HW 6

a. What is the problem the article's trying to address?

Cancer drug treatments have low effectiveness and high relapse rates due to the heterogeneity of cancer cells. Single-cell RNA sequencing (scRNA-seq) can reveal gene expression patterns in different cell populations, but existing methods for predicting drug responses cannot be directly applied to single-cell data. Deep learning models have been successful in analyzing large-scale scRNA-seq data, but the limited amount of benchmark data hinders their training. Transfer learning can address this issue by leveraging knowledge from bulk RNA-seq data. In this study, the authors develop scDEAL, a neural network model that predicts drug responses using both bulk and single-cell data. scDEAL incorporates a large amount of bulk-level data from databases to optimize the model and ensures transferability of drug response labels between bulk and single cells. It also considers cell clusters and integrates gradient interpretation to enhance model interpretability. The authors demonstrate the accuracy of scDEAL in predicting drug responses in diverse cell types and identify gene signatures related to drug sensitivity or resistance. This tool can aid in preliminary studies on drug development and selection in cancer treatment.

b. What are the related works in the field and why there is still a need to propose new solutions?

AI can help decipher the varied gene expressions of cancer subpopulations in response to drugs using scRNA-sequencing. Existing drug-response prediction methods for bulk data aren't applicable for highly complex single-cell data, necessitating the development of computational approaches to address this challenge.

c. What do the authors propose? Describe their solution, input data, processing, metrics etc.

The authors propose a deep transfer learning framework. scDEAL is a framework that models the relationship between gene expression and drug response at the bulk level. It identifies the shared low-dimensional feature space between single-cell and bulk data to harmonize the two data types. A DTL model is trained to optimize these relations. The framework involves five steps: extracting bulk gene features, predicting drug response in each bulk cell line, extracting single-cell gene features, training and updating models, and applying the trained model to scRNA-seq data. Two denoising autoencoders are used to extract low-dimensional gene features and reduce reconstruction loss. The DTL model updates the models simultaneously in a multi-task learning manner. The output is the predicted potential drug response of individual cells. Strategies are used to maintain single-cell heterogeneity in the training process (Fig. 1 from article).

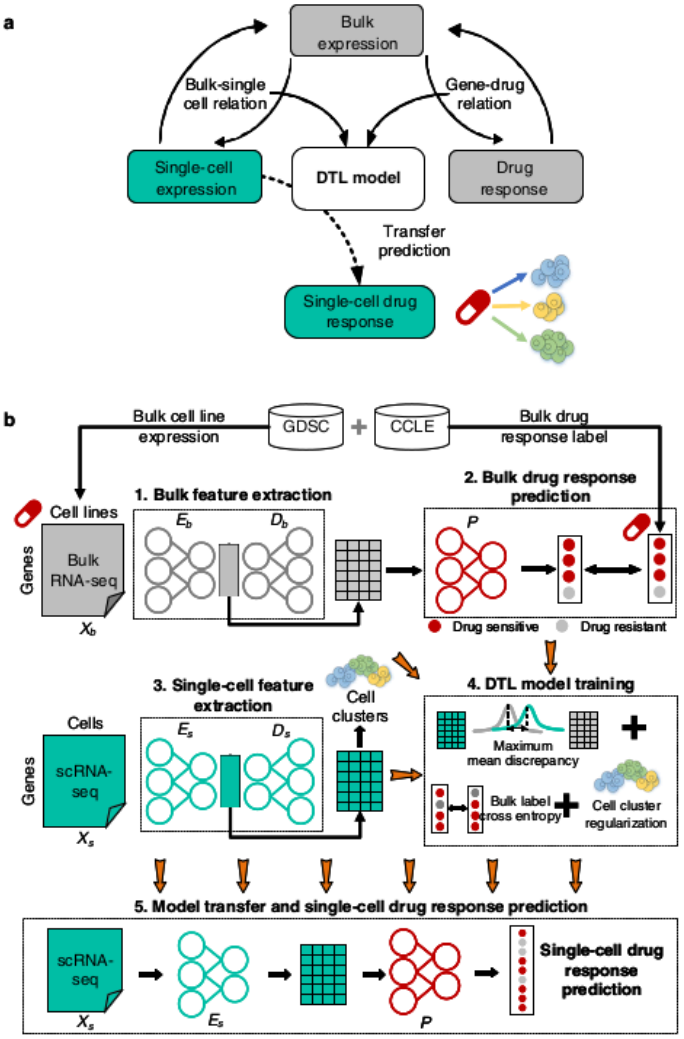

Fig. 1 | The scDEAL framework. a scDEAL trains the model to align two relations: (*i*) bulk–single-cell relations and (*ii*) gene–drug response relations at the bulk level. The trained model will then be transferred to be directly applied to the scRNA-seq data and to predict the single-cell drug responses. Green-colored elements represent single-cell related data, and grey-colored elements represent bulk-related data. Different colors of cells represent different cell types. **b** Bulk RNA-seq data and the corresponding drug response labels are obtained from the GDSC and CCLE databases. Five steps are then applied. A DAE is used to induce noises into the bulk data. It uses an encoder (E_b) and a decoder (D_b) to obtain low-dimensional features. The bulk feature × cell-line matrix is then input to a fully connected predictor (P) to predict cell-line drug responses. A similar strategy is used for single-cell feature election using a separated DAE (E_s and D_s). The overall framework will be trained by considering the maximum mean discrepancy between the low-dimensional feature spaces of single-cell and bulk data, the cross-entropy loss between predicted bulk cell-line drug responses and ground-truth labels, and the regularization of cell clusters predicted from scRNA-seq data. By achieving the minimum overall loss, E_b, E_s, and P will be updated and optimized simultaneously. scDEAL transfers the well-trained E_s and P to predict single-cell drug responses from the scRNA-seq data. Abbreviations: deep transfer learning (DTL), Genomics of Drug Sensitivity in Cancer (GDSC), Cancer Cell Line Encyclopedia CCLE.

The drug response prediction performances of scDEAL were evaluated on six public scRNA-seq datasets treated with five drugs: Cisplatin, Gefitinib, I-BET-762, Docetaxel, and Erlotinib. The datasets had ground-truth drug response annotations for individual cells, indicating whether they were drug-sensitive or drug-resistant. The scDEAL predictions were evaluated using seven metrics, including F1-score, AUROC, AP score, precision, recall, AMI, and ARI. The results showed that scDEAL had high performance in single-cell drug response prediction, with average scores of 0.892 (F1-score), 0.898 (AUROC), 0.944 (AP score), 0.926 (precision), 0.899 (recall), 0.528 (AMI), and 0.608 (ARI) across all datasets. UMAP visualizations and Sankey plots were used to showcase the prediction results, which aligned well with the ground truth and revealed distinct cell cluster differences. Comparison tests were also conducted to validate the effectiveness of the scDEAL framework, including transfer learning, bulk data integration, and the use of autoencoders and cell-type regularization. The results demonstrated that these components contributed to the improved performance of scDEAL. Finally, the robustness of scDEAL was evaluated through randomly stratified sampling tests, which showed consistent performance across multiple runs.

d. What are the major results and achievements of the proposed solution? How do they relate to the existing methods and what are the limitations?

The scDEAL was successful in predicting drug response in leukemia cells treated with I-BET. The predictions matched the original labels for both drug-resistant and drug-sensitive cells. scDEAL provided continuous probability scores and binary labels to determine drug response. Additionally, a gene score was introduced to reflect gene expression levels in different cell clusters. The correlation between predicted and ground truth gene scores was high, indicating accurate prediction. An empirical null model test confirmed the significance of the correlation.

The model can identify critical genes responsible for drug response, as demonstrated by scDEAL analysis of oral squamous cell carcinoma treated with Cisplatin. Cisplatin induces DNA crosslinks, interfering with replication and causing double-strand breaks, leading to apoptosis. Genes that enhance DNA repair or inhibit apoptosis can render cancer cells resistant to Cisplatin. Using scDEAL, 85% of cells were accurately predicted for sensitivity or resistance to Cisplatin. The analysis identified 936 drug-sensitive genes and 868 drug-resistant genes, including BCL2A124 and DKK125, which mediate Cisplatin resistance. Pathway enrichment analysis revealed the significance of DNA repair and cell division pathways in Cisplatin resistance. Additional pathways associated with resistance were also identified.

Drug response prediction in scDEAL was validated using pseudotime analysis. Monocle346 was used for trajectory inference on Data 6 (I-BET treated samples). Pseudotime analysis shows a trend starting from DMSO samples towards I-BET treated samples. The drug response (probability score) on the same diffusion UMAP shows increased resistance from DMSO control to treated samples. Expression levels of CGs (Eid2, Galnt17) and DEGs (Emilin1, Ramp1) match the trajectory of pseudotime analysis and predicted drug response probability scores. scDEAL predictions have strong correlations to drug response development.

Limitations:

- The level of complexity of the model might necessitate a considerable amount of computational resources.
- AI technology requires the use of extensive collections of bulk and single-cell RNA-seq data in order to train the model efficiently.
- The ongoing difficulty in forecasting drug response at a cellular level is accurately predicting responses in various species.

e. What are the conclusions?

scDEAL is a tool that enhances the analysis of single-cell RNA sequencing (scRNA-seq) data by incorporating bulk gene expression data. It can predict drug responses in cell populations based on scRNA-seq data in cancer and other diseases. The neural networks used in scDEAL are trained on large volumes of bulk cell-line data, allowing for the prediction of drug sensitivity from scRNA-seq data alone, without the need for cell type or drug response labels. The performance of scDEAL was evaluated on six drug-treated scRNA-seq datasets and demonstrated robust predictions of drug response labels and identification of gene signatures. Additional features of scDEAL include the ability to explain and interpret genetic features between bulk and scRNA-seq data using IG scores and CG identification. The tool is adaptable to different datasets and can be used for integrative drug

combination predictions. There is potential to further improve scDEAL by incorporating more bulk gene expression data and experimentally validated drug response scRNA-seq data.