# Deep transfer learning of cancer drug responses by integrating bulk and single-cell RNA-seq data——复现报告

Junyi Chen, Xiaoying Wang, Anjun Ma, Qi-En Wang, Bingqiang Liu, Lang Li, Dong Xu & Qin Ma

19 October 2022

**摘要**

该论文提出了 scDEAL（Single-Cell Drug rEsponse AnaLysis），一个基于深度迁移学习的框架，用于预测单细胞水平的癌症药物响应。该框架通过整合大规模的细胞系数据（bulk RNA-seq）和单细胞 RNA 测序（scRNA-seq）数据，训练模型以预测药物响应。scDEAL 的关键特点是通过集成梯度特征解释来推断药物抗性机制的标志性基因。研究者在六个 scRNA-seq 数据集上对 scDEAL 进行了基准测试，并通过对药物响应标签预测、基因签名识别和伪时间分析的三个案例研究来展示其模型可解释性。scDEAL 有望帮助研究细胞重编程、药物选择和再利用，以提高治疗效果。

**关键词：**单细胞 RNA 测序（scRNA-seq）；深度迁移学习（DTL）；基因表达；癌症治疗

## 1 引言

癌症治疗的个体化需求日益增长，特别是在精准医学的背景下，对癌症细胞对药物的响应进行精确预测变得至关重要。传统的基于细胞系的癌症药物筛选方法虽然提供了大量关于药物敏感性的数据，但这些数据往往无法直接应用于单细胞水平的药物响应预测。单细胞 RNA 测序（scRNA-seq）技术的出现，使得研究者能够观察到癌症细胞群体中药物响应的异质性 [8]，这对于理解癌症的复杂性和开发更有效的治疗策略具有重要意义。然而，scRNA-seq 数据的高维度和公共数据库中有限的标记数据样本，限制了基于 scRNA-seq 数据的药物响应预测模型的开发和应用。

本文提出的 scDEAL 框架，通过深度迁移学习技术，有效地整合了 bulk RNA-seq 数据和 scRNA-seq 数据，解决了单细胞数据训练样本不足的问题。这一方法不仅提高了单细胞药物响应预测的准确性，而且通过集成梯度特征解释，增强了模型的可解释性，有助于揭示药物抗性的关键基因。这对于理解癌症细胞的生物学特性、优化药物选择和开发新的治疗策略具有重要的研究意义。此外，scDEAL 的提出也为癌症治疗领域提供了一个强大的工具，有助于推动精准医学的发展，特别是在药物开发和疗效优化方面。通过 scDEAL，研究者可以更深入地理解癌症细胞的异质性和药物响应机制 [2]，为个体化治疗提供更精准的指导，从而提高癌症治疗的成功率和患者的生活质量。

选择复现这篇论文的原因有两点。一是研究方向契合，我的研究方向正是与这篇论文密切相关的领域——细胞与癌症药物治疗的关联性。论文涉及的主题与我当前的研究兴趣高度一致，这将使得复现工作不仅能够加深我对该领域的理解，也能够对我未来的研究提供宝贵的经验和启示。二是迁移学习应用，迁移学习是这篇论文的一个重要训练策略，而我的研究中同样需要运用迁移学习来迁移细胞之间的关系。通过复现这篇论文，我将深入学习并掌握论文中采用的迁移学习方法，为我的研究提供实用经验和参考。

# 2 相关工作

## 2.1 RNA 测序

单细胞 RNA 测序（scRNA-seq）和批量 RNA 测序（bulk RNA-seq）是两种不同的 RNA 测序方法，它们在样本处理、数据分析和应用领域等方面存在显著差异。针对不同的应用需求，可以选择不同的测序方式。

### 2.1.1 单细胞 RNA 测序

scRNA-seq 是一种用于测定单个细胞内 RNA 分子的测序技术 [6]。与传统的 bulk RNA-seq 不同，scRNA-seq 可以揭示细胞群体中个体细胞的基因表达差异，提供了更为细致的生物信息。在应用方面，单细胞 RNA 测序主要适用于研究个体细胞之间的异质性，揭示不同细胞亚群的存在、细胞分化过程、细胞状态变化等。这种测序方法的特点在于数据量较小，处理过程相对复杂，因为需要从单个细胞中提取 RNA、构建文库，存在较多的技术挑战。通常使用微流控芯片或微滴式分选技术以单细胞为单位进行操作 [11]。

### 2.1.2 批量 RNA 测序

bulk RNA-seq 是一种对整个细胞群体进行 RNA 测序的方法，将细胞总 RNA 提取并进行测序，从而获得群体平均的基因表达水平 [9]。在应用方面，批量 RNA 测序适用于对细胞总体基因表达状况进行分析，用于研究组织、器官或细胞群体的平均表达情况，如疾病与对照组的比较。这种测序方法的特点在于数据量较大，处理相对简单，适用于高通量测序技术。但因为是对整体细胞群体的表达进行分析，所以在一定程度上掩盖了细胞异质性。

## 2.2 细胞异质性

细胞异质性是指细胞群体内个体细胞之间的差异性，表现为基因表达、形态、功能等方面的多样性。细胞异质性在生物体内普遍存在，是生命体系的一个重要特征。癌症治疗就与细胞异质性有关，它是癌症治疗中治疗抵抗性、肿瘤复发、靶向治疗挑战存在的原因。癌症细胞内部存在不同亚群的表现，这些亚群可能表现出不同的治疗敏感性。某些亚群可能对特定治疗更为敏感，而其他亚群则可能表现出抵抗性。因此，细胞异质性的存在使得一种治疗方法难以对所有癌症细胞产生一致的疗效 [7]。在治疗期间，某些亚群的癌细胞可能逃脱治疗而残留，这些残留的癌细胞可能具有更高的恶性程度，导致肿瘤复发。

# 3 本文方法

## 3.1 本文方法概述

本文提出方法的创新点在于迁移学习的应用 [4]，通过对 bulk RNA-seq 数据的迁移，将 bulk 的知识迁移到单细胞层面，进而实现更高的预测准确率。

开发用于预测单细胞药物反应的基于深度学习的工具的主要障碍是由于公共领域的基准数据数量有限，训练能力不足。直观地说，药物相关的 bulk RNA-seq 数据可以作为推断基因表达-药物反应关系的有效补充资源，支持单细胞水平上的药物反应预测。幸运的是，深度迁移学习 (DTL, deep transfer learning) 可以将知识和关系模式从 bulk 数据转移到单细胞数据，这可以成为克服训练数据有限问题的一种手段。使用迁移学习，我们可以在单细胞水平上使用 bulk 水平的初步训练模型解决特定任务，可以全部或部分解决。

这里对 scDEAL 框架进行介绍。scDEAL 方法的框架如图 1所示，首先在 bulk 层面上模拟基因表达特征与药物反应之间的关系，然后，通过识别单细胞数据和批量数据之间共享的低维特征空间，协调两种数据类型之间的关系并捕获 bulk 层面上的基因表达-药物反应关系。训练 DTL 模型来学习上述两个关系的优化解。最后，通过 DTL 模型中单细胞水平基因表达、bulk 层面的基因表达与药物反应的元关系，构建单细胞-药物反应关系。

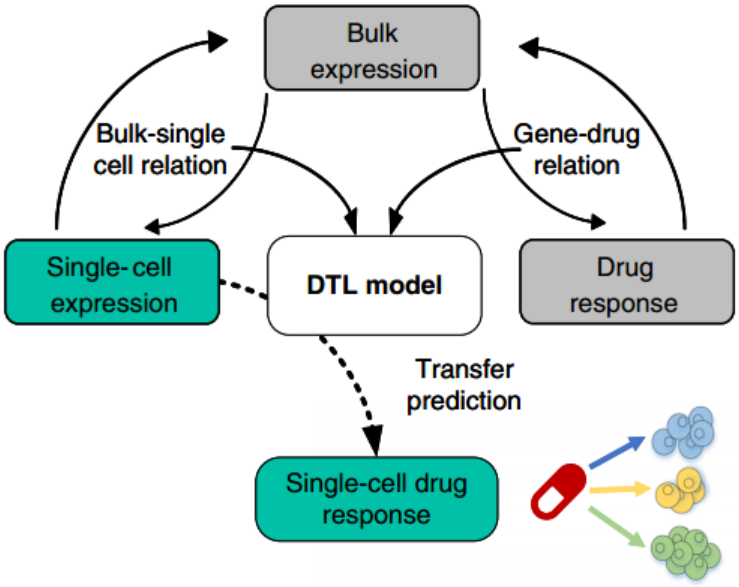


图 1. 模型框架概览

## 3.2 低维特征提取模块

与自编码器 AE 的学习只是简单地保留原始输入数据的信息，并不能确保获得一种有用的特征表示不同，本方法使用的去噪自编码器 DAE 能够学习叠加噪声的原始数据，并且学习到的特征与从未叠加噪声的数据学到的特征几乎一样 [14]。这样一来，去噪自编码器从叠加噪声的输入中学习得到的特征更具鲁棒性，并且可以避免像自编码器一样可能学习不到特别有用信息特征的问题。

本方法使用去噪自编码器的一大原因是 bulk RNA-seq 和 scRNA-seq 数据中的噪声特征非常不同，使用 DAE 模型，在特征降维之前在在两种数据类型中诱导噪声。通过这种方式，

可以避免不平衡训练的风险，不平衡训练只会迫使 scRNA-seq 数据中的基因表达接近于 bulk RNA-seq 数据，进而缺失了单细胞所特有的异质性。

### 3.3 深度迁移学习

在本论文方法中，深度迁移学习分别在以下方面发挥了作用。

- 领域自适应：bulk RNA-seq 和 single-cell RNA-seq 数据具有明显的差异，例如数据的维度、噪声水平、稀疏性等。深度迁移学习可以通过领域自适应的方法来减小不同数据领域之间的差异，提高模型在不同数据类型上的泛化性能。
- 共享特征学习：由于 bulk RNA-seq 和 single-cell RNA-seq 数据共享相似的生物学特征，通过深度迁移学习可以在一个任务上学到的共享特征，从而在另一个任务上提升性能。在 scDEAL 方法中，通过在 bulk RNA-seq 数据上进行预训练，然后将学到的特征迁移到 single-cell RNA-seq 任务上，可以加速模型的收敛并提高性能。
- 跨组织学习：在生物医学研究中，不同组织的 RNA-seq 数据可能存在差异。深度迁移学习可以帮助模型在一个组织上学到的知识迁移到另一个组织上，从而更好地适应不同组织的特异性。
- 单细胞异质性研究：在单细胞 RNA-seq 中，细胞异质性是一个重要的研究方向。深度迁移学习可以帮助识别和理解不同细胞类型之间的共享和特异特征，为深入研究单细胞异质性提供支持。
- 多任务学习：在同时处理 bulk RNA-seq 和 single-cell RNA-seq 数据时，深度迁移学习的多任务学习方法可以使模型更好地从两者中受益，提高模型整体性能。

## 4 复现细节

### 4.1 与已有开源代码对比

复现过程中使用了作者已经开源的代码，同时自己实现了单细胞数据部分药物反应预测的实现以及预测结果可视化、交叉验证降低过拟合以及单细胞数据增强部分代码。

在结果可视化部分，利用 UMAP 图，可以将药物反应的模式以一种清晰的方式呈现出来。在图中使用颜色编码来表示样本的药物灵敏性或抗性，使得观察者能够直观地识别哪些细胞对于特定药物是敏感的，哪些是耐药的。

交叉验证降低过拟合，通过使用五折交叉验证，可以更可靠地评估模型的性能，减少因数据划分不同而引入的随机性。这对于确保模型在不同数据子集上都能表现出良好的泛化性能非常重要。

由于单细胞测序的技术问题，难以测出低表达率的基因数据，因此会对这些数据进行认为的赋零值，这被称为 “dropout” 现象，这个现象影响了数据的完整性和对下游任务的分析，如差异表达分析、细胞聚类等，因此对单细胞基因表达数据进行数据增强是很有必要的。我使用了基于图嵌入的神经网络模型来对单细胞数据进行增强，对下游任务进行分析，可以得到数据增强的效果。

复现过程中，除了将作者给定的六个参数模型进行复现外，我还修改了部分参数，额外训练了 6 个迁移模型，对迁移的效果进行比较。对数据集进行分析处理，并对两种数据进行增强。

### 4.2 实验环境搭建

| | | |
|---|---|---|
| 硬件环境 | CPU | Intel(R) Core(TM) i5-8265U CPU @ 1.60GHz 1.80 GHz |
| | GPU | NVIDIA GeForce Mx230 |
| | 内存 | 8GB |
| 操作系统 | OS | Windows 10 家庭中文版 |
| 开发环境 | 编程语言 | Python 3.7 |
| | 深度学习框架 | Pytorch |
| | 开发工具 | PyCharm 2021.2.3 (Community Edition) |

### 4.3 创新点

复现时对单细胞基因表达数据进行数据增强即对基因表达数据的 dropout 现象进行了数据补全。dropout 现象是指单个细胞某基因的含量过低，未能检测到这些基因的表达，导致人为的零表达，使得基因表达数据变得稀疏。dropout 现象普遍存在于单细胞数据集中，在 bulk 细胞数据集中存在较少，因为 bulk 细胞数据集中记录的一个组织或器官中一群细胞的基因含量，细胞的数量升高，会使得基因的表达量也提升，进而 dropout 现象存在较少。

对 dropout 零值进行数据补全对于单细胞数据的分析至关重要。准确的数据补全可以提高下游分析的可靠性，如差异表达基因（DEGs）的识别、细胞类型注释、细胞轨迹分析等。这些分析对于理解细胞状态的变异、发现新的生物标记物以及揭示疾病发展的机制等方面具有重要意义。尤其是在这篇论文的应用当中，数据补全的意义重大。在癌症治疗中，了解细胞在不同治疗条件下的基因表达差异对于理解治疗响应和耐药机制至关重要。首先数据补全可以帮助更准确地识别差异表达基因，从而揭示潜在的治疗相关的生物学特征。其次可以更好地捕捉细胞状态的动态变化，从而支持细胞轨迹分析。最后，数据补全有助于更准确地预测个体单细胞对于特定药物治疗的反应，这对于实现个性化治疗和减少治疗的不良反应具有重要意义。

为了评估数据增强的效果，选取 GSE117872 数据集，对该数据集进行置零并增强操作。分别置零了该数据集中的 10%、20%、40% 的数据（针对非零值进行置零），对三个置零后的数据集进行模型训练，得到数据增强后的结果，并将生成的数据集与原始的数据集进行比较 RMSE [1]，L1_distance [12] 以及皮尔逊系数（pearson_corr）[13] 计算，其中 RMSE，L1_distance 的值越低，皮尔逊系数越高说明数据增强的效果越好。表 1为数据增强后与初始数据进行比较得到的结果，可以看出，新生成的数据与原始数据相似性高，没有丢失单细胞基因表达数据的相关性。

## 5 实验结果分析

我对作者提供的六个不同的数据集以及对应的药物进行了 bulk 数据集以及 single-cell 数据集的模型复现，其中 6 个数据集信息如表 2 所示，并选择了迁移效果较为优秀的进行结果

表 1. 数据增强效果比较

| Drop rate(%) | RMSE | l1_distance | pearson_corr |
|---|---|---|---|
| 10 | 0.613 | 0.230 | 0.951 |
| 20 | 0.708 | 0.272 | 0.931 |
| 40 | 0.999 | 0.398 | 0.859 |

展示。下面的结果展示都是关于 GSE149383 数据集以及用于治疗肺癌的埃罗替尼药物的。

表 2. 六个训练数据集信息

| | Drug | GEO access | Cells | Cancer type |
|---|---|---|---|---|
| Data 1&2 | Cisplatin | GSE117872 | 548 | Oral squamous cell carcinomas |
| Data 3 | Gefitinib | GSE112274 | 507 | Lung cancer |
| Data 4 | Docetaxel | GSE140440 | 324 | Prostate Cancer |
| Data 5 | Erlotinib | GSE149383 | 1496 | Lung cancer |
| Data 6 | I-BET-762 | GSE110894 | 1419 | Acute myeloid leukemia |

下面对迁移前后的预测结果以及 ground truth 进行了可视化。从图 2中可以看到颜色越浅代表单细胞中基因对药物越敏感，相反，颜色越深即颜色越接近黑色，代表该细胞基因对药物表现为耐药性。最右边的图为 ground truth，此时对药敏感的细胞都分布在上方，对药表现为抗性的细胞分布在图片下方。再看迁移以前的灵敏预测细胞分布图，可以看到更多的细胞被预测为对药敏感，只有左下角一部分细胞被预测为耐药性。看中间的敏感细胞分布图，经过 bulk 细胞知识的迁移以后，敏感或抗药细胞的分布更接近于 ground truth，在敏感细胞分布集中的区域里，被预测为耐药的细胞数也比迁移以前的少了很多。

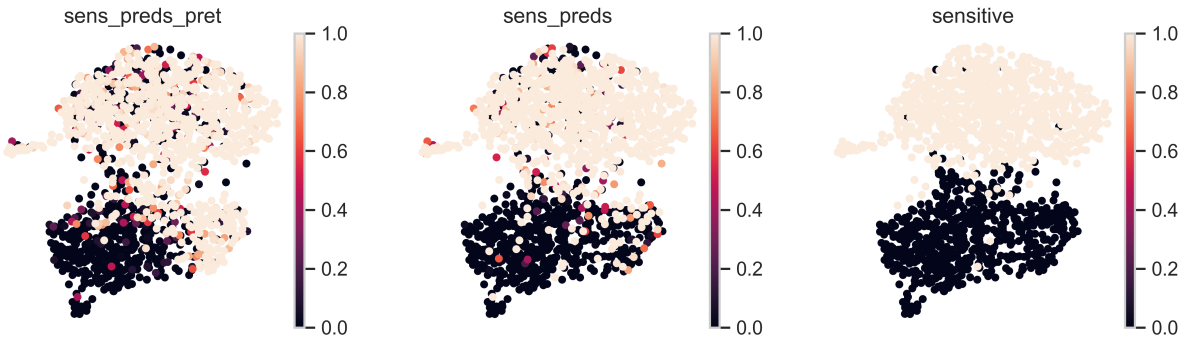


图 2. 迁移前后灵敏预测比较图

其中迁移前后的预测结果评价指标如表 3所示。从表格数据中可以看到，迁移学习在药物反应预测任务中呈现出显著的提升，模型性能改善。综合来看，迁移前后 F1 Score、平均精确率（AP）、ROC 曲线下面积（AUROC）等指标均显示出明显的增长。这说明迁移学习有助于提升模型在预测药物反应方面的全面表现，使其更为准确和可靠。

表 3. 迁移前后评价指标评价

| | f1 | ap | auroc |
|---|---|---|---|
| pretrain | 0.7468656 | 0.8103518 | 0.7394002 |
| transfer | 0.8922971 | 0.9568624 | 0.8889841 |

为了更清晰的看到预测结果与 ground truth 的区别，使用 UMAP 图 [5]，如图 4所示，通过突出显示错误预测，可以更直观地了解模型在单细胞数据中的性能。从图 4可以看出，错误预测即图中蓝色的点数量较少，说明经过迁移学习以后，模型的准确率高，能更好的适配于单细胞基因表达的数据集。

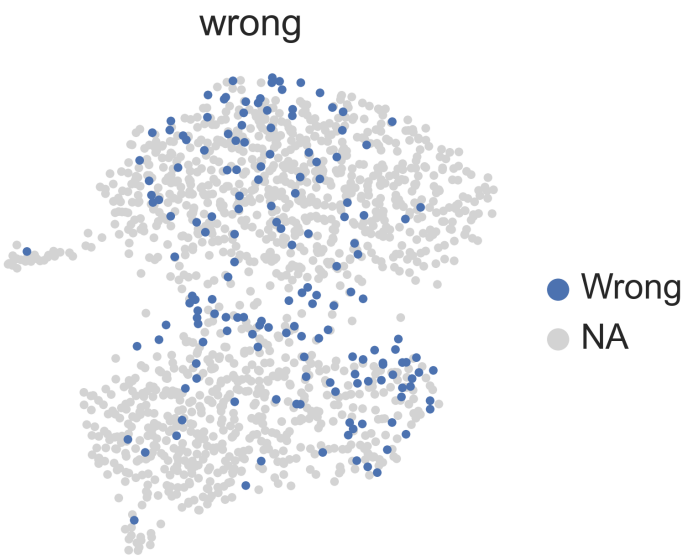


图 3. 错误预测高亮图

图 2中点与点之间的分布较为紧凑，只能从整体上看出各部分之间的差别，为了使预测结果细胞分布与真实分布的区别更加明显，我们使用图来进行比较。从图中可以发现，细胞在 UMAP 图中的分布整体上是一致的，除了在敏感细胞聚集区中会预测错一些细胞为抗性细胞，在抗性细胞聚集区会预测错一些细胞为敏感细胞，但整体上是不太影响预测的结果区分的。

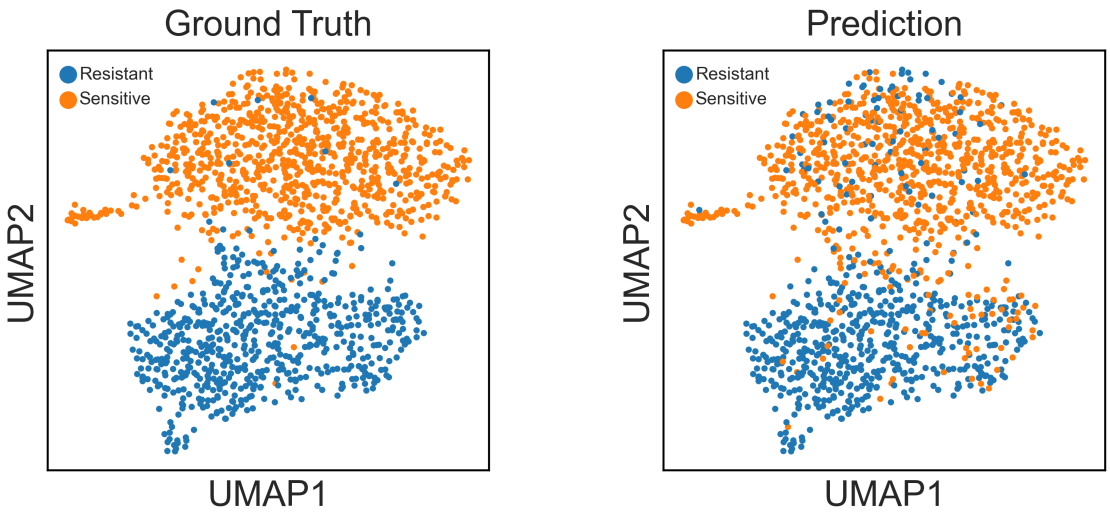


图 4. 结果比较图

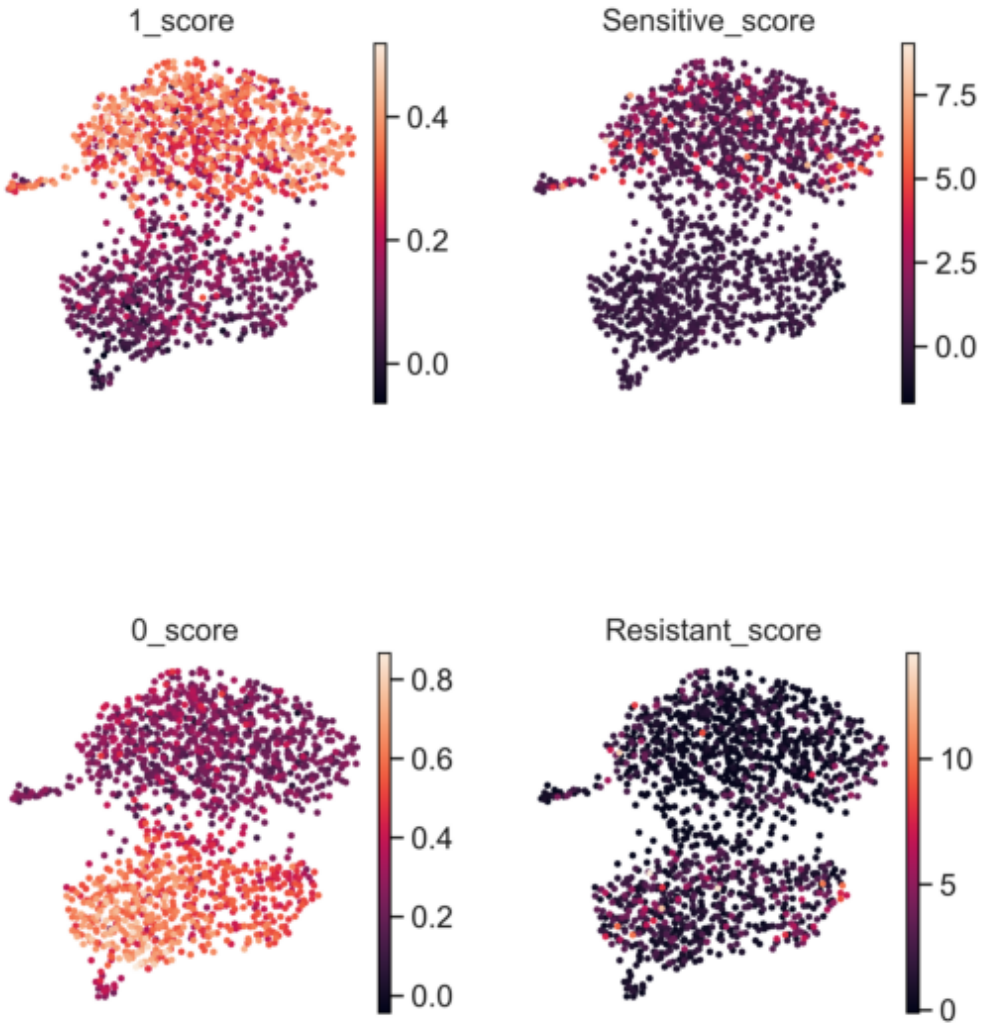


图 5. 基因得分图

在上面的内容中我们展示了敏感和抗药细胞的分布，接下来引入一个基因评分来反映在

敏感 (或抗性) 细胞簇中鉴定的差异表达基因的总体基因表达水平。如图 5所示, 上面两幅图展示的是细胞敏感性基因得分，下面两幅图是细胞耐药性基因得分。其中颜色越浅代表基因得分越高即基因表达水平越高，颜色越深代表基因得分越低即基因表达水平越低，因此在上面和下面分别代表敏感和耐药的图中，可以发现颜色是相反的，因为敏感基因表达水平越高意味着耐药基因的表达水平越低。

DEG 为差异表达基因 [10]。对于敏感的 DEG 列表，预测 DEG 分数与真实 DEG 分数之间的相关性高达 $R^2 = 0.89$ ，对于抗性的 DEG 列表，相关性高达 $R^2 = 0.66$。x 轴表示差异表达基因评分的经验相关性，y 轴表示频率，红色虚线表示 scDEAL 结果。这里进行了经验零模型检验来评估相关性的显著性。我们随机选择与预测 DEG 相同数量的基因，并按照上述 1000 次计算相关性。实证检验 (n = 1000) 结果显示，敏感和抗性 DEG 评分相关性的 p 值低于 0.001，表明相关性显著且具有统计学意义。

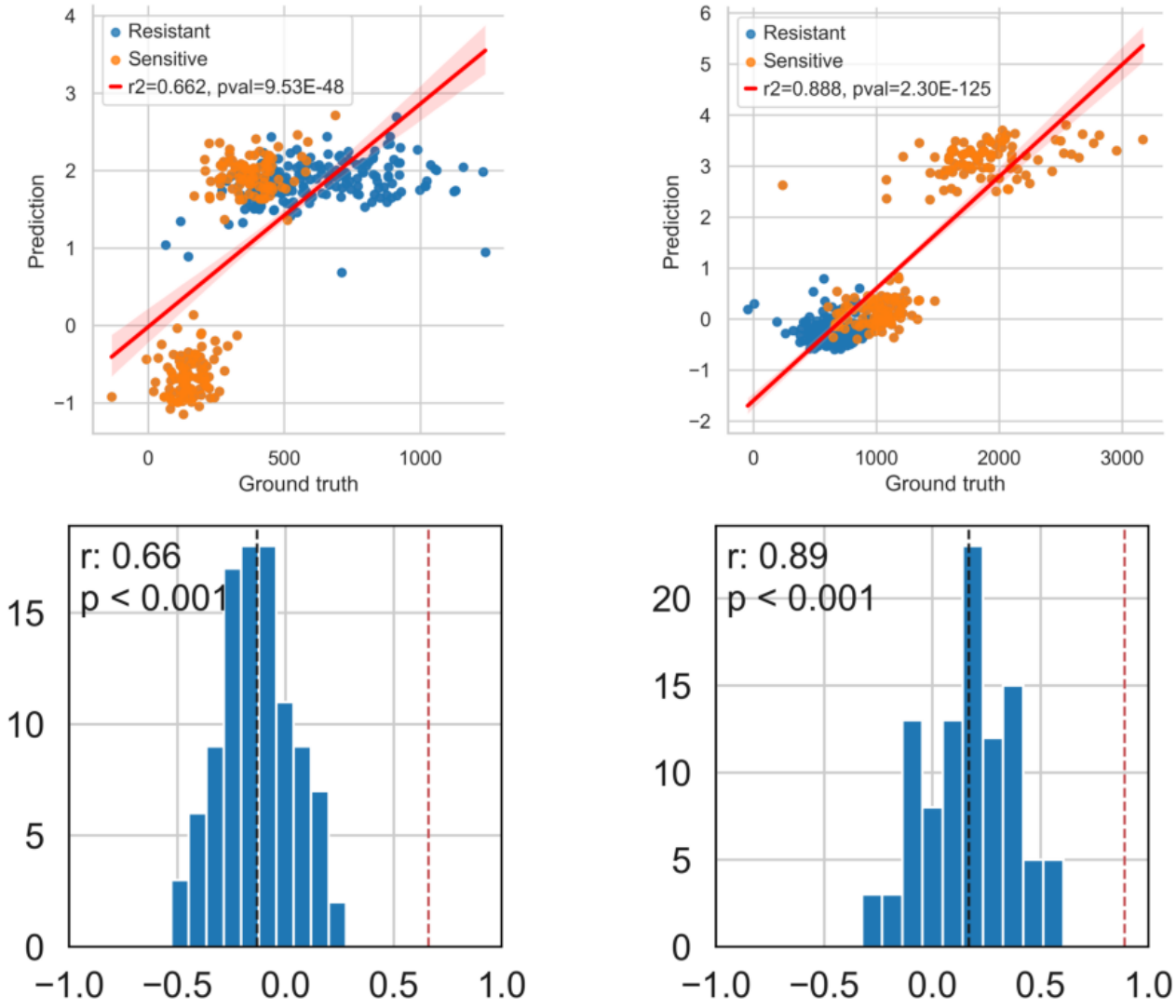


图 6. 数据相关性及 p 值

# 6 总结与展望

在前面五个章节里，分别介绍了复现的论文的不同内容。第一章节介绍复现论文的研究背景、意义以及我选择复现这篇论文的原因；第二章节我对本论文涉及的相关工作进行了介绍，例如 RNA 测序以及细胞异质性的作用；在第三章节中，对复现论文的模型框架进行了大致的介绍，同时介绍模型中使用到的低维特征提取模块以及深度迁移学习在该方法中发挥的作用；第四章节介绍了论文复现的细节，如复现环境、方法的创新点等；第五章节对我训练得到的模型进行结果分析，结果可视化等。

复现过程中发现模型验证部分使用的不是交叉验证 [3]，这可能造成模型在训练集上过度拟合，导致在新数据集上表现不佳，我将验证方式改为了五折交叉验证，并对修改前后的模型进行评价指标的分析，发现结果并无较大差异。

未来将进一步研究数据增强这部分，现有的单细胞基因表达数据增强方法是仅在单细胞层面的增强，没有考虑到 bulk 层面的数据增强，我们相信在对 bulk 基因表达数据进行学习并知识迁移以后，单细胞基因表达数据的增强效果会得到较大的提升。这是我目前的研究内容，目前已经做了一部分实验，预计在未来一段时间内对实验进行完善后，将研究成果投稿至 Bioinformatics 期刊。

# 参考文献


[1] Tianfeng Chai and Roland R Draxler. Root mean square error (rmse) or mean absolute error (mae). *Geoscientific model development discussions*, 7(1):1525–1534, 2014.

[2] Ibiayi Dagogo-Jack and Alice T Shaw. Tumour heterogeneity and resistance to cancer therapies. *Nature reviews Clinical oncology*, 15(2):81–94, 2018.

[3] Tadayoshi Fushiki. Estimation of prediction error by using k-fold cross-validation. *Statistics and Computing*, 21:137–146, 2011.

[4] Jian Hu, Xiangjie Li, Gang Hu, Yafei Lyu, Katalin Susztak, and Mingyao Li. Iterative transfer learning with neural network for clustering and cell type classification in single-cell rna-seq analysis. *Nature machine intelligence*, 2(10):607–618, 2020.

[5] Dmitry Kobak and George C Linderman. Initialization is critical for preserving global data structure in both t-sne and umap. *Nature biotechnology*, 39(2):156–157, 2021.

[6] Malte D Luecken and Fabian J Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019.

[7] Nemanja D Marjanovic, Robert A Weinberg, and Christine L Chaffer. Cell plasticity and heterogeneity in cancer. *Clinical chemistry*, 59(1):168–179, 2013.

[8] Corbin E Meacham and Sean J Morrison. Tumour heterogeneity and cancer cell plasticity. *Nature*, 501(7467):328–337, 2013.

[9] Amarinder Singh Thind, Isha Monga, Prasoon Kumar Thakur, Pallawi Kumari, Kiran Dindhoria, Monika Krzak, Marie Ranson, and Bruce Ashford. Demystifying emerging bulk rna-seq applications: the application and utility of bioinformatic methodology. *Briefings in bioinformatics*, 22(6):bbab259, 2021.

[10] S Udhaya Kumar, D Thirumal Kumar, R Bithia, Srivarshini Sankar, R Magesh, Mariem Sidenna, C George Priya Doss, and Hatem Zayed. Analysis of differentially expressed genes and molecular pathways in familial hypercholesterolemia involved in atherosclerosis: a systematic and bioinformatics approach. *Frontiers in Genetics*, 11:734, 2020.

[11] Juexin Wang, Anjun Ma, Yuzhou Chang, Jianting Gong, Yuexu Jiang, Ren Qi, Cankun Wang, Hongjun Fu, Qin Ma, and Dong Xu. scgnn is a novel graph neural network framework for single-cell rna-seq analyses. *Nature communications*, 12(1):1882, 2021.

[12] Zehao Xiong, Jiawei Luo, Wanwan Shi, Ying Liu, Zhongyuan Xu, and Bo Wang. scgcl: an imputation method for scrna-seq data based on graph contrastive learning. *Bioinformatics*, 39(3):btad098, 2023.

[13] Junlin Xu, Lijun Cai, Bo Liao, Wen Zhu, and JiaLiang Yang. Cmf-impute: an accurate imputation tool for single-cell rna-seq data. *Bioinformatics*, 36(10):3139–3147, 2020.

[14] Chu Zhang, Lei Zhou, Yiying Zhao, Susu Zhu, Fei Liu, and Yong He. Noise reduction in the spectral domain of hyperspectral images using denoising autoencoder methods. *Chemometrics and Intelligent Laboratory Systems*, 203:104063, 2020.