Report on "Deep Transfer Learning of Cancer Drug Responses by Integrating Bulk and Single-cell RNA-seq Data"

a. Problem Addressed

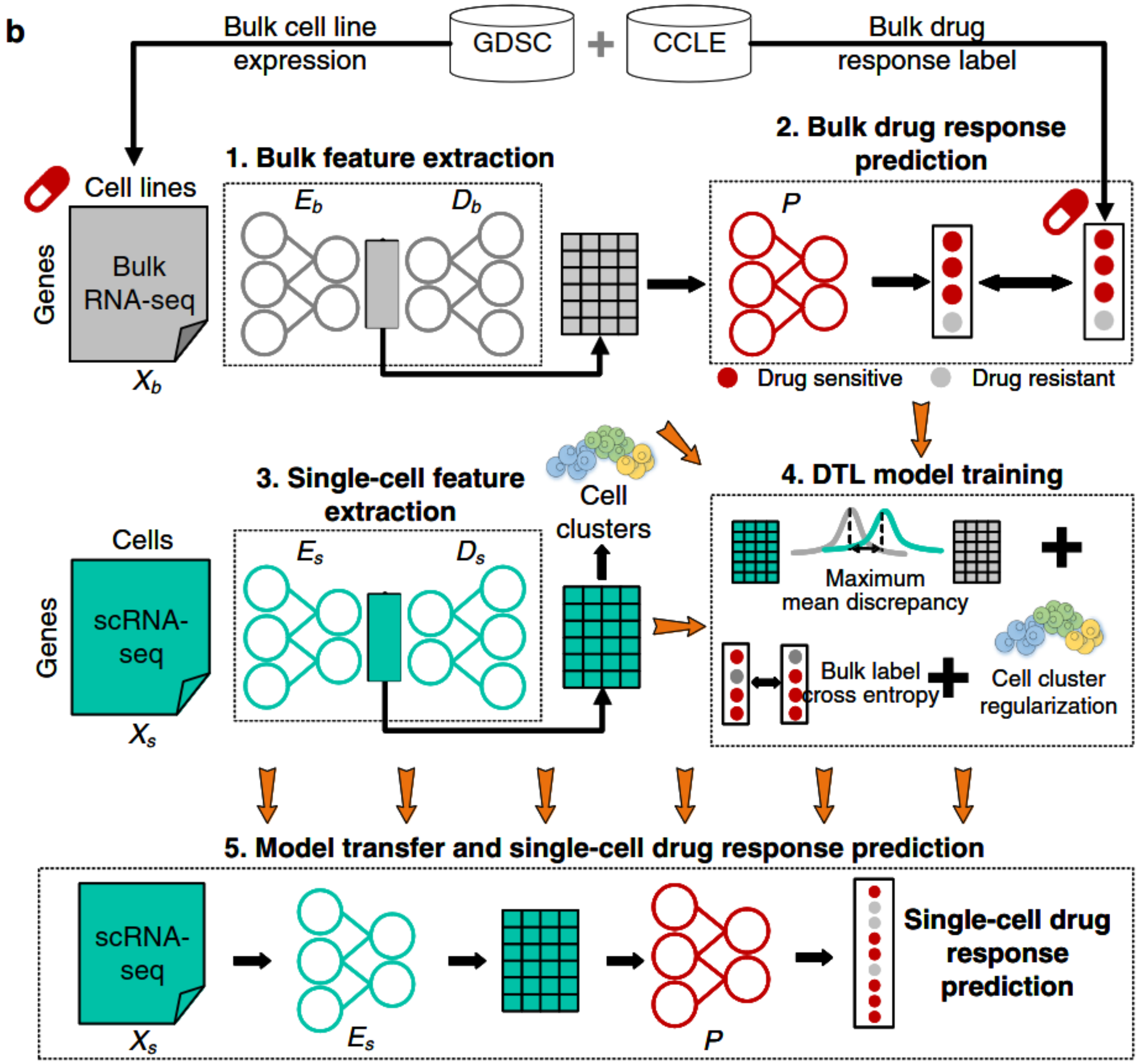
The article communicates the challenge of predicting cancer drug responses at the single-cell level by integrating bulk RNA-seq data with scRNA-seq data. The primary problem is the heterogeneity of cancer cells, which leads to varied responses to drugs, resulting in low treatment efficacy and high relapse rates. Current bulk RNA-seq-based prediction methods fail to capture this heterogeneity, necessitating a novel approach to accurately predict drug responses at the single-cell level.

b. Related Works

The main obstacle in developing deep learning-based tools for predicting single-cell drug responses is the limited number of benchmarked single-cell data available in the public domain. Traditional approaches have primarily relied on bulk RNA-seq data, which, while extensive, fail to capture the cellular heterogeneity that impacts drug responses. For example, studies like those by Barretina et al. (2012) and Yang et al. (2013) used bulk RNA-seq data from databases such as the Genomics of Drug Sensitivity in Cancer (GDSC) and the Cancer Cell Line Encyclopedia (CCLE) to develop predictive models of drug responses. These models typically employed machine learning techniques to correlate gene expression profiles with drug sensitivity or resistance across different cancer cell lines.

c. Proposed Solution

The authors propose **scDEAL** (single-cell Drug rEsponse AnaLysis), a deep transfer learning framework designed to predict drug responses at the single-cell level by integrating bulk RNA sequencing (RNA-seq) data with single-cell RNA sequencing (scRNA-seq) data. The key components of scDEAL and its core concepts are illustrated and described in several figures within the article.



- **1. Bulk Feature Extraction:**
  Input Data: Bulk RNA-seq data from the GDSC and CCLE databases are used. These databases provide comprehensive bulk cell line expression profiles and corresponding drug response labels.
  Processing: Gene features are extracted from the bulk RNA-seq data using a denoising autoencoder (DAE) consisting of an encoder E_b and a decoder D_b.

This step reduces the data dimensionality and noise, resulting in a low-dimensional representation of gene expression features X_b. The noise operation introduces random noise to the input data to improve the robustness of the feature extraction.

- **2. Bulk Drug Response Prediction:**
  Input Data: The processed bulk gene features X_b are used.
  Processing: A predictor model P is trained on the bulk data to correlate gene expression profiles with drug responses. The model learns to classify each cell line as drug-sensitive or drug-resistant based on the provided labels. The classification is optimized using cross-entropy loss, which measures the difference between the predicted and actual drug response labels.
- **3. Single-cell Feature Extraction:**
  Input Data: Single-cell RNA-seq (scRNA-seq) data.
  Processing: Similar to the bulk data, a denoising autoencoder (DAE) is used to extract low-dimensional gene features from the scRNA-seq data. The encoder E_s and decoder D_s create a noise-reduced, low-dimensional representation of single-cell gene expression features X_s. This step also involves the introduction of noise to handle the variability and noise characteristics of single-cell data.
- **4. DTL Model Training:**
  Processing: The Domain-adaptive Neural Network (DaNN) is trained to align the feature spaces of bulk and single-cell data. This involves:
  Maximum Mean Discrepancy (MMD) Loss: Measures the similarity between bulk and single-cell features to ensure that the knowledge from bulk data can be transferred to single-cell data.
  Cross-entropy Loss: Used to optimize the accuracy of drug response predictions by comparing predicted labels to actual labels.
  Regularization with Cell Clusters: Incorporates cell clustering information to maintain the heterogeneity of single-cell data during training. This ensures that the model does not oversimplify the diverse gene expression profiles of individual cells.
  Denoising Autoencoders (DAE): Both bulk and single-cell data are processed through DAEs to handle the distinct noise characteristics in these datasets and to ensure robust feature extraction.

- **5. Model Transfer and Single-cell Drug Response Prediction:**

  Processing: The trained model, which now contains generalized knowledge from bulk data, is transferred to single-cell data. The single-cell RNA-seq data X_s are processed through the trained single-cell encoder E_s and predictor P, enabling the prediction of drug responses at the single-cell level. The output is the predicted drug response for individual cells, classified as drug-sensitive or drug-resistant.

- **Performance Metrics:** The performance of scDEAL is evaluated using several metrics:

  F1-score: Measures the balance between precision and recall.

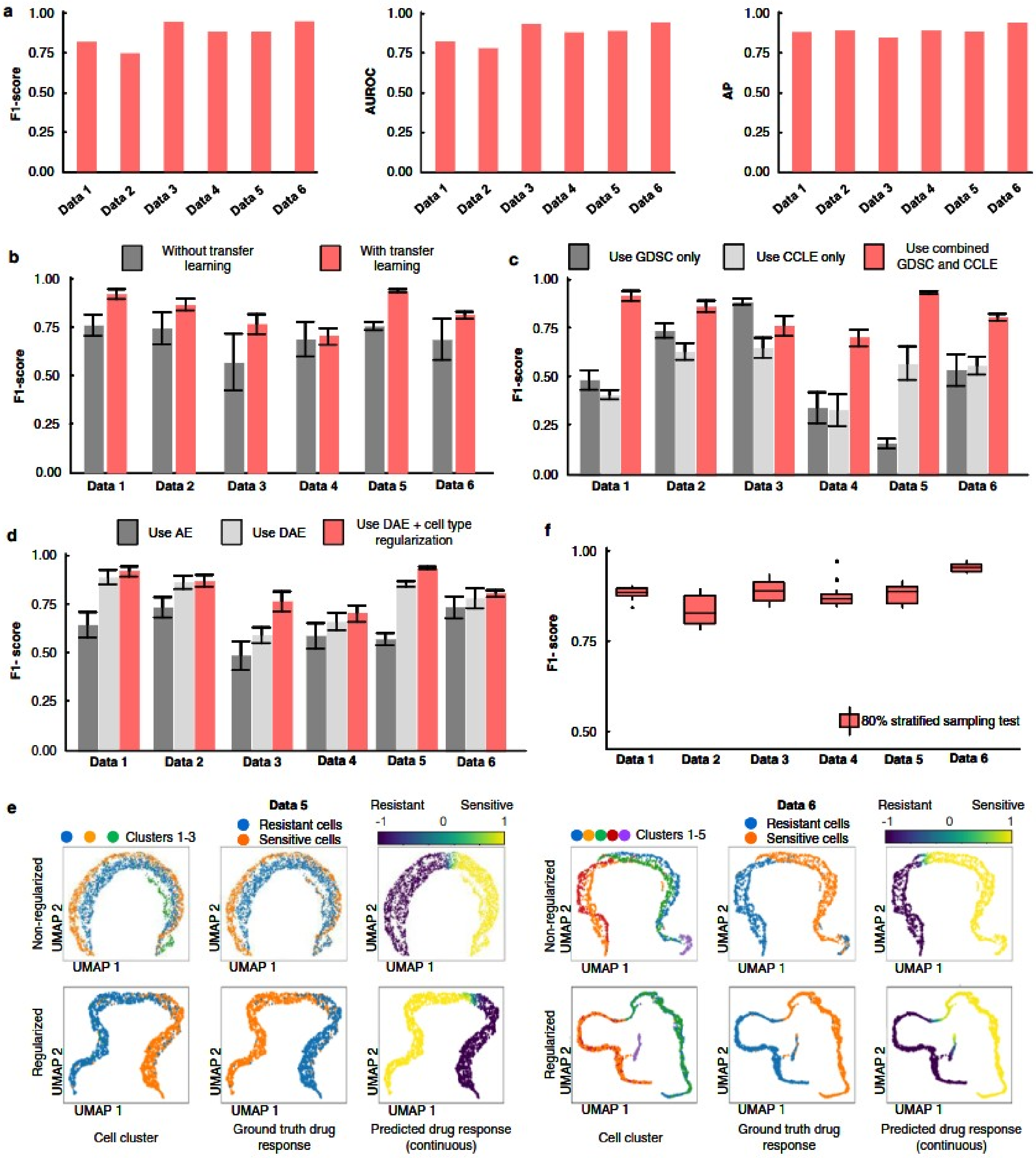
  Area Under the Receiver Operating Characteristic (AUROC): Assesses the model's ability to distinguish between drug-sensitive and drug-resistant cells.

  Average Precision (AP) Score: Evaluates the precision-recall trade-off.

  Precision and Recall: Measures the accuracy and completeness of the predictions.

  Adjusted Mutual Information (AMI) and Adjusted Rand Index (ARI): Evaluate the clustering performance, accounting for the agreement between predicted and true labels.

d. Major Results and Achievements



a. Performance Metrics:

- **F1-score, AUROC, and AP:**
  - The bar graphs show high performance metrics across all six datasets, with F1-scores averaging 0.892, AUROC scores averaging 0.898, and AP scores averaging 0.944. These metrics indicate robust and accurate drug response predictions by scDEAL, with high F1-scores suggesting a good balance between precision and recall.

b. Transfer Learning Comparison:

- **With vs. Without Transfer Learning:**
    - The comparison of F1-scores with and without transfer learning shows a significant improvement when transfer learning is applied. Specifically, there is an average increase of 19% in F1-scores across the datasets when using transfer learning, highlighting the effectiveness of transferring knowledge from bulk RNA-seq data to single-cell RNA-seq data.

c. Data Source Comparison:

- **GDSC only, CCLE only, Combined:**
    - Using combined GDSC and CCLE data yields better F1-scores than using either dataset alone. The average F1-scores show a 130% and 69% increase compared to using only the GDSC or CCLE datasets, respectively. This indicates that integrating multiple bulk RNA-seq data sources significantly enhances the model's predictive accuracy.

d. Model Architecture Comparison:

- **Autoencoder (AE) vs. Denoising Autoencoder (DAE) vs. DAE with Cell Type Regularization:**
    - The framework with DAE and cell type regularization outperforms other architectures, showing the highest F1-scores. Specifically, using DAE and cell type regularization shows an average increase of 36% and 9% in F1-scores compared to using AE or DAE alone. This suggests that incorporating denoising and regularization strategies effectively preserves single-cell heterogeneity and improves prediction accuracy.

e. UMAP Visualizations:

- **Cell Clusters, Ground Truth Drug Response, Predicted Drug Response:**
    - UMAP plots illustrate the clustering of cells and alignment between ground truth and predicted drug responses. The regularized models show more distinct and compact clusters, indicating better preservation of cellular heterogeneity. The visualizations confirm that the predicted drug response labels of most cells are well-aligned with the ground truth, showcasing distinct cell cluster differences.

f. Robustness Test:

- **80% Stratified Sampling Test:**
  - The box plots demonstrate the robustness of scDEAL across multiple runs of random sampling, with consistent F1-scores and minimal variation. The variations in F1-score, AUROC, and AP scores are 0.031, 0.046, and 0.027, respectively. This confirms the reliability of the model in predicting drug responses across different datasets and sampling conditions.

e. Conclusions

The authors conclude that scDEAL successfully leverages bulk RNA-seq data to enhance drug response predictions at the single-cell level. This approach addresses the limitations of existing methods by capturing the heterogeneity of cancer cells, providing accurate predictions that align well with experimental data. The framework's ability to identify critical genes involved in drug responses has significant implications for cancer treatment, including drug selection, repurposing, and understanding resistance mechanisms. Future work will focus on integrating more comprehensive bulk datasets and exploring cross-species prediction capabilities.