Enhancing Breast Cancer Treatment Response Prediction with Single-Cell RNA Sequencing and Large Language Models

Yiming (Emmett) Peng[1], Victoria Truong[1], Aoqi Xie[1], Yu Shi[1]

[1]Dalla Lana School of Public Health, University of Toronto

Project Highlights

- **Developed PRECISE**, a novel framework that uses large language models and cell-type-specific markers to predict treatment response.
- **Outperforms existing models**, including Seurat-based pipelines and published benchmarks (e.g., PD-L1, InteractPrint).
- **Demonstrates consistent accuracy** across independent datasets.

Background and Objectives

Background

Accurately predicting breast cancer treatment response can increase pathologic complete response (pCR) rates, an indicator of potential cure, and reduce unnecessary toxicity.

Challenges

- The substantial heterogeneity of breast cancer can only be fully captured through single-cell RNA sequencing (scRNA-seq) data.
- Existing prediction models often rely on bulk-level features and overlook cell-type-specific signals that may drive treatment response.

Objectives

Develop a predictive framework that uses cell-type-specific marker genes from scRNA-seq data to improve treatment response (pCR) prediction in breast cancer.

Discovery and Validation Datasets

Discovery Dataset – Bassez et al. Cohort (2021)

- Sample size: 29 patients (9 achieve pCR, 20 did not).
- scRNA-seq dimensions: 157,760 cells × 25,291 genes.
- Bulk gene expression data: aggregated from the scRNA-seq data using PRECISE model.

Validation Dataset – I-SPY2 Trial Cohort Treated with anti-PD1 Treatment

- Sample size: 69 patients (31 achieve pCR, 38 did not).
- Bulk gene expression data dimensions: 69 patients × 19,134 genes.

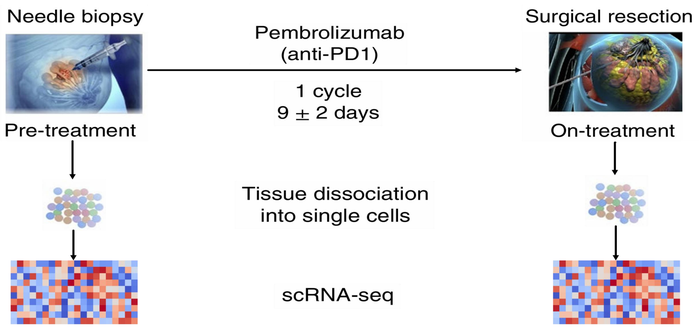

Figure 1. Sampling process of the discovery dataset (modified from Bassez et al., 2021).

Methods

PRECISE Framework (Prediction of REsponse using Cell-type Inference and Single-cell Embedding)

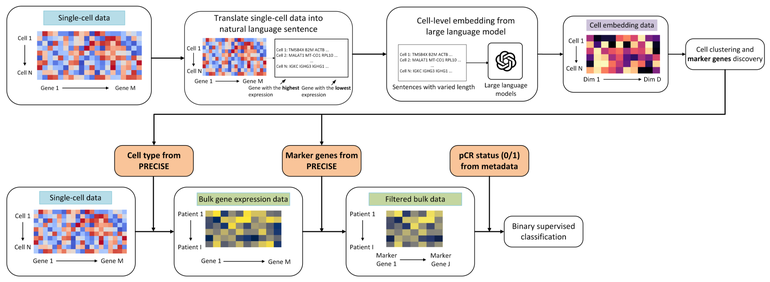

Figure 2. Overview of the PRECISE framework, which leverages large language models and cell-type-specific markers for treatment outcome prediction.

- Use large language models (LLMs) to generate cell-level embeddings.
- Cluster cells and identify marker genes per cell type using *Louvain algorithm* and *Seurat.*
- Cell type–specific marker genes are used to predict pCR.

Workflow 1: Generating Cell Embeddings with LLMs

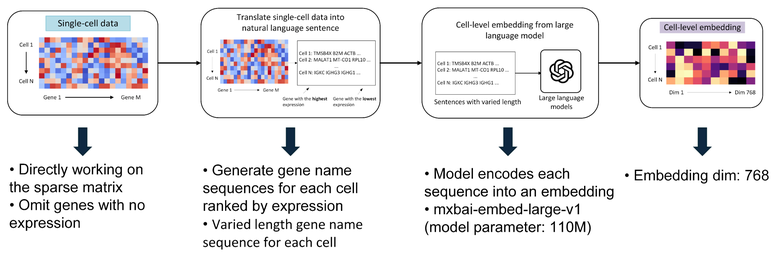

Figure 3. Workflow for generating cell-level embeddings from single-cell RNA-seq data using LLMs.

- Sparse expression matrix transformed into ranked gene lists per cell.
- Gene lists converted into sentences for LLM input.
- Resulting embeddings capture complex cell-level patterns.

Workflow 2: Predicting Treatment Outcomes from Bulk Gene Expression

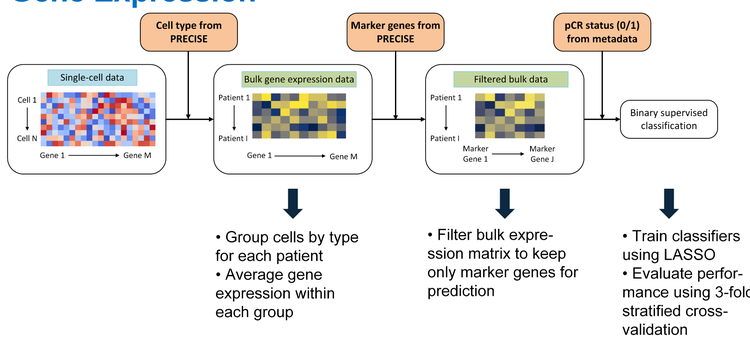

Figure 4. Workflow for predicting treatment outcomes using bulk gene expression derived from single-cell RNA-seq and PRECISE-identified marker genes.

- Gene expression averaged within groups to generate pseudo-bulk profiles.
- PRECISE-identified marker genes used for classification.

Results

- PRECISE's macrophage marker genes achieved high AUCs across all four settings, outperforming Seurat-derived markers (**Figure 5**).
- PRECISE outperformed PD-L1 and T Cell InteractPrint baselines with AUCs of 0.861 (Bassez) and 0.917 (I-SPY2) (**Figure 6**).
- PRECISE offers reliable uncertainty estimates via conformal prediction, with few low-confidence cases near the 0.5 threshold (**Figure 7**).

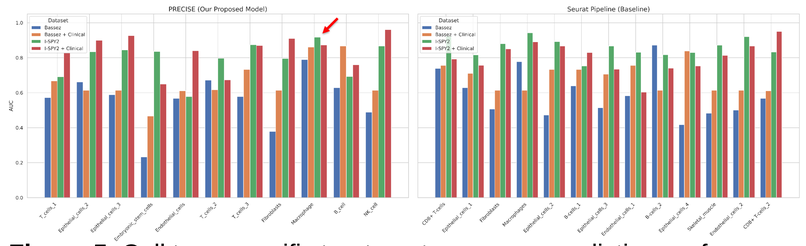

Figure 5. Cell type-specific treatment response prediction performance across models and datasets (LASSO).

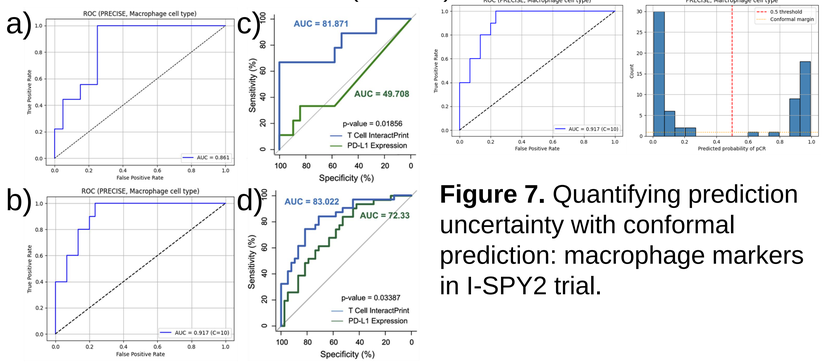

Figure 6. Comparison of PRECISE macrophage-based prediction with existing published models.
(a, b) ROC curves from PRECISE using macrophage markers on the Bassez dataset (a) and I-SPY2 trial (b).
(c, d) ROC curves from Xu et al., Cell Reports Medicine (2024), comparing T Cell InteractPrint and PD-L1 Expression on the same datasets.

Figure 7. Quantifying prediction uncertainty with conformal prediction: macrophage markers in I-SPY2 trial.

Conclusions

- PRECISE improves treatment outcome prediction, outperforming published models and clinically used approaches across datasets.
- PRECISE consistently outperforms the Seurat pipeline, with embedding-based features from foundation models generalizing well across datasets.

Limitation and Future Work

- Reproduce PD-L1 Expression and InteractPrint pipelines for statistical comparison.
- Analyze top marker genes for biological and clinical relevance.

References and Acknowledgements

Special thanks to the Hu Lab for their invaluable support and guidance.

References