Report about Deep learning article [1] by Smutin Daniil

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

The article aims to address the challenge of predicting cancer drug responses at the single-cell level, considering the heterogeneity of drug responses among cancer cell subpopulations. Current drug response prediction methods developed for bulk data are not directly applicable to single-cell data due to its complexity and scale, necessitating the development of computational approaches for single-cell drug response inference. So, the article tries to find a better way of single-cell and bulk RNAseq integration, to predict drug responses accurately at the single-cell level, aiming to improve therapeutic efficacy and understand drug resistance mechanisms.

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

Existing drug response prediction methods have been developed for bulk data but are not directly applicable to the complexity and scale of single-cell data, highlighting the need for computational approaches tailored to single-cell drug response inference. On the other hand, deep learning methods can deal well with scRNAseq data [2,3], so they can be possibly used for integrating it with bulk data. When the baseline method works only with the data part, and you possibly have more data, any type of non-random combining should be prediction efficient because you just will have more predictor variables.
Previously, DTL models were efficiently used for integrating multiple bulk RNAseq data sources dealing with batch effect and sequencing bias [4]. So, authors try to integrate both data sources and use them for making better drugs predictions. When there are so many possible types of cancer and so many more possible drugs, standard algorithms or human analysis of the data is often of little use. Standard classification algorithms based on machine learning suffer from insufficient data in the training sample and the number of features, so deep learning looks promising.

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

Solution
The authors propose scDEAL, a deep transfer learning framework for predicting cancer drug responses at the single-cell level by integrating large-scale bulk cell-line data. scDEAL harmonizes drug-related bulk RNA-seq data with single-cell RNA-seq data to transfer models trained on bulk data for predicting drug responses in single cellsю. The framework includes integrated gradient feature interpretation to infer signature genes associated with drug resistance mechanisms. scDEAL is benchmarked on six single-cell RNA-seq datasets, demonstrating its model interpretability through case studies focusing on drug response label prediction, gene signature identification, and pseudotime analysis.
In major, scDEAL involves training a model on drug-related bulk RNA-seq data and transferring this model to predict drug responses in single-cell RNA-seq data (figure 1).

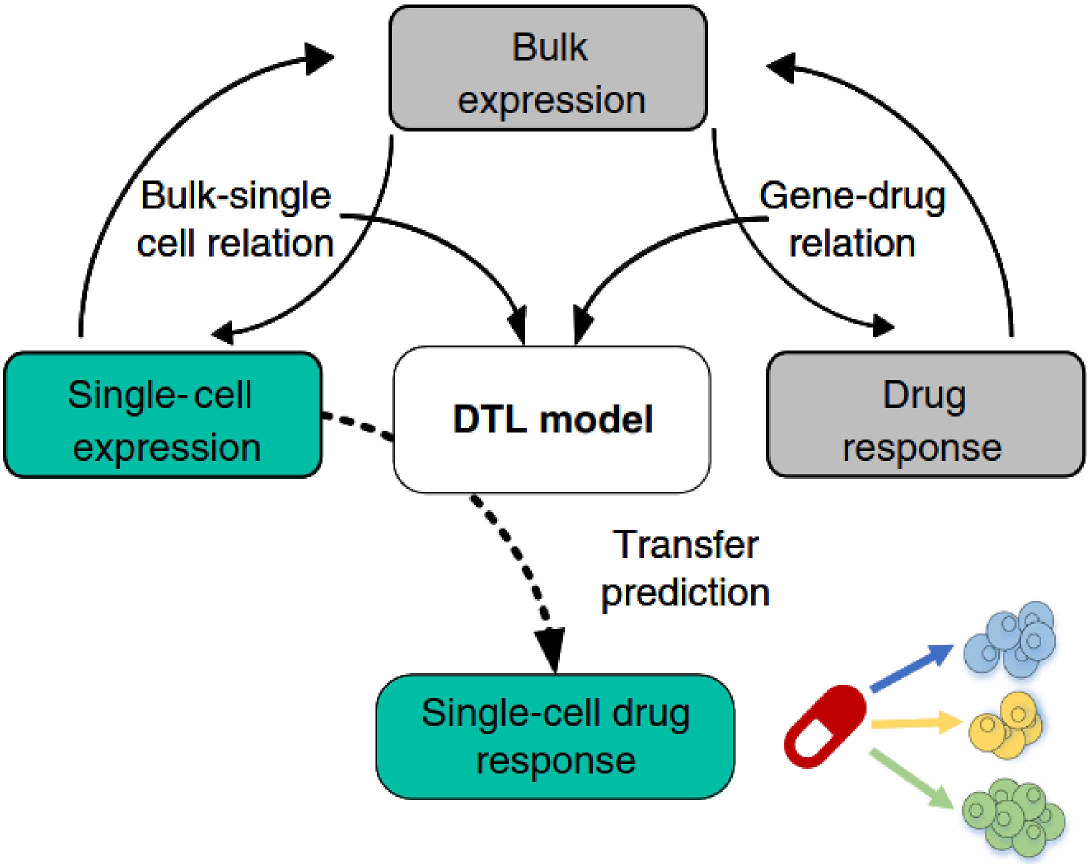

Figure 1. DTL model relations between bulk-single cell data and gene-drug data. Pre-trained on bulk data scDEAL should be possible to transfer directly to scRNAseq data to predict possible drug response.

Model description

The scDEAL framework consists of two main components: supervised learning for building a model to predict response labels at the bulk level and transferring this model for label prediction at the single-cell level.

The framework involves several key steps:

- Bulk feature extraction using a Denoising Autoencoder (DAE) to learn a low-dimensional representation from bulk expression data.
- Training a predictor model based on a fully connected multi-layer perceptron (MLP) for bulk drug response prediction.
- An integrated gradient (IG) feature interpretation method is applied in scDEAL to infer critical input gene features associated with drug resistance mechanisms, enhancing the interpretability of the model.
- Single-cell feature extraction using a similar DAE model to extract low-dimensional features from single-cell RNA-seq data.
- Deep Transfer Learning (DTL) model training to adapt gene features from bulk and single-cell levels for sensitivity prediction in cells through the predictor model.
- The DTL training incorporates a DaNN model that introduces an extra loss function called Maximum Mean Discrepancy (MMD) to estimate the similarity between the extracted features from bulk and single-cell data.

Data and processing

The research utilized bulk gene expression data from the GDSC database, which includes drug response annotations such as IC50 and AUC for 1280 cancer cell lines, 1557 drugs, and their expression profiles on 15,962 genes. Additionally, the CCLE cell line expression profile and PRISM cell line viability assay data were collected and integrated with the GDSC data to create a comprehensive dataset for analysis.

Quality control and preprocessing of the single-cell RNA-seq data were conducted using *scanpy*, filtering out cells with less than 200 detected genes and genes detected in less than 3 cells. The trajectory inference for the single-cell RNA-seq data was preprocessed using Monocle3, projecting the read count matrix to a 2-dimensional UMAP space and constructing a graph topology from the reduced dimension space based on the reversed graph embedding algorithm.

Deep learning and model performance

Bulk RNA-seq data was splitted into training, validation, and testing sets, with 64, 16, and 20 subsets, respectively, for initial model development and evaluation. A comparison test was conducted by training the model solely on bulk data and directly using it for single-cell data prediction without transfer learning, repeating the experiment 50 times to assess model performance.

The scDEAL framework utilized seven key metrics to evaluate the prediction performance of drug responses in single-cell RNA-seq data, including F1-score, area under the receiver operating characteristic (AUROC), AP score, precision, recall, Adjusted Mutual Information (AMI), and Adjusted Rand Index (ARI) . The average scores across the six datasets were reported as follows:

- F1-score (0.892)
- AUROC (0.898)
- AP score (0.944)
- precision (0.926)
- recall (0.899)
- AMI (0.528)
- ARI (0.608)

These metrics collectively provided a comprehensive evaluation of the scDEAL model's ability to predict drug responses in single-cell RNA-seq data, demonstrating its effectiveness in identifying drug-sensitive and drug-resistant cells with high accuracy and reliability. To be honest, these metrics are truly incredible for such complex data.

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?

Overall, the proposed solution as was mentioned above demonstrates really high sensitivity and precision (figure 2.) The results demonstrated a significant increase in F1-scores when employing the transfer learning strategy compared to without it, with an average 19% increase in F1-score achieved by the model. That is really a big DEAL).

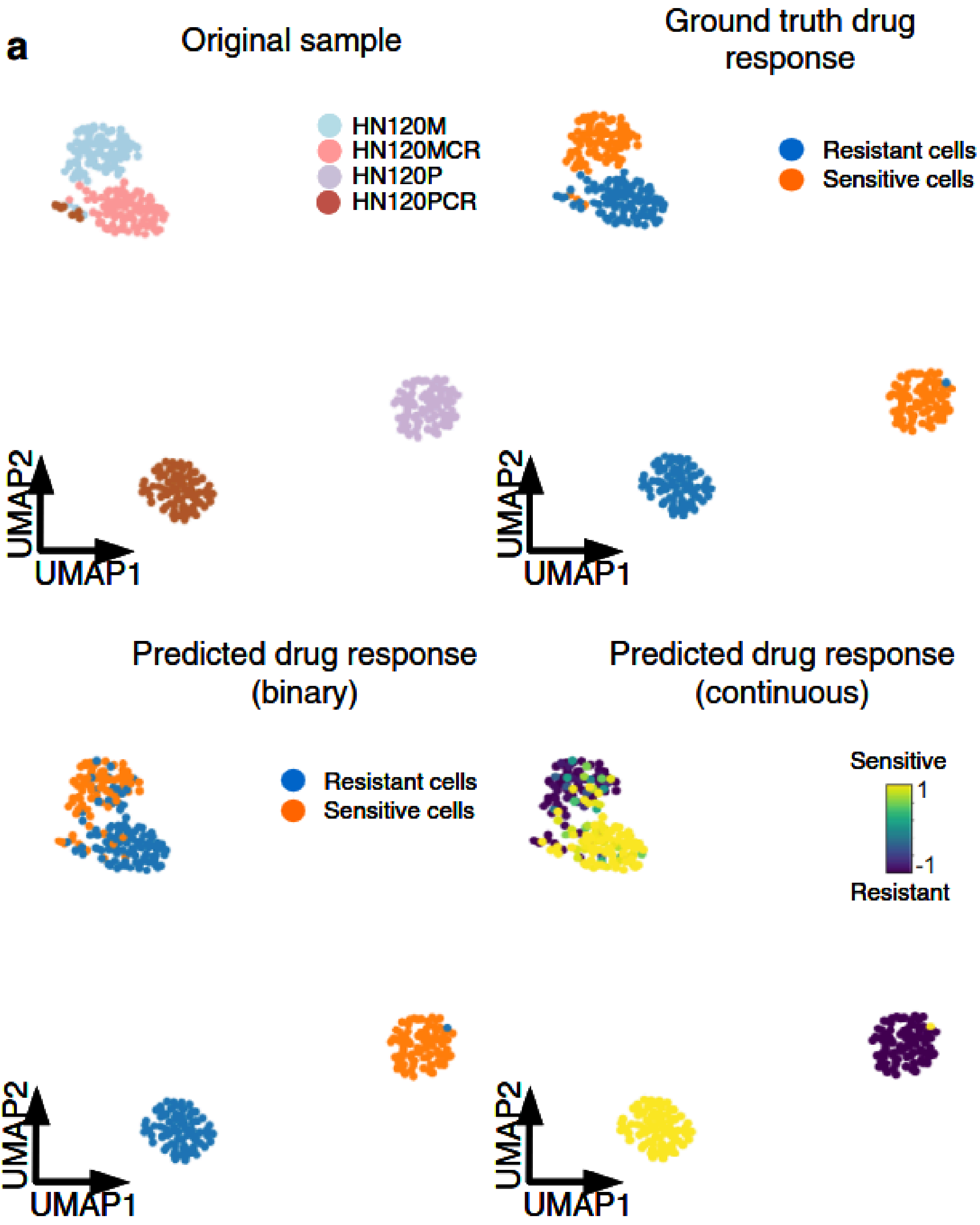

Figure 2. UMAP comparison between ground-truth labels and predicted binary response labels in scDEAL. Overall, very good predictions, but could be seen a little bit of false-negatives. With almost 0 true-negative rate, which is insane. Even dealing with possible outlier (right down cluster).

Integration of bulk and single-cell RNA-seq data in the scDEAL framework enabled the prediction of cancer drug responses at the single-cell level, showcasing the potential of leveraging large-scale bulk data to enhance predictions in single-cell data.

The framework's integrated gradient feature interpretation method successfully identified signature genes associated with drug resistance mechanisms, providing valuable insights into the molecular basis of drug responses in cancer cells. Pseudotime analysis conducted in the study revealed a correlation between scDEAL's predicted drug responses and the progression of drug treatment, indicating that the model's predictions align with the development of drug resistance over time (figure 3). The correlation analysis between predicted drug response scores and ground truth sensitive and resistant gene expression scores showed strong correlations, further validating the accuracy of scDEAL in predicting drug resistance at the single-cell level

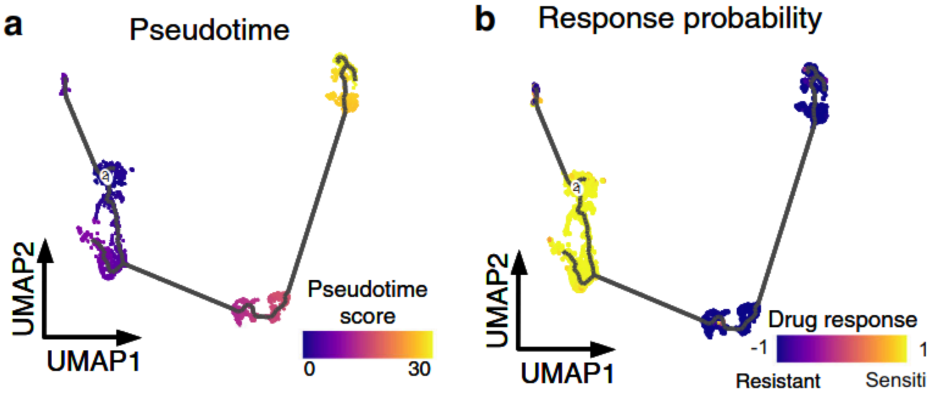

Figure 3. Validating predicted drug response with pseudotime trajectory. a Cell UMAP plot colored as per pseudotime scores predicted from the raw scRNA-seq data. b Same UMAP plot colored as per predicted continuous drug response probability scores in scDEAL

e. What are the conclusions?

The scDEAL framework, integrating bulk and single-cell RNA-seq data, demonstrates a powerful approach for predicting cancer drug responses at the single-cell level, offering insights into drug resistance mechanisms and heterogeneity within cancer cell populations. The model's interpretability, through integrated gradient feature interpretation, enables the identification of signature genes associated with drug resistance, enhancing our understanding of the molecular basis of drug responses in cancer cells. scDEAL's competitive performance in predicting drug responses across multiple scRNA-seq datasets underscores its potential for precision medicine applications, aiding in the selection and repurposing of drugs to improve therapeutic efficacy.

The ideas of data transfer and overlap in omics research are not new. For example, there is a large project, mixOmics, designed for this purpose, but any such data suffer greatly from the batch effect. The authors' model shows amazing efficiency, and similar approaches can be used in the future to integrate other multiomics data, e.g. genomic single-cell CV, SNP and proteomic data.

References

1. Chen, J. *et al.* Deep transfer learning of cancer drug responses by integrating bulk and single-cell RNA-seq data. *Nat. Commun.* **13**, 6494 (2022).
2. Ma, Q. & Xu, D. Deep learning shapes single-cell data analysis. *Nat. Rev. Mol. Cell Biol.* **23**, 303–304 (2022).
3. Wu, Z. *et al.* Single-Cell Techniques and Deep Learning in Predicting Drug Response.

Trends Pharmacol. Sci. **41**, 1050–1065 (2020).

4. Dhruba, S. R., Rahman, R., Matlock, K., Ghosh, S. & Pal, R. Application of transfer learning for cancer drug sensitivity prediction. *BMC Bioinformatics* **19**, 497 (2018).