Adversarial Transfer Learning for Predicting Drug Sensitivity in Single-Cell Data

Huawei Zhang[1,2], Shaoshuai Zhu[1,2], Fei Lin[3], Qinghua Zhang[4], Fan Yang[2(✉)], and Qingke Zhang[1(✉)]

[1] School of Information Science and Engineering, Shandong Normal University, Jinan, China
tsingke@sdnu.edu.cn
[2] School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan, China
fanyang@sdu.edu.cn
[3] School of Continuing Education, Shandong University, Jinan, China
[4] Chongqing University of Posts and Telecommunications, Chongqing, China

Abstract. Drug resistance remains a primary cause of cancer treatment failure, while single-cell-level heterogeneity within the tumor microenvironment poses significant challenges for predicting drug sensitivity. This study proposes a framework for single-cell drug sensitivity prediction based on adversarial transfer learning. Our approach addresses the scarcity of annotated single-cell data by transferring drug response knowledge from bulk cell-line data (source domain) to single-cell data (target domain) through adversarial domain adaptation. We first establish a drug response classification model via supervised learning in the source domain, followed by adversarial alignment of feature distributions between domains. The experimental results demonstrate that our framework surpasses existing models in cross-domain transfer learning tasks, achieving 94% classification accuracy.

Keywords: adversarial domain adaptation · transfer learning · single cell · drug sensitivity

1 Introduction

Drug resistance and the resulting ineffectiveness of drug treatments lead to up to 90% of cancer-related deaths [1]. In addition to intrinsic resistance in tumor subpopulations, cancer cells may develop resistance through multiple mechanisms including pathway activation [2], alternative splicing [3], and drug efflux [4]. Within the tumor microenvironment, distinct cancer subpopulations exhibit marked heterogeneity in gene expression profiles and functional states, directly influencing drug response diversity. Consequently, there is an urgent need to predict drug sensitivity at single-cell resolution and identify causal genetic determinants of therapeutic outcomes [5]. Single-cell RNA sequencing (scRNA-seq) enables transcriptomic profiling at cellular resolution, providing new insights into

Q. Zhang et al. (Eds.): IJCRS 2025, LNAI 15710, pp. 242–256, 2025.
https://doi.org/10.1007/978-3-031-92741-6_18

tumor heterogeneity and drug resistance mechanisms. The field encounters challenges in acquiring drug sensitivity annotations at single-cell resolution, primarily due to the substantial experimental material requirements and prohibitively high equipment costs associated with single-cell sequencing technologies.

In transfer learning applications, negative transfer [6] represents an important challenge, characterized by performance degradation when transferring knowledge between domains. This phenomenon may arise from source domain overfitting (leading to poor target domain generalization) or inappropriate feature selection (transferring irrelevant or counterproductive features).

To address these challenges, this study proposes an adversarial transfer learning framework for single-cell drug sensitivity prediction. The proposed architecture employs a domain discriminator for adversarial training to enable domain-invariant feature learning, while simultaneously aligning feature distributions between domains using Maximum Mean Discrepancy (MMD) loss in the target domain encoder. This dual mechanism effectively mitigates negative transfer effects. Leveraging available drug sensitivity labels from bulk cell line data, our framework demonstrates successful knowledge transfer through adversarial domain adaptation, enabling accurate drug response predictions at single-cell resolution. Experimental validation shows that our framework achieves 94% classification accuracy in single-cell drug sensitivity prediction, significantly outperforming conventional transfer learning approaches.

The primary contributions of this work are as follows.

- This study introduces novel application of adversarial transfer learning to single-cell drug sensitivity prediction, effectively addressing the challenge of missing drug response annotations at cellular resolution. The adversarial learning mechanism in the target domain enhances transfer generalization through enriched feature encoding.
- In our target domain encoder, the Maximum Mean Discrepancy (MMD) distance is used to reduce the distributional disparities between the source and target domains, thereby facilitating the learning of domain invariance.
- Asymmetric feature mapping enables more accurate simulation of low-level feature discrepancies between domains, facilitating improved cross-domain feature space alignment (Fig. 1).

2 Related Work

The cellular heterogeneity unveiled through single-cell data analysis is crucial for improving the accuracy of drug sensitivity prediction and significantly promotes the design of combination drug strategies [7–9]. Although single-cell technology is relatively new, researchers have increasingly adopted these approaches to investigate disease heterogeneity. Single-cell RNA sequencing (scRNA-seq) has proven particularly valuable in uncovering transcriptional variations associated with drug resistance mechanisms [10]. It has been demonstrated in a study that scRNA-seq technology is capable of effectively characterizing and predicting the

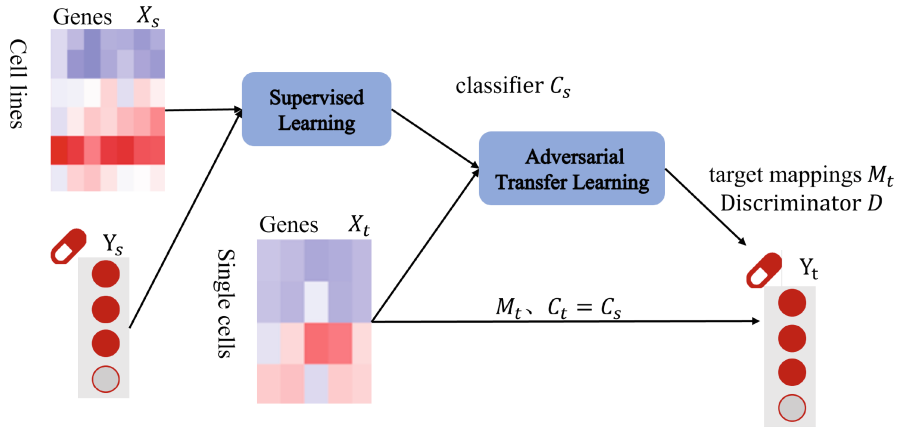


Fig. 1. Overall framework. The model obtains a classifier C_s through supervised learning on X_s (source domain), undergoes adversarial training on X_t (target domain) to obtain a target domain mapping M_t, and finally migrates the classifier C_s to single-cell data (the target domain) to predict drug sensitivity.

most aggressive tumor subpopulations in lung adenocarcinoma [11]. Comparative analyses between single-cell and conventional bulk sequencing approaches reveal that transcriptional signatures of drug-resistant subpopulations are frequently obscured by dominant cell populations, resulting in critical biological insights being masked. These findings underscore the necessity of transitioning from bulk tissue analysis to single-cell resolution when studying cellular heterogeneity [12]. The scope of single-cell research has expanded from simply identifying cell types to elucidating the biological mechanisms that lead to drug resistance in resistant subpopulations. Drug resistance can arise from both genetic [13] and non-genetic [14] factors. Genetic drug resistance usually stems from heritable mutations in cellular DNA, which make cells and their descendants resistant to existing drug treatments. The task of this paper is to predict drug sensitivity based on single-cell transcriptional data.

In the academic field of domain transfer learning, there have been a large number of previous research achievements [15–21]. For unlabeled target domains, the prevailing paradigm involves minimizing feature distribution discrepancies between domains to guide representation learning [22]. To achieve this goal, some methods have adopted the Maximum Mean Discrepancy (MMD) loss [23]. MMD measures the distribution difference by calculating the norm of the difference between the means of the two domains. The Deep Domain Confusion (DDC) method [17] introduces MMD on the basis of the conventional source domain classification loss to learn representations that are both discriminative and domain-invariant. The Deep Adaptation Networks (DAN) [24]apply MMD to layers embedded in the reproducing kernel Hilbert space, effectively matching the higher-order statistics of the two distributions. A novel architecture known as the Variational Auto-Encoder-Long-Short-Term-Memory Network-Local Weighted

Deep Sub-Domain Adaptation Network (VLSTM-LWSAN) [25] has been proposed for RUL prediction. Through the local weighted deep sub-domain adaptation network, this architecture compresses the input data into an interpretable latent space and aligns the fine-grained features across subdomains. A novel metric that integrates MMD and CORAL is introduced to amplify domain confusion [26]. Other methods choose adversarial losses to minimize domain shift and learn representations that can distinguish source labels but not domains. Literature [22] proposed adding a domain classifier that predicts the binary domain labels of the input (a single fully connected layer) and designed a domain confusion loss to encourage its predictions to be as close as possible to the uniform distribution of binary labels. The ReverseGrad algorithm [27] similarly views domain invariance as a binary classification task. However, it achieves this by inverting the gradients of the domain classifier, thereby directly increasing the classifier's loss. Researchers proposed CoCoGAN [28], a conditional coupled generative adversarial network, to capture the joint distribution of dual-domain samples for domain adaptation. It uses source-domain samples from a relevant task (RT) to provide high-level concepts approximating the target domain, and dual-domain samples from an irrelevant task (IRT) to capture shared correlations between the two domains. Recent studies [29] propose a novel framework that uses a reconstruction loss to align source and target distributions while preserving target domain labels. Due to the sparsity and heterogeneity of single-cell gene expression data, reconstruction loss may not be able to accurately measure the performance of the model. It is difficult to have a relatively stable and predictable structure as in image data.

3 Preliminaries

3.1 Transfer Learning

Before defining transfer learning, we first revisit the concepts of a domain and a task [30].

Domain: A domain $\mathcal{D}$ comprises two components: a feature space $\mathcal{X}$ and a marginal distribution $P(X)$. Formally, $\mathcal{D} = \{\mathcal{X}, P(X)\}$. Here, $\mathcal{X}$ represents an instance set, defined as $\mathcal{X} = \{\mathbf{x}_i \in \mathcal{X}, i = 1, \ldots, n\}$. There are two domains in this paper: the source domain $\mathcal{D}_s$ (cell line data) and the target domain $\mathcal{D}_t$ (single-cell data).

Task: A task $\mathcal{T}$ consists of a label space $\mathcal{Y}$ and a decision function f, expressed as $\mathcal{T} = \{\mathcal{Y}, f\}$. In my drug sensitivity classification prediction model, the output is the predicted conditional probability distribution of drug sensitivity classes for each instance. The decision function f is implicit and is learned from sample data. In this case, $f(\mathbf{x}_j) = \{P(y_k|\mathbf{x}_j)|y_k \in \mathcal{Y}, k = 1, \ldots, |\mathcal{Y}|\}$, where y_k represents the labels of the drug sensitivity class and $\mathbf{x}_j$ is the feature vector for the j-th instance.

In practice, a domain is typically observed through a set of instances, which may or not include label information. In the context of this paper, drug sensitivity labels for cell line data are easily obtainable $\mathcal{D}_S = \{(\mathbf{x}_i, y_i)|\mathbf{x}_i \in \mathcal{X}^S, y_i \in \mathcal{Y}^S, i = 1, \ldots, n^S\}$, while drug sensitivity labels for single-cell data are missing due to cost and other issues.

Transfer Learning: Given the source domain $\mathcal{D}_S = \{(\mathbf{x}_i^S, y_i^S)\}_{i=1}^{n_S}$, where $\mathbf{x}_i^S \in \mathcal{X}^S$ is the feature vector of the source domain, and $y_i^S \in \mathcal{Y}^S$ is the corresponding label; the target domain $\mathcal{D}_T = \{(\mathbf{x}_j^T, y_j^T)\}_{j=1}^{n_T}$, where $\mathbf{x}_j^T \in \mathcal{X}^T$ is the feature vector of the target domain, and $y_j^T \in \mathcal{Y}^T$ is the corresponding label (which may be partially unknown). The goal of transfer learning is to use the knowledge from the source domain $\mathcal{D}_S$ to learn a model for the target domain $\mathcal{D}_T$, thereby maximizing the predictive performance on the target domain.

3.2 Adversarial Losses

The adversarial losses train the adversarial discriminator using a standard classification loss $\mathcal{L}_{\text{adv}_D}$. However, they differ in the loss used to train the mapping $\mathcal{L}_{\text{adv}_M}$.

$$\mathcal{L}_{\text{adv}_M} = -\mathcal{L}_{\text{adv}_D}. \tag{1}$$

This optimization aligns with the true minimax objective for generative adversarial networks. Nevertheless, this objective can pose challenges, as the discriminator tends to converge rapidly in the early stages of training, leading to vanishing gradients. Creating two distinct optimization objectives, one for the generator component and one for the discriminator component. In this case, $\mathcal{L}_{\text{adv}_D}$ stays the same, while $\mathcal{L}_{\text{adv}_M}$ is modified to:

$$\mathcal{L}_{\text{adv}_M}(\mathbf{X}_s, \mathbf{X}_t, D) = -\mathbb{E}_{\mathbf{x}_t \sim \mathbf{X}_t}[\log D(M_t(\mathbf{x}_t))]. \tag{2}$$

This objective shares the same fixed-point characteristics with the minimax loss function, yet it offers more robust gradients for the target mapping. It is important to note that in this scenario, separate mappings are employed for the source and target domains, and M_t is trained adversarially. This method is similar to the GAN [31] framework, in which the source distribution stays unchanged while the generator's distribution is trained to match it.

The GAN loss function is typically used when the generator aims to replicate a static distribution. However, when both distributions are dynamic, this goal may cause instability—once the mapping reaches its optimal state, the discriminator may simply invert its prediction. To address this, a study [22] introduced the domain confusion objective. With this objective, the mapping is trained using a cross-entropy loss function based on a uniform distribution.

$$\mathcal{L}_{\text{adv}_M}(\mathbf{X}_s, \mathbf{X}_t, D) = \\ -\sum_{d \in \{s,t\}} \mathbb{E}_{\mathbf{x}_d \sim \mathbf{X}_d} \left[\frac{1}{2} \log D(M_d(\mathbf{x}_d)) + \frac{1}{2} \log(1 - D(M_d(\mathbf{x}_d))) \right]. \tag{3}$$

This loss function plays a crucial role in ensuring that the adversarial discriminator perceives the two domains as identical.

4 Adversarial Transfer Learning

The framework of this paper consists of two main components: (1) Supervised learning, where a model is established to predict response label classification at the source domain batch level (cell-lines). (2) Adversarial learning, the discriminator determines whether a sample is from the source domain or the target domain, and the target mapping ensures the mapping space of the target domain as close as possible to the mapping space of the source domain. The classifier are transferred to the single-cell level for label prediction (Fig. 2).

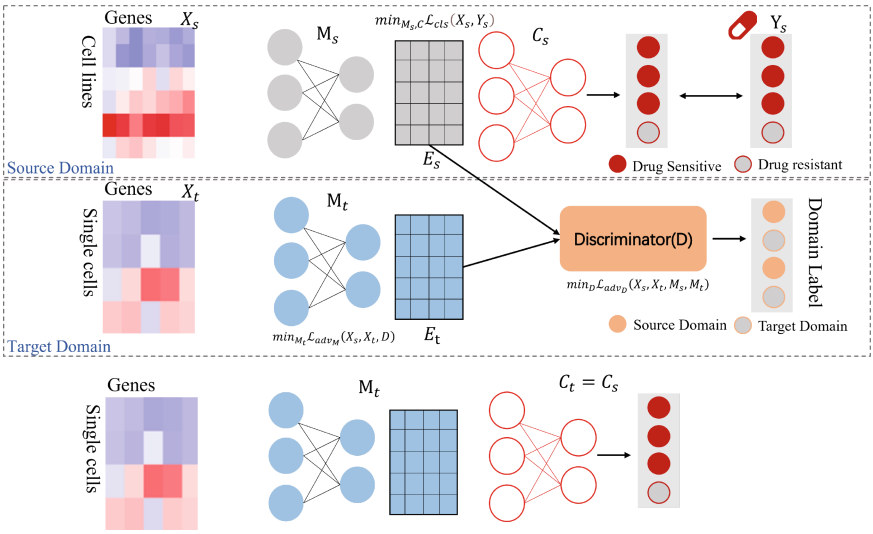


Fig. 2. Adversarial transfer learning framework. During the adversarial learning phase, the target domain encoder learns to generate mappings that are as similar as possible to the source domain feature space, while the discriminator serves to distinguish the origin of the features, identifying them as coming from either the source domain or the target domain. Both are trained adversarially until the discriminator can no longer distinguish.

In unsupervised and adaptive learning, it is assumed that cell line gene expression data X_s and drug sensitivity labels Y_s are extracted from the source domain distribution $p_s\,(x, y)$, and single-cell gene expression data X_t is extracted from the target distribution $p_t\,(x, y)$, with no label observations. The objective is to develop a target representation M_t and classifier C_t to classify target data into 2 classes, only sensitive and insensitive cases. Given that direct supervised

learning on the target domain is infeasible, our method learns a source representation mapping M_s and a source classifier C_s. Subsequently, it adapts this model to fit the target domain.

In the adversarial domain adaptation approach, the primary objective is to regulate the learning process of the source mappings M_s and target mappings M_t. This study opts for a design where the source and target mappings are independent, achieved through the use of non-shared weights. This strategy offers greater flexibility, as it enables the model to learn more domain-specific feature extraction techniques. Given that the target domain lacks labels, and without weight sharing, improper initialization or training could lead the target model to quickly converge on suboptimal solutions. To address this, the study employs a pre-trained source model as the initial basis for the target representation space. During adversarial training, the source model is kept fixed, while M_t makes efforts to minimize the distance between the empirical distributions of the source mappings $M_s\left(X_s\right)$ and target mapping $M_t\left(X_t\right)$.

Supervised Learning in the Source Domain. The source classification model C_s is applicable to the target representation directly, eliminating the necessity to develop a distinct target classifier, so it is set that $C = C_s = C_t$.

$$min_{M_s,C}\mathcal{L}_{cls}\left(X_s, Y_s\right) = \mathbb{E}_{(x_s,y_s)\backsim(X_s,Y_s)} - \sum_{k=1}^{K} 1_{[k=y_s]} logC(M_s(x_s)). \quad (4)$$

Adversarial Learning in the Target Domain. The domain discriminator D discriminates whether the data point is from the source domain or the target domain, while the target mapping tries to make the target domain data as close as possible to the source domain data mapping, and they are updated through adversarial training. Therefore, it is optimized according to the standard supervised loss $\mathcal{L}_{adv_D}\left(X_s, X_t, M_s, M_t\right)$, and M_t is optimized according to the loss $\mathcal{L}_{adv_M}\left(X_s, X_t, D\right)$, defined as follows:

$$\begin{aligned}\min_D \mathcal{L}_{adv_D}(X_s, X_t, M_s, M_t) = \\ - \mathbb{E}_{x_s\sim X_s}[\log D(M_s(X_s))] - \mathbb{E}_{x_t\sim X_t}[\log(1 - D(M_t(X_t)))]. \quad (5)\end{aligned}$$

$$\begin{aligned}&\mathcal{L}_{\mathrm{adv}_M}(\mathbf{X}_s, \mathbf{X}_t, D) = \\ &- \sum_{d\in\{s,t\}} \mathbb{E}_{\mathbf{x}_d\sim\mathbf{X}_d}\left[\frac{1}{2}\log D(M_d(\mathbf{x}_d)) + \frac{1}{2}\log(1 - D(M_d(\mathbf{x}_d)))\right] + MMD(X_s, X_t).\end{aligned} \quad (6)$$

The distance between two distributions is defined as:

$$MMD(X,Y) = \left\| \frac{1}{n}\sum_{i=1}^{n}\Phi(x_i) - \frac{1}{m}\sum_{j=1}^{m}\Phi(y_j) \right\|_H^2 . \quad (7)$$

The target representation M_t and classifier C_t, obtained through the aforementioned steps, are applied to the single-cell RNA sequencing data X_t to derive Y_t, which represents the predicted results of single-cell drug sensitivity.

5 Data and Preprocessing

Data: The dataset from the GDSC database[1] includes drug response annotations such as area under the dose-response curve (AUC) and half-maximal inhibitory concentration (IC50). Additionally, gene expression data for cell lines (RMA-normalized base expression profiles) is also available on the GDSC website. This paper additionally gathered the CCLE cell line expression profiles and the PRISM cell line viability assays[2]. The GDSC database and the CCLE database both provide gene expression data and drug sensitivity information. To ensure the consistency and comparability of the data when integrating the two databases, it is crucial to retain shared genes as well as cell lines and drugs that have no missing values. Although some information may be lost during the integration process, the proportion of this lost information is relatively small in the overall dataset, and its impact on the experimental results is limited. Overall, this paper collected drug response data with 1280 cancer cell-lines, 15,962 genes and 83 drugs. All scRNA-seq data can be obtained from Gene Expression Omnibus (GEO)[3] (Table 1).

Table 1. Description of Datasets.

Data	Samples	Features	Introduction
GSE110894	1504	25921	Leukemia cells treated with the I.ET.762 inhibitor
GSE149383	2740	18380	PC9 lung cancer cells treated with ERLOTINIB
GSE140440	324	60555	Prostate cancer cells treated with DOCETAXEL

Preprocessing [32]**:** The classification of drug sensitivity for cell lines in batch datasets is based on AUC values obtained from the GDSC and CCLE databases. For each drug, the AUC scores are converted into binary labels to indicate whether the cell lines are sensitive or resistant to the drug. Cell lines that respond positively to a particular drug are assigned a label of 1, whereas those that are

[1] https://www.cancerrxgene.org/.
[2] https://depmap.org/portal/.
[3] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi.

resistant to the drug are assigned a label of 0. To organize this data, a waterfall approach is employed to arrange the AUC values of the cell lines from highest to lowest, creating an AUC-cell line curve. In this graphical representation, the x-axis corresponds to the different cell lines, while the y-axis indicates the AUC values. The AUC cutoff value is established using two distinct methods: First, for linear curves—characterized by a Pearson correlation coefficient of the regression line exceeding 0.95—the AUC cutoff between sensitivity and resistance is set at the median AUC value across all cell lines. Second, for non-linear curves, the cutoff is determined by identifying the AUC value of a boundary data point that lies farthest from the line segment connecting the data points with the highest and lowest AUC values.

Using SCANPY53 [33], we conducted quality control and preprocessing of the scRNA-seq data. Specifically, the "filter cells" and "filter genes" functions are used to filter out cells with less than 200 detected genes (indicating no cells in the droplet) and genes that can only be detected in less than 3 cells. Cells with a mitochondrial gene expression percentage higher than 10% are removed. Then, the expression values are scaled using the "preprocessingMinMaxScaler" from the sklearn [34] package. After preprocessing and filtering, the GSE110894 dataset retained 1,135 cells and 3,935 genes; the GSE149383 dataset retained 1,947 cells and 3,097 genes; the GSE140440 dataset retained 324 cells and 10,906 genes.

6 Results

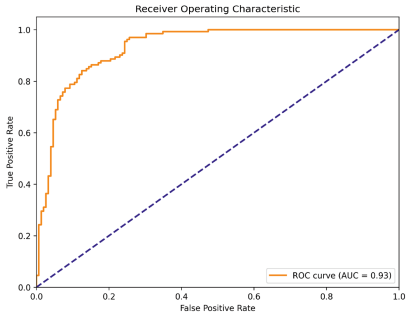
To ensure the robustness of our evaluation, we employed 5-fold cross-validation [35]. In this method, the dataset is randomly divided into five equal parts. For each iteration, one part is held out as the validation set while the remaining four parts are used for training the model. This process is repeated five times, with each of the five parts used exactly once as the validation set. The performance metrics reported are the average values obtained from these five iterations, providing a more reliable and generalized assessment of our model's performance. This paper uses ACC (Accuracy), AUC (Area Under Curve), Pre (Precision), Recall, and F1 Score as metrics to evaluate model performance. The calculation formulas for each evaluation metrics are as follows.

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}, \tag{8}$$

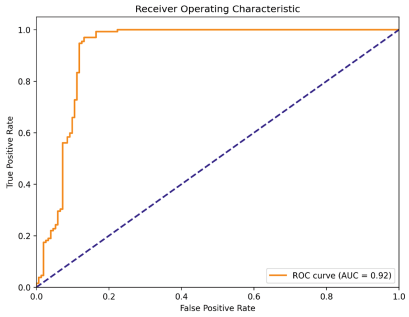
where TP denotes the number of samples that are truly sensitive and correctly predicted as sensitive. TN indicates the number of samples that are truly drug-resistant and correctly predicted as drug-resistant. FP refers to the number of samples that are actually drug-resistant but incorrectly predicted as sensitive. FN signifies the number of samples that are truly sensitive but incorrectly predicted as drug-resistant. Accuracy measures the ratio of correctly classified samples to the total number of samples.

The Area Under the Curve (AUC) serves as a metric for evaluating the performance of classification models across different threshold settings. It quantifies

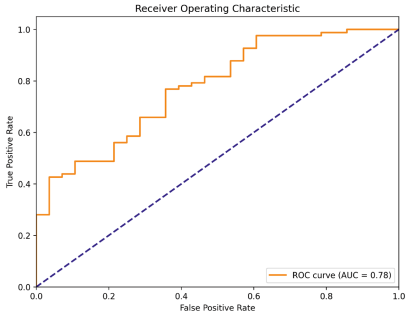
a model's capacity to differentiate between classes and is visually represented by the area beneath the Receiver Operating Characteristic (ROC) curve. The ROC curves of our method on the three datasets are shown in Fig. 3. The ROC curve is constructed by plotting the true positive rate (TPR) against the false positive rate (FPR) at various thresholds. An AUC of 1.0 signifies a flawless model, whereas an AUC of 0.5 indicates a model that performs no better than random guessing. Unlike the other two datasets, GSE140440 has only 324 samples and 10,906 features post - processing. This high dimensionality and small sample size somewhat restrict the model's performance. Nevertheless, comparative experiments demonstrate that our model still surpasses others.



(a) GSE110894



(b) GSE149383



(c) GSE140440

Fig. 3. The ROC curve of the model for GSE110894, GSE149383 and GSE140440.

$$\text{Pre} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{9}$$

Precision describes the proportion of positive identifications that were actually correct. It measures the accuracy of the positive predictions made by the model.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{10}$$

Recall describes the proportion of actual positives that were correctly identified. It measures the model's ability to find all relevant cases within the data.

$$\text{F1 Score} = 2 \times \frac{\text{Pre} \times \text{Recall}}{\text{Pre} + \text{Recall}}. \tag{11}$$

The F1 Score is the harmonic mean of Precision and Recall, used to measure a test's accuracy. It provides a single measure that balances both precision and recall, and is particularly useful when dealing with imbalanced datasets (Tables 2, 3 and 4).

Table 2. Results of different model using GSE110894.

Model	ACC	AUC	Pre	Recall	F1score
Source only	0.68	0.66	0.80	0.40	0.54
DANN	0.73	0.81	0.74	0.69	0.71
Domain Confusion	0.90	0.92	0.91	0.89	0.89
our proposed	0.94	0.94	0.93	0.95	0.94

Table 3. Results of different model using GSE149383.

Model	ACC	AUC	Pre	Recall	F1score
Source only	0.58	0.63	0.60	0.56	0.50
DANN	0.80	0.89	0.81	0.76	0.77
Domain Confusion	0.81	0.88	0.80	0.78	0.79
our proposed	0.93	0.93	0.90	0.95	0.92

Table 4. Results of different model using GSE140440.

Model	ACC	AUC	Pre	Recall	F1score
Source only	0.45	0.44	0.41	0.35	0.38
DANN	0.64	0.63	0.60	0.56	0.58
Domain Confusion	0.73	0.76	0.74	0.69	0.69
our proposed	0.80	0.78	0.80	0.78	0.79

To verify the effectiveness of the method proposed in this paper, we compared it with three other models: source only model, DANN, and Domain Confusion, across three different target domain datasets. The results indicate that

our method significantly outperforms the other models in all performance metrics.

Taking the GSE110894 dataset as an example, the accuracy (ACC) of our method reached 0.94, which is significantly higher than that of source only model at 0.68 and the DANN model at 0.73, with improvements of 0.26 and 0.21 respectively ($p < 0.01$). Additionally, compared to the Domain Confusion model, our method also scored 0.04 higher ($p < 0.05$). The p-values were calculated using Welch's t-test.

(1) Source only: The training of this model relies solely on data sourced from the source domain. It does not attempt to make any adjustments or optimizations for the data in the target domain, so it can serve as a benchmark for evaluating the performance of other domain adaptation methods. Therefore, the performance on the GSE140440 dataset is relatively poor, while it is comparatively better on the GSE110894 dataset, due to the latter's data dimension being closer to that of the source domain. (2) DANN (Domain Adversarial Training of Neural Networks) [15]: The main idea of DANN is the ReverseGrad algorithm, which treats domain invariance as a binary classification problem, but by reversing the gradient of the domain discriminator, it directly maximizes the loss of the domain classifier. Adversarial training itself is relatively complex; a simple gradient reversal essentially provides convenience in coding, but it is extremely unstable during the training process. (3) Domain confusion: The Deep Domain Confusion (DDC) method [17] introduces the MMD on the basis of the conventional source domain classification loss to learn representations that are both discriminative and domain-invariant, mapping the features of the source and target domains together into a new shared space for symmetric mapping. The DANN and Domain Confusion models can alleviate the issues caused by dimensionality differences to some extent, thereby enhancing the model's performance on the target domain. However, on the GSE140440 dataset, due to the significant dimensionality differences, the improvement in performance is limited. (4) Ours: This paper's method first learns discriminative representations using labels in the source domain, and then maps the target data to the source feature space using an asymmetric mapping learned through domain adversarial loss, which can better simulate the differences in low-level features than symmetric mapping. The method proposed in this paper demonstrates the best performance across all datasets, indicating that the method has good adaptability to differences in the number of samples and genes. It can effectively handle high-dimensional, small-sample data, thereby achieving good results on various datasets.

Parameters. This paper maps single-cell gene expression data and cell line gene expression data to a 512-dimensional latent space, which is passed to the discriminator for adversarial learning. Gene expression data is high-dimensional, which makes computations complex and prone to overfitting. A dimensionality of 512 reduces the complexity of the data while retaining sufficient feature information, which helps the model learn the essential characteristics of the data. This dimension has demonstrated good model performance and stability in preliminary

experiments. During the training process, this dimension allows the model to effectively learn the feature representations of both the source and target domain data, and achieve feature distribution alignment through adversarial learning, thereby improving the model's predictive accuracy. The dimensions of the layers of the source domain neural network are [4096,2048,1024,512], and a pre-trained source model is used as the initialization of the target representation space. And the source model is fixed during adversarial training. The number of layers and dimensions of the target domain neural network need to be adjusted according to the dimensions of the preprocessed data. The dropout for the target domain encoder is set to 0.1, and the dropout for the discriminator is set to 0.2. This is because, during the initial stages of adversarial training, the discriminator learns quickly, which can lead to the target domain encoder failing to learn useful knowledge, thereby affecting the model's performance. An appropriate dropout value can prevent overfitting to some extent and ensure the model's generalization capability. The learning rate is set to 1×10^{-4}, which is a relatively small learning rate that helps the model to converge stably during training. Especially in the early stages of training, a smaller learning rate can prevent drastic fluctuations in model parameters, allowing the model to gradually learn the complex patterns and features in the data.

7 Conclusions and Outlook

In this study, we proposed a novel framework based on adversarial transfer learning for predicting single-cell drug sensitivity. One of the key innovations of our framework is the adversarial learning mechanism, which ensures that the feature distributions of the source and target domains are aligned. This alignment is crucial for the successful transfer of knowledge from the labeled cell line data to the unlabeled single-cell data. Through our experiments, we have demonstrated that our model outperforms existing methods in terms of prediction accuracy and robustness.

Despite the progress made in this study, there are several avenues for future work that we are actively exploring. One of the primary directions is the integration of multi-modal data. Currently, our framework relies primarily on single-cell RNA sequencing data; however, incorporating additional modalities such as proteomics, metabolomics and imaging data could provide a more comprehensive understanding of the cellular mechanisms underlying drug resistance. Multi-modal data integration has the potential to reveal hidden relationships between different biological layers and enhance the accuracy and reliability of causal gene identification.

Another promising direction is the extension of our framework to handle multi-class drug sensitivity predictions. Currently, our model focuses on binary classification of drug sensitivity. Expanding this to multi-class classification would allow for the prediction of varying levels of sensitivity to multiple drugs, providing a more detailed and nuanced understanding of drug responses at the single-cell level. This could lead to more effective and personalized treatment strategies in the future.

References

1. Wang, X., Zhang, H., Chen, X.: Drug resistance and combating drug resistance in cancer. Cancer Drug Resist. **2**(2), 141 (2019)
2. Siegfried, Z., Karni, R.: The role of alternative splicing in cancer drug resistance. Curr. Opin. Genet. Dev. **48**, 16–21 (2018)
3. Konieczkowski, D.J., Johannessen, C.M., Garraway, L.A.: A convergence-based framework for cancer drug resistance. Cancer Cell **33**(5), 801–815 (2018)
4. Panda, M., Biswal, B.K.: Cell signaling and cancer: a mechanistic insight into drug resistance. Mol. Biol. Rep. **46**(5), 5645–5659 (2019). https://doi.org/10.1007/s11033-019-04958-6
5. Ramsey, J., Glymour, M., Sanchez-Romero, R., Glymour, C.: A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. Int. J. Data Sci. Anal. **3**, 121–129 (2017)
6. Zhang, W., Deng, L., Zhang, L., Dongrui, W.: A survey on negative transfer. IEEE/CAA J. Automatica Sinica **10**(2), 305–329 (2022)
7. Gambardella, G., Viscido, G., Tumaini, B., Isacchi, A., Bosotti, R., Di Bernardo, D.: A single-cell analysis of breast cancer cell lines to study tumour heterogeneity and drug response. Nat. Commun. **13**(1), 1714 (2022)
8. Schmidt, F., Efferth, T.: Tumor heterogeneity, single-cell sequencing, and drug resistance. Pharmaceuticals **9**(2), 33 (2016)
9. Van de Sande, B., et al.: Applications of single-cell rna sequencing in drug discovery and development. Nat. Rev. Drug Disc. **22**(6), 496–520 (2023)
10. Shapiro, E., Biezuner, T., Linnarsson, S.: Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat. Rev. Genet. **14**(9), 618–630 (2013)
11. Kim, K.-T., et al.: Single-cell mrna sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. Genome Biol. **16**, 1–15 (2015)
12. Kempa, E.E., Hollywood, K.A., Smith, C.A., Barran, P.E.: High throughput screening of complex biological samples with mass spectrometry-from bulk measurements to single cell analysis. Analyst **144**(3), 872–891 (2019)
13. Goldie, J.H., Coldman, A.J.: The genetic origin of drug resistance in neoplasms: implications for systemic therapy. Cancer Res. **44**(9), 3643–3653 (1984)
14. Marine, J.-C., Dawson, S.-J., Dawson, M.A.: Non-genetic mechanisms of therapeutic resistance in cancer. Nat. Rev. Cancer **20**(12), 743–756 (2020)
15. Ganin, Y., et al.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17**(59), 1–35 (2016)
16. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: International Conference on Machine Learning, pp. 97–105. PMLR (2015)
17. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: maximizing for domain invariance. arXiv preprint arXiv:1412.3474 (2014)
18. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Mach. Learn. **79**, 151–175 (2010)
19. Sun, B., Feng, J., Saenko, K.: Return of frustratingly easy domain adaptation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30 (2016)
20. Fang, Y., Yap, P.-T., Lin, W., Zhu, H., Liu, M.: Source-free unsupervised domain adaptation: a survey. Neural Netw. **174**, 106230 (2024)

21. Zhao, Z., Alzubaidi, L., Zhang, J., Duan, Y., Yuantong, G.: A comparison review of transfer learning and self-supervised learning: definitions, applications, advantages and limitations. Expert Syst. Appl. **242**, 122807 (2024)
22. Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4068–4076 (2015)
23. Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.: Covariate shift and local learning by distribution matching. In: Dataset Shift in Machine Learning, pp. 131–160 (2008)
24. Long, M., Cao, Y., Wang, J., Jordan, M.I.: Learning transferable features with deep adaptation networks. arXiv: Learning (2015)
25. Zhenyu, W., Zhang, H., Guo, J., Ji, Y., Pecht, M.: Imbalanced bearing fault diagnosis under variant working conditions using cost-sensitive deep domain adaptation network. Expert Syst. Appl. **193**, 116459 (2022)
26. Qian, Q., Qin, Y., Luo, J., Wang, Y., Fei, W.: Deep discriminative transfer learning network for cross-machine fault diagnosis. Mech. Syst. Signal Process. **186**, 109884 (2023)
27. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International Conference on Machine Learning, pp. 1180–1189. PMLR (2015)
28. Wang, J., Jiang, J.: Conditional coupled generative adversarial networks for zero-shot domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3375–3384 (2019)
29. Sun, Y., Yang, G., Ding,D., Cheng, G., Xu, J., Li, X.: A gan-based domain adaptation method for glaucoma diagnosis. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2020)
30. Panigrahi, S., Nanda, A., Swarnkar, T.: A survey on transfer learning. In: Mishra, D., Buyya, R., Mohapatra, P., Patnaik, S. (eds.) Intelligent and Cloud Computing. SIST, vol. 194, pp. 781–789. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-5971-6_83
31. Goodfellow, I., et al.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020)
32. Chen, J., et al.: Deep transfer learning of cancer drug responses by integrating bulk and single-cell rna-seq data. Nat. Commun. **13**(1), 6494 (2022)
33. Wolf, F.A., Angerer, P., Theis, F.J.: SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. **19**, 1–5 (2018)
34. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
35. Fushiki, T.: Estimation of prediction error by using k-fold cross-validation. Stat. Comput. **21**, 137–146 (2011)