Deep transfer learning of cancer drug responses by integrating bulk and single-cell RNA-seq data

Report

Polezhaeva Valeria

Main problem:

Traditional cancer treatments continue to struggle with low effectiveness and high relapse rates due to the diverse states and cell fates within cancer heterogeneity.

The emergence of single-cell RNA-sequencing (scRNA-seq) opens up new possibilities for uncovering gene expression diversity within cancer subpopulations in response to specific drugs.

Current drug response prediction methods designed for bulk data are not directly applicable to the complexity of single-cell datasets.

Therefore, there is an urgent need for computational approaches to predict cancer drug responses at the single-cell level.

This heterogeneity results in varying responses of individual cells to drugs, often leaving behind residual cancer cells and ultimately leading to cancer relapse.

In vitro drug screening studies have generated valuable data on drug responses in various cancer cell lines.

Related works in the field and the need for a new solution

- It has previously been shown that deep learning models applied to scRNA-seq data have achieved competitive results in calculating gene expression, cell clustering, batch correction, and similar tasks.

However!

- The limited number of benchmarked data in the public domain is the reason for insufficiently trained deep learning models working with scRNA.

Solution:

- Bulk RNA-seq data can enhance single-cell drug response predictions through deep transfer learning (DTL), transferring knowledge from bulk to single-cell data.

This scale model, based on a Domain-adaptive Neural Network, predicts drug responses using both bulk and single-cell RNA-seq.

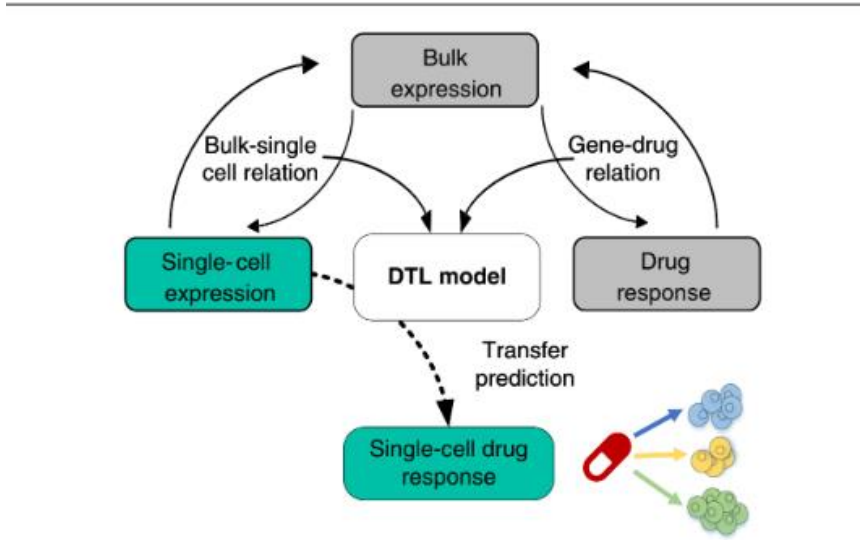
What do the authors propose?

- scDEAL model predicts drug responses using both bulk and single-cell RNA-seq data. It leverages data from the GDSC and CYCLE databases, harmonizing single-cell and bulk data structures for accurate predictions. By incorporating cell cluster labels and interpreting signature genes, scDEAL achieves high accuracy in predicting cell-type drug responses. It identifies gene signatures influencing drug sensitivity/resistance and aligns predicted responses with treatment trajectories.

Results

Overview of the scDEAL framework

ScDEAL establishes connections between gene expression features and drug response at the bulk level, where cell line annotations are provided. It then identifies a shared low-dimensional feature space that allows for the harmonization of these relationships between single-cell and bulk data. The gene expression-drug response associations at the bulk level are captured through this shared low-dimensional feature space. A DTL model is then trained to find the optimal solution for these connections. Ultimately, by leveraging the meta-relationship between gene expression at the single-cell level, gene expression at the bulk level, and drug response in the DTL model, scDEAL can predict drug responses for individual cells without needing supervised training at the single-cell level.



Results

Overview of the scDEAL framework

The scDEAL framework involves five major steps:

(1) extracting bulk gene features,

(2) predicting drug response in each bulk cell line using features extracted in step 1,

(3)extracting single-cell gene features,

(4) jointly training and updating all the models in the previous steps,

(5) transferring and applying the trained model to scRNA-seq data to predict drug responses.

Results
Overview of the scDEAL framework

Two denoising autoencoders (DAEs) are trained to extract low-dimensional gene features separately from bulk and single-cell RNA sequencing data. These features are then used to represent the original gene expressions efficiently. The initial training phase helps establish neuron weights within the DTL model, with a fully connected predictor linked to the trained bulk feature extractor for forecasting bulk-level drug responses.

The DTL model simultaneously updates two DAE models and the predictor model using multi-task learning. The first task minimizes differences between gene features from two extractors to connect bulk and scRNA-seq data. The second task minimizes differences between prediction results and database-provided drug responses through cross-entropy loss.

The output of scDEAL is the predicted potential drug response of individual cells.

Results

Overview of the scDEAL framework

One of the main challenges during model training is to preserve the unique characteristics of individual cells while aligning single-cell RNA sequencing data with bulk data. To address this, two strategies were implemented.

Firstly, recognizing the distinct noise patterns in bulk RNA-seq and scRNA-seq data, we opted for a Denoising Autoencoder (DAE) model instead of a standard autoencoder or variational autoencoder. This allowed us to introduce noise into both bulk and scRNA-seq data before reducing features, preventing the bias that could arise from pushing gene expressions in scRNA-seq data towards bulk RNA-seq data.

Secondly, cell clustering results was incorporated to regulate the overall loss function of scDEAL, ensuring that cellular diversity was preserved throughout the training process.

Results

Benchmarking single-cell drug response predictions in scDEAL

The accuracy of drug response predictions was tested on six public scRNA-seq datasets treated with five drugs:

Cisplatin, Gefitinib, I-BET-762, Docetaxel, and Erlotinib. Each dataset includes known drug response annotations for individual cells (0 for resistant, 1 for sensitive). ScDEAL predictions were compared with ground truth labels using various metrics: F1-score, AUROC, AP score, precision, recall, AMI, and ARI. Average scores across the datasets were: 0.892 (F1-score), 0.898 (AUROC), 0.944 (AP score), 0.926 (precision), 0.899 (recall), 0.528 (AMI), and 0.608 (ARI). Visualizations: UMAPs and Sankey plots highlighted the model's accuracy in predicting drug responses at the single-cell level.

Bulk-level results also demonstrated the model's effectiveness in analyzing scRNA-seq data.

Results

Benchmarking single-cell drug response predictions in scDEAL

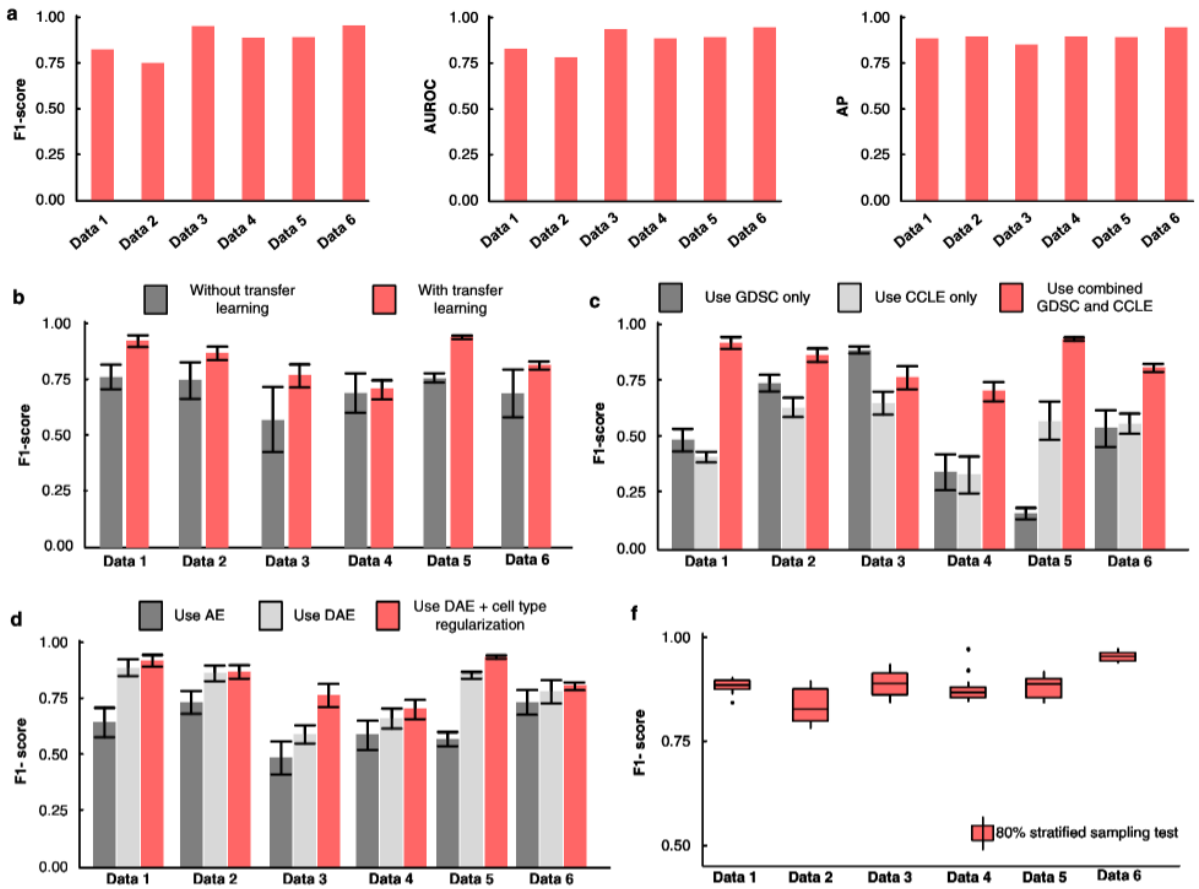
The scDEAL model showed high performance in single-cell drug response prediction across all six datasets. By testing different components of the scDEAL framework, it was found that transfer learning significantly improved F1-scores by 19% on average compared to models without transfer learning. Combining bulk data from GDSC and CCLE databases increased prediction power by 130% and 69% on average. Additionally, using DAE and cell-type regularization in the framework led to a 36% and 9% increase in F1-scores compared to other options. The cell-type regularizer helped preserve the heterogeneity of scRNA-seq data, as shown in UMAP visualizations.

The random stratified sampling test further proves the robustness of scDEAL, showing consistent performance across multiple runs. It's important to consider adjusting parameters like bulk sampling methods and bottleneck dimensions for optimal prediction performance, especially when working with new datasets. Random stratified sampling test further proves the robustness of scDEAL, showing consistent performance across multiple runs. It's important to consider adjusting parameters like bulk sampling methods and bottleneck dimensions for optimal prediction performance, especially when working with new datasets.

Results

Benchmarking single-cell drug response predictions in scDEAL

The scDEAL model showed high performance in single-cell drug response prediction across all six datasets. By testing different components of the scDEAL framework, it was found that transfer learning significantly improved F1-scores by 19% on average compared to models without transfer learning. Combining bulk data from GDSC and CCLE databases increased prediction power by 130% and 69% on average. Additionally, using DAE and cell-type regularization in the framework led to a 36% and 9% increase in F1-scores compared to other options. The cell-type regularizer helped preserve the heterogeneity of scRNA-seq data, as shown in UMAP visualizations.



Results

scDEAL advantages

The scDEAL analysis on Data 6 (GSE110894) showcased its efficacy in predicting drug responses in MA9 leukemic cells treated with an I-BET inhibitor. It achieved high accuracy rates of 97.1% for drug-resistant cells and 95.8% for drug-sensitive cells, providing continuous probability scores and binary labels for prediction. The gene scores accurately reflected differential gene expression levels, showing a strong correlation with ground truth data ($R2=0.90$ for sensitive DEGs, $R2=0.77$ for resistant DEGs), which was statistically significant ($p<0.001$) based on an empirical null model test.

In another instance on Data 118, scDEAL successfully predicted responses to Cisplatin in OSCC cells with 85% accuracy. It identified critical genes impacting drug response, distinguishing 936 drug-sensitive genes and 868 drug-resistant genes. Key resistant genes like BCL2A1 and DKK1, known for anti-apoptotic activity, were linked to Cisplatin resistance. Pathway analysis emphasized DNA repair and cell division processes in resistance, with genes related to these functions identified as crucial in predicting drug responses. This highlights scDEAL's ability to identify key genes relevant to drug response mechanisms.

Plans and limitations

- There are plans to enhance scDEAL by integrating more bulk gene expression databases and validated drug response single-cell RNA sequencing data. This will help improve the model's performance and potentially lead to the development of direct single-cell-to-single-cell deep transfer learning models.
- Some limitations of scDEAL include its reliance on the quality and availability of bulk gene expression data and the scarcity of experimentally validated drug response RNA-seq data.

Conclusions

- scDEAL is a tool that enhances the analysis of single-cell RNA sequencing data by incorporating bulk gene expression data. It can predict drug responses in cell populations based on cancer single-cell RNA sequencing data and other diseases. By training neural networks on bulk cell-line data, scDEAL can predict drug sensitivity from single-cell RNA sequencing data without the need for labels like cell type or drug response.
- The results showed that scDEAL excels in predicting drug response labels and identifying gene signatures.
- Reliability was tested by reassembling datasets with potential sample swaps, and the results indicated that scDEAL maintains competitive drug response prediction accuracy. This demonstrates the robustness and effectiveness of our approach in predicting drug responses.