] FENG J L, LI D L, DANG W, et al. Changes in storm surges based on a bias-adjusted reconstruction dataset from 1900 to 2010[J]. Journal of hydrology, 2023, 617: 128759.

[9] BUTLER A, HEFFERNAN J E, TAWN J A, et al. Extreme value analysis of decadal variations in storm surge elevations[J]. Journal of marine systems, 2007, 67(1/2): 189–200.

[10] MARCOS M, JORDÀ G, GOMIS D, et al. Changes in storm surges in southern Europe from a regional model under climate change scenarios[J]. Global and planetary change, 2011, 77(3/4): 116–128.

[11] GUO Z, CAO A Z, MA W Y, et al. Spatiotemporal features of storm surges in the Zhoushan Archipelago from 2001 to 2020[J]. Continental shelf research, 2025, 291: 105492.

[12] 吴亚楠, 王智峰, 董胜, 等. 山东沿海台风暴潮数值模拟与统计分析[J]. 自然灾害学报, 2015, 24(3): 169–176.

[13] WANG Y P, MAO X Y, JIANG W S. Long-term hazard analysis of destructive storm surges using the ADCIRC-SWAN model: a case study of Bohai Sea, China[J]. International journal of applied earth observation and geoinformation, 2018, 73: 52–62.

[14] CHOI B H, KIM K O, EUM H M. Digital bathymetric and topographic data for neighboring seas of Korea[J]. Journal of Korean society of coastal and ocean engineers, 2002, 14(1): 41–50.

[15] CHEN C S, LIU H D, BEARDSLEY R C. An unstructured grid, finite-volume, three-dimensional, primitive equations ocean model: application to coastal ocean and estuaries[J]. Journal of atmospheric and oceanic technology, 2003, 20(1): 159–186.

[16] 陈波昌, 魏皓. FVCOM 模型在渤海湾潮流潮汐模拟中的应用[J]. 天津科技大学学报, 2013, 28(4): 40–43.

[17] DING Y M, WEI H. Modeling the impact of land reclamation on storm surges in Bohai Sea, China[J]. Natural hazards, 2017, 85: 559–573.

[18] 丁玉梅, 丁磊. 渤海风暴潮过程的数值模拟研究[J]. 天津科技大学学报, 2018, 33(3): 57–62.

[19] 史道济. 实用极值统计方法[M]. 天津: 天津科学技术出版社, 2006.

[20] 赵鹏. 渤海寒潮风暴潮增水风险的数值研究[D]. 青岛: 中国海洋大学, 2010.

责任编辑： 周建军

(上接第 26 页)

stein generative adversarial networks[C]//PMLR. International Conference on Machine Learning. New York: PMLR, 2017: 214–223.

[15] XIE Y J, WANG X F, WANG R J, et al. A fast proximal point method for computing exact Wasserstein distance[C]//PMLR. Uncertainty in Artificial Intelligence. New York: PMLR, 2020: 433–453.

[16] FUSTERO-TORRE C, JIMÉNEZ-SANTOS M J, GARCÍA-MARTÍN S, et al. Beyondcell: targeting cancer therapeutic hetero-geneity in single-cell RNA-seq data[J]. Genome medicine, 2021, 13(1): 187.

[17] PELLECCHIA S, VISCIDO G, FRANCHINI M, et al. Predicting drug response from single-cell expression profiles of tumours[J]. BMC Medicine, 2023, 21(1): 476.

[18] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139–144.

[19] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of artificial intelligence research, 2002, 16: 321–357.

[20] KINKER G S, GREENWALD A C, TAL R, et al. Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity[J]. Nature genetics, 2020, 52(1): 1208–1218.

[21] SINHA S, VEGESNA R, MUKHERJEE S, et al. Perception predicts patient response and resistance to treatment using single-cell transcriptomics of their tumors[J]. Nature cancer, 2024, 5(6): 938–952.

责任编辑： 郎婧