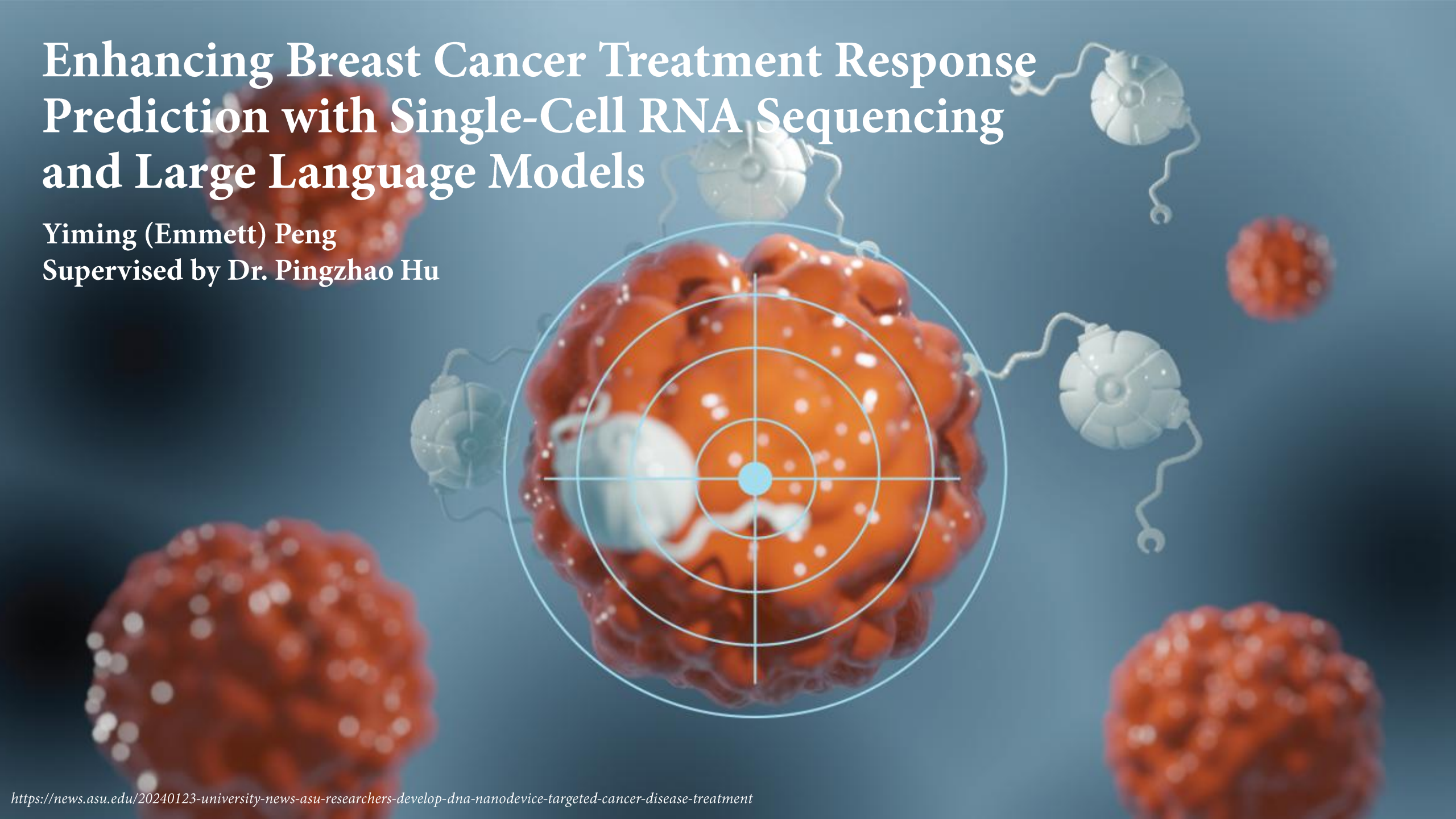
Enhancing Breast Cancer Treatment Response Prediction with Single-Cell RNA Sequencing and Large Language Models

Yiming (Emmett) Peng
Supervised by Dr. Pingzhao Hu



https://news.asu.edu/20240123-university-news-asu-researchers-develop-dna-nanodevice-targeted-cancer-disease-treatment

Background

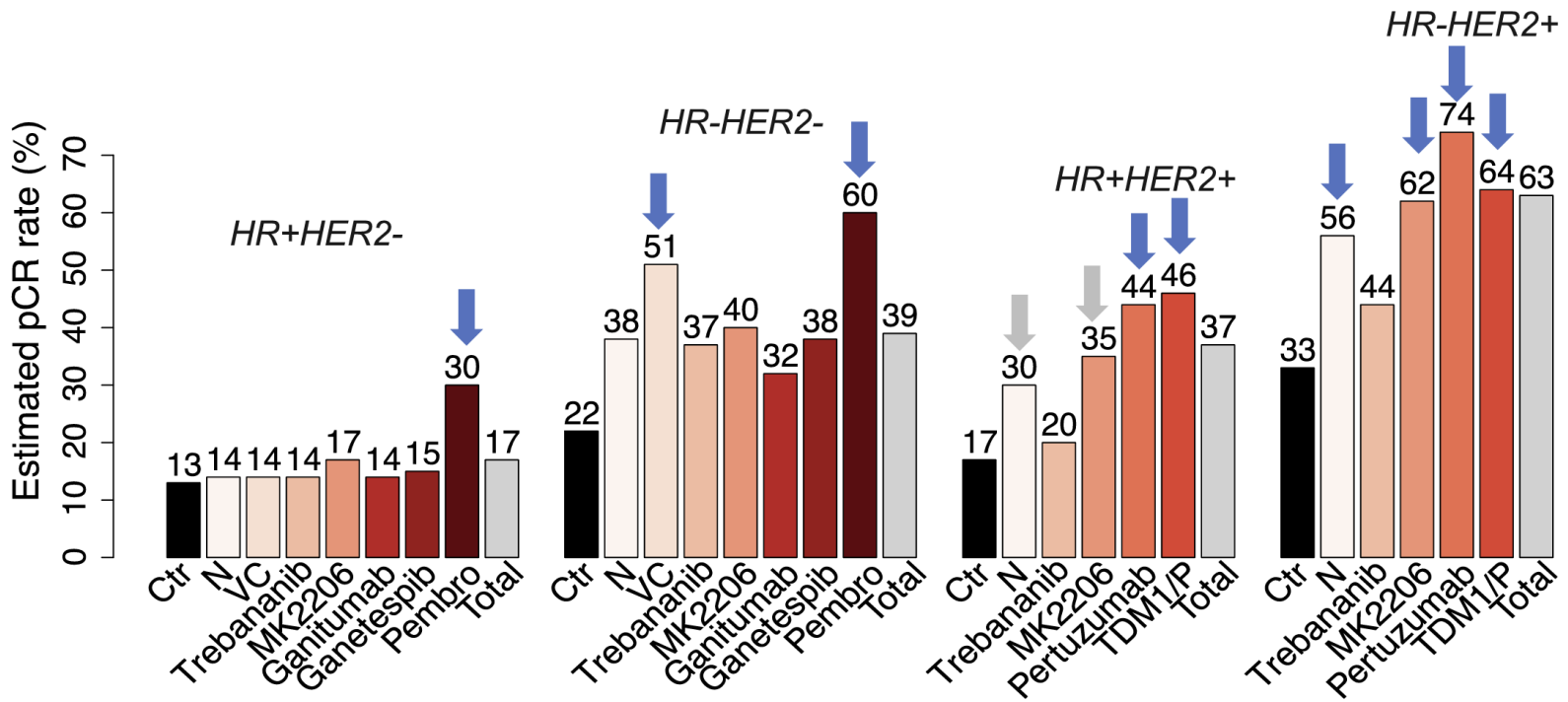
Challenges in Breast Cancer Treatment Response Prediction

- Breast cancer is the most common cancer among women worldwide, **one in three patients dies from the disease globally** (DeSantis et al., 2015)
- **Accurate treatment prediction** is crucial to ensure each patient receives the most effective therapy, maximizing the chance of achieving pCR and reducing unnecessary side effects
  - Pathologic Complete Response (pCR): absence of invasive cancer in the breast after treatment (indicator of cure)
- Breast cancer is known to have substantial **heterogeneity**, which can only be fully understood when analyzing **single-cell RNA sequence data**

Challenges in Breast Cancer Treatment Response Prediction

- Current Challenge in Breast Cancer Treatment: existing treatment strategies fail to achieve desirable pCR rates across breast cancer subtypes

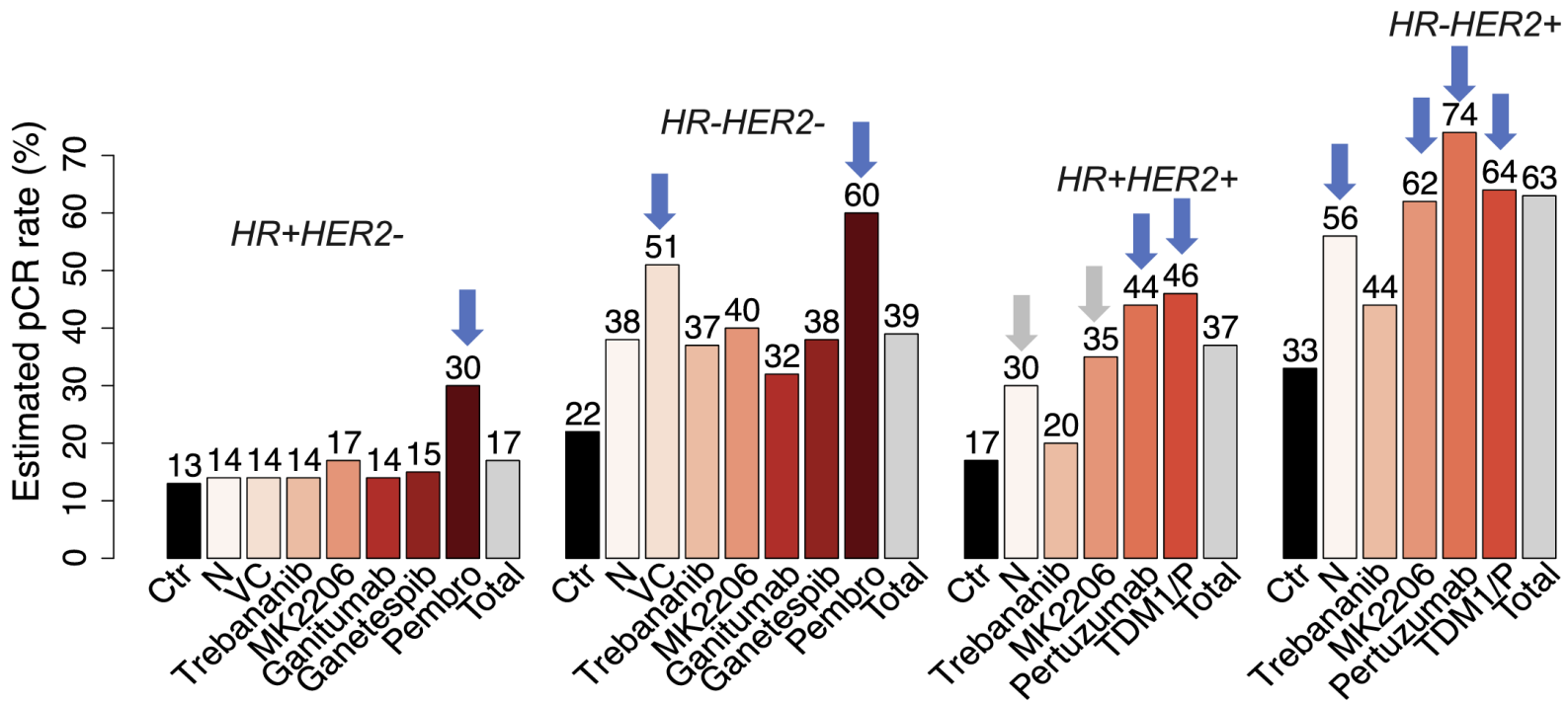
Estimated pCR rate across arms by receptor subtype (adapted from Wolf et al., 2022)



Challenges in Breast Cancer Treatment Response Prediction

- Current Challenge in Breast Cancer Treatment: existing treatment strategies fail to achieve desirable pCR rates across breast cancer subtypes

Estimated pCR rate across arms by receptor subtype (adapted from Wolf et al., 2022)



A new treatment outcome prediction structure is needed

Current Treatment Response Prediction Methods

- In 2022, Wolf et al. use bulk RNA-seq and protein data to refine breast cancer subtypes, incorporating biomarkers through logistic regression (**RPS-5**) to predict the treatment response
  - Bulk RNA-seq: averages gene expression across cells
  - Biomarkers: e.g., specific genes/proteins
  - Limitation: Relies on bulk data, which overlooks cellular heterogeneity present in breast cancer
- Xu et al. (2024) integrate scRNA-seq datasets to study heterogeneity and develop **InteractPrint** that computes weighted scores for cell-cell interactions for treatment response prediction
  - Limitations: **InteractPrint** is limited to known interactions

Current Treatment Response Prediction Methods

- In 2022, Wolf et al. use bulk RNA-seq and protein data to refine breast cancer subtypes, incorporating biomarkers through logistic regression (**RPS-5**) to predict the treatment response
  - Bulk RNA-seq: averages gene expression across cells
  - Biomarkers: e.g., specific genes/proteins
  - Limitation: Relies on bulk data, which overlooks cellular heterogeneity present in breast cancer
- Xu et al. (2024) integrate scRNA-seq datasets to study heterogeneity and develop **InteractPrint** that computes weighted scores for cell-cell interactions for treatment response prediction
  - Limitations: **InteractPrint** is limited to known interactions

A new cell-level prediction framework is needed for accurately predict pCR

Research Objective

Research Objective

To utilize single-cell RNA-seq data to identify cell-type-specific marker genes for predicting treatment response (pCR) in breast cancer patients, addressing cellular heterogeneity and improving upon current predictive methods

Methods

PRECISE Framework for Treatment Response Prediction

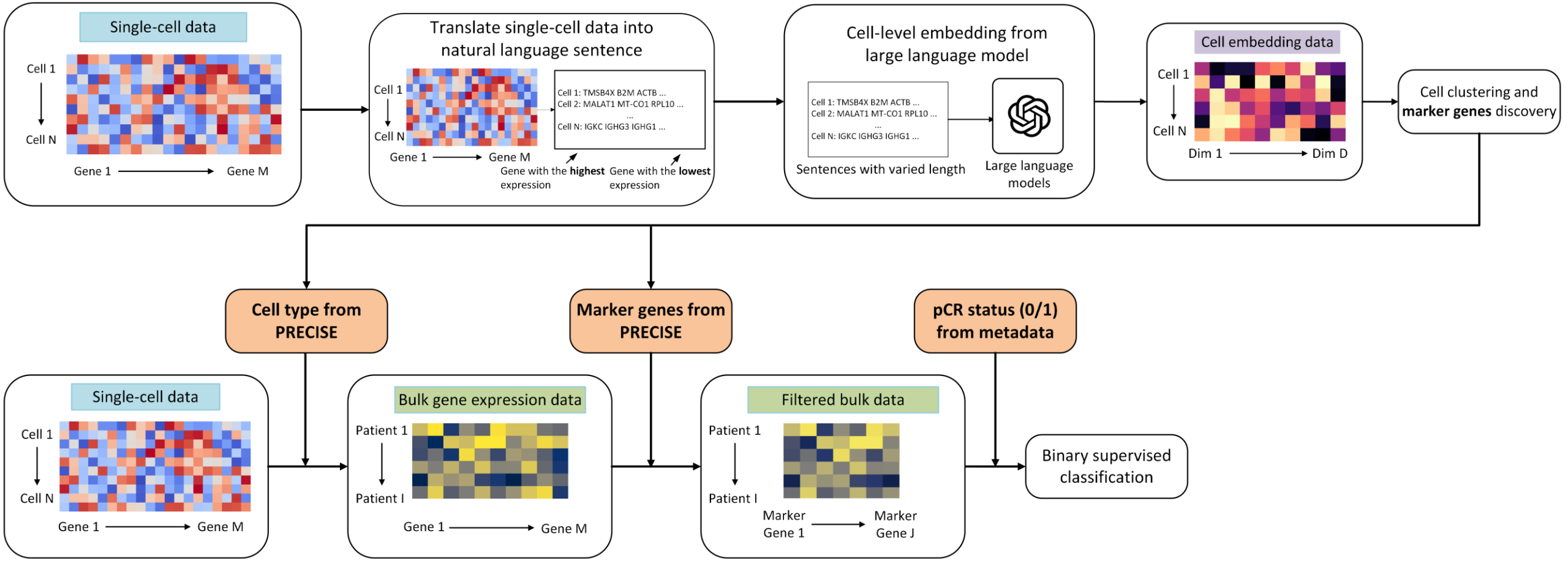
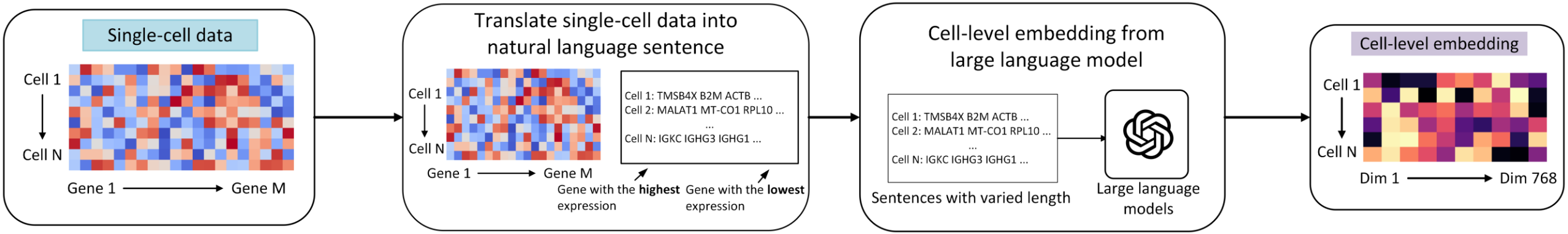


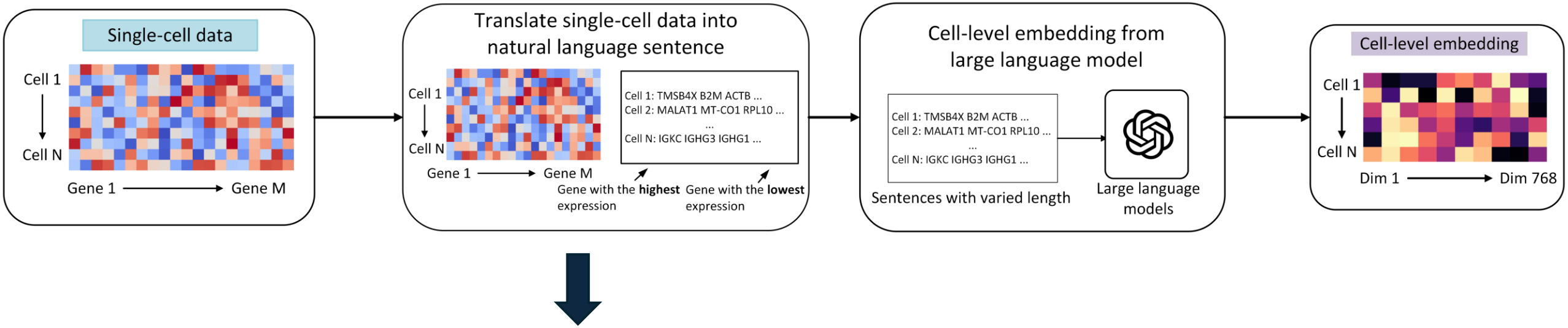
Figure 1: Overview of the PRECISE framework (**P**rediction of **RE**sponse using **C**ell-type **I**nference and **S**ingle-cell **E**mbedding), which leverages large language models and cell-type-specific markers for treatment outcome prediction

Generating Cell Embeddings with Large Language Models (LLMs)



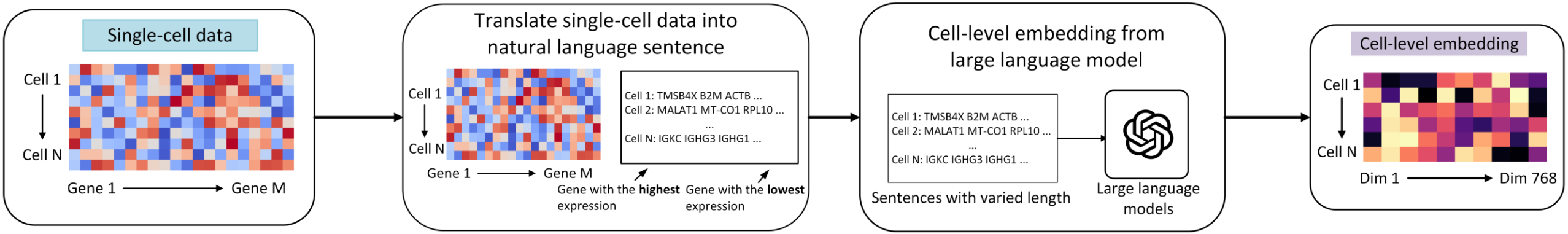
- Directly working on the sparse matrix
- Omit genes with no expression

Generating Cell Embeddings with Large Language Models (LLMs)



- Generate gene name sequences for each cell ranked by expression
- Varied length gene name sequence for each cell

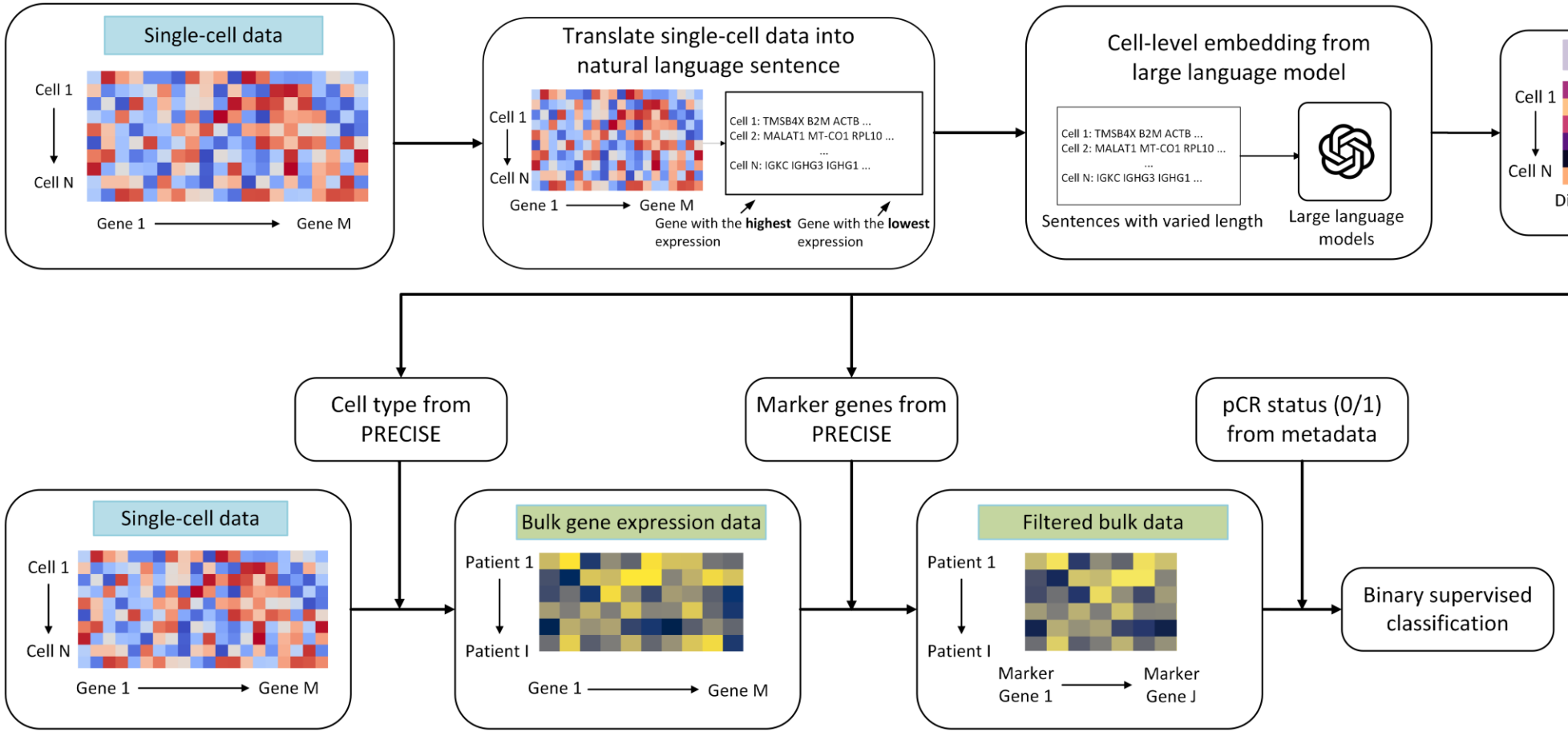
Generating Cell Embeddings with Large Language Models (LLMs)



- Model encodes each sequence into an embedding
- mxbai-embed-large-v1 (model parameter: 110M)

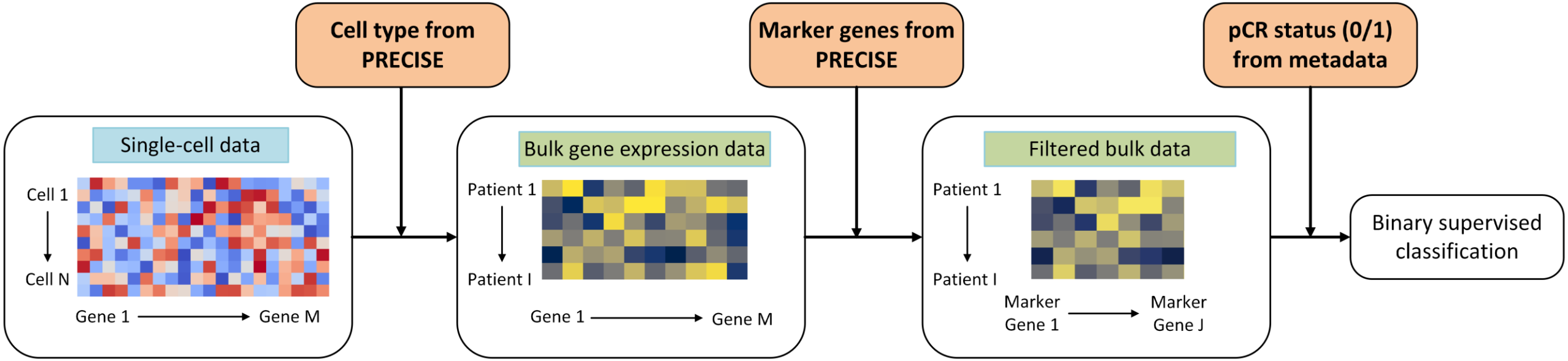
- Embedding dim: 768

Identifying Cell-Type-Specific Marker Genes

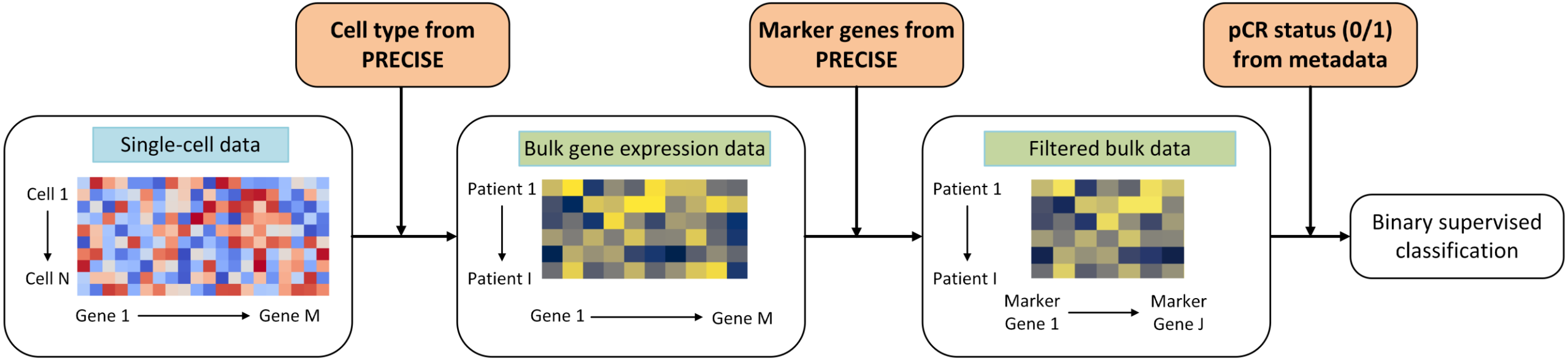


- **Cell clustering** was performed using the Louvain algorithm
- **Marker genes** for each cluster were identified using Seurat
- **Cell-type annotations** were assigned using *SingleR*

Predicting Treatment Outcomes from Bulk Gene Expression

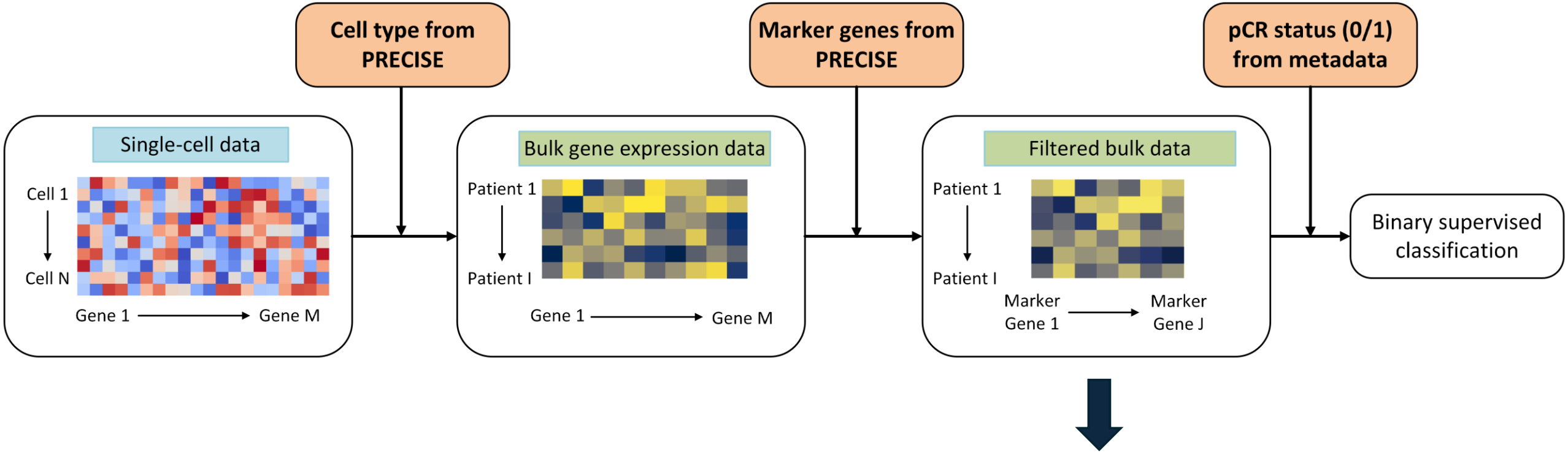


Predicting Treatment Outcomes from Bulk Gene Expression



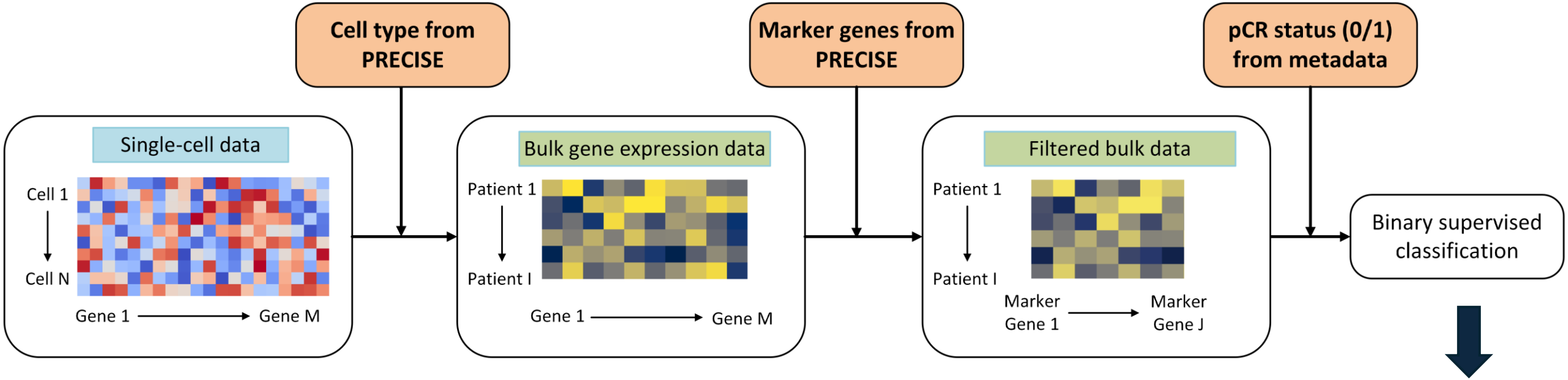
- Group cells by type for each patient
- Average gene expression within each group

Predicting Treatment Outcomes from Bulk Gene Expression



- Filter bulk expression matrix to keep only marker genes for prediction

Predicting Treatment Outcomes from Bulk Gene Expression



- Train classifiers using LASSO
- Evaluate performance using 3-fold stratified cross-validation

Results and Future Work

Discovery Dataset - Bassez et al. Cohort (2021)

- Study design:
  - Paired scRNA-seq data were collected before and after anti-PD1 treatment
  - scRNA-seq dimensions: 157,760 cells × 25,291 genes
- Motivation for our project:
  - Better understand the mechanisms underlying breast cancer by using **single-cell gene expression data**
  - Improve treatment outcome prediction for **anti-PD1 treatment**
  - Identify novel **marker genes** predictive of treatment response

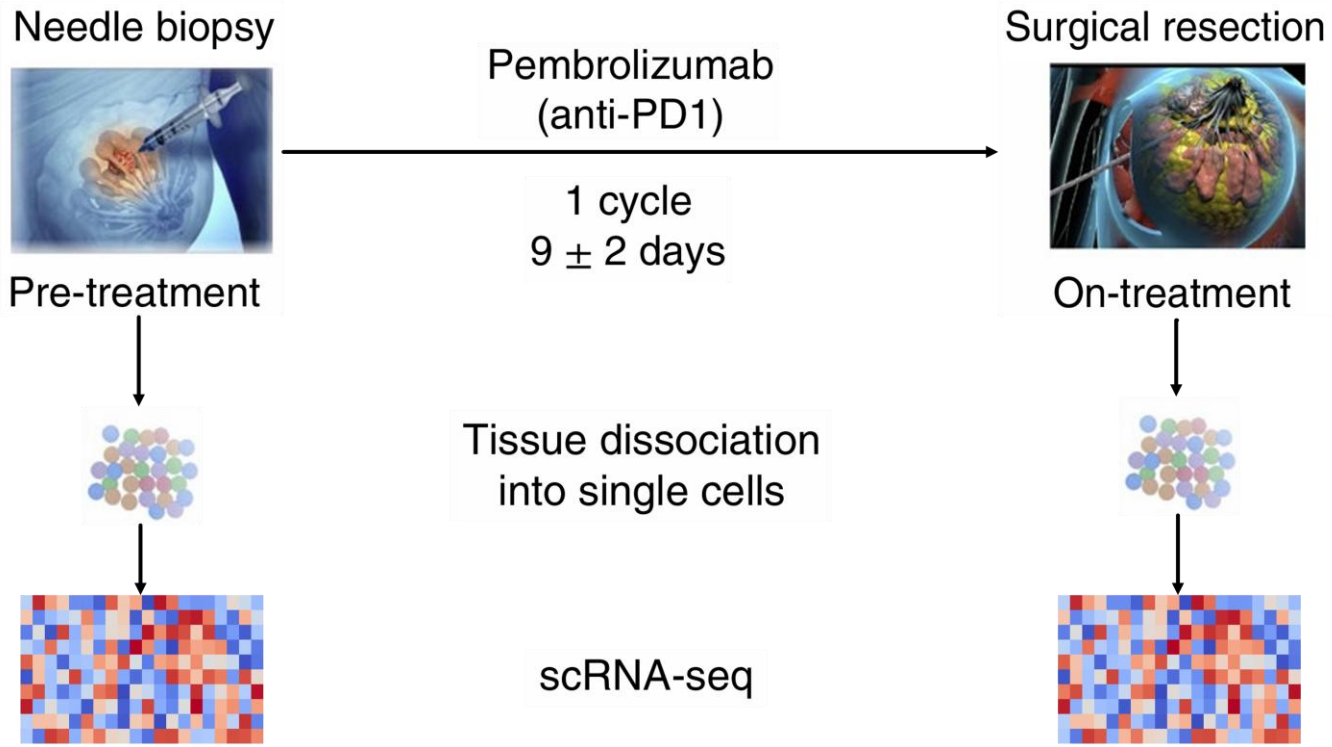


Figure 2: Sampling process of the discovery dataset (modified from Bassez et al., 2021)

Bulk Gene Expression Dataset for Treatment Response Prediction

- Discovery dataset: *Bassez et al. Cohort* (2021)
  - Treatment: anti-PD1 treatment (pembrolizumab + paclitaxel)
  - Sample size: **29 patients** (9 achieve pCR, 20 did not)
  - Bulk gene expression data: **aggregated from the cell-by-gene matrix using the PRECISE model**, 29 patients × 25,291 genes
- 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

Treatment Outcome Prediction: PRECISE vs. Seurat

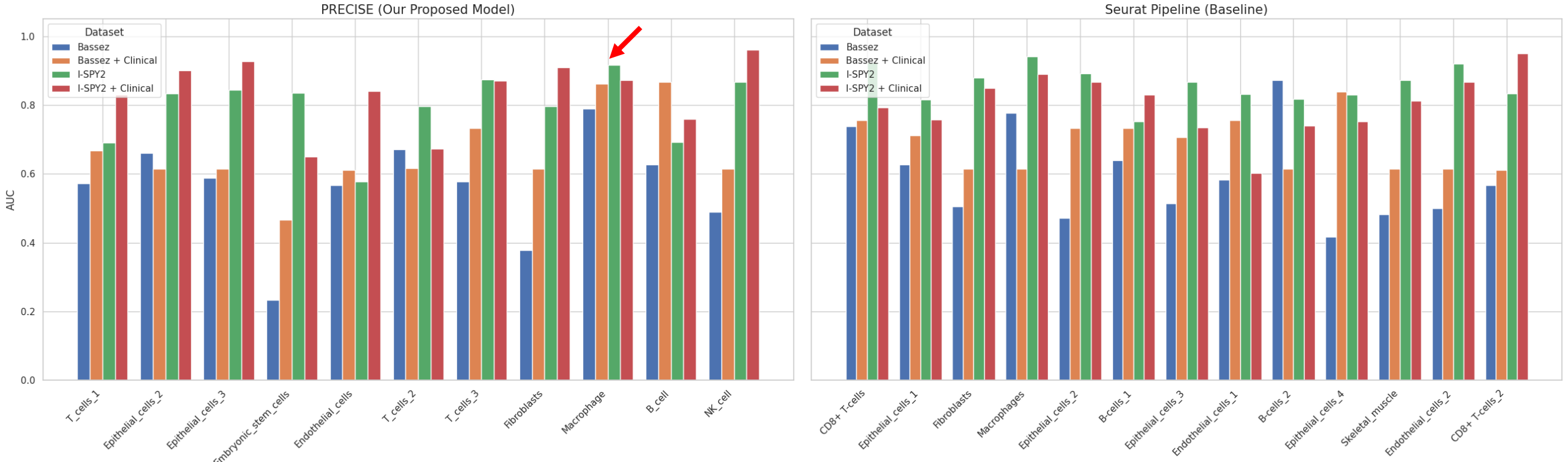


Figure 3: Cell type-specific treatment response prediction performance across models and datasets (LASSO)

PRECISE-identified **Macrophage** marker genes showed **consistently high AUC** across all datasets

- Strong performance on both **Bassez** and **I-SPY2** and Maintained performance even after incorporating **clinical covariates**
- Outperformed the baseline **Seurat pipeline** in nearly all conditions for macrophages
- Fine-tuned **XGBoost** and **SVM** models underperformed compared to **LASSO**, yielding lower AUC scores

Treatment Outcome Prediction: PRECISE vs InteractPrint

PRECISE-identified **Macrophage** marker genes **Outperforms Published Models for Treatment Outcome Prediction**

- PRECISE achieved **AUC = 0.861** on the Bassez et al. dataset (top left) and **AUC = 0.917** on the I-SPY2 trial (bottom left)
- Outperforms **PD-L1 expression** and **T Cell InteractPrint** baselines from Xu et al., 2024 (right panel)

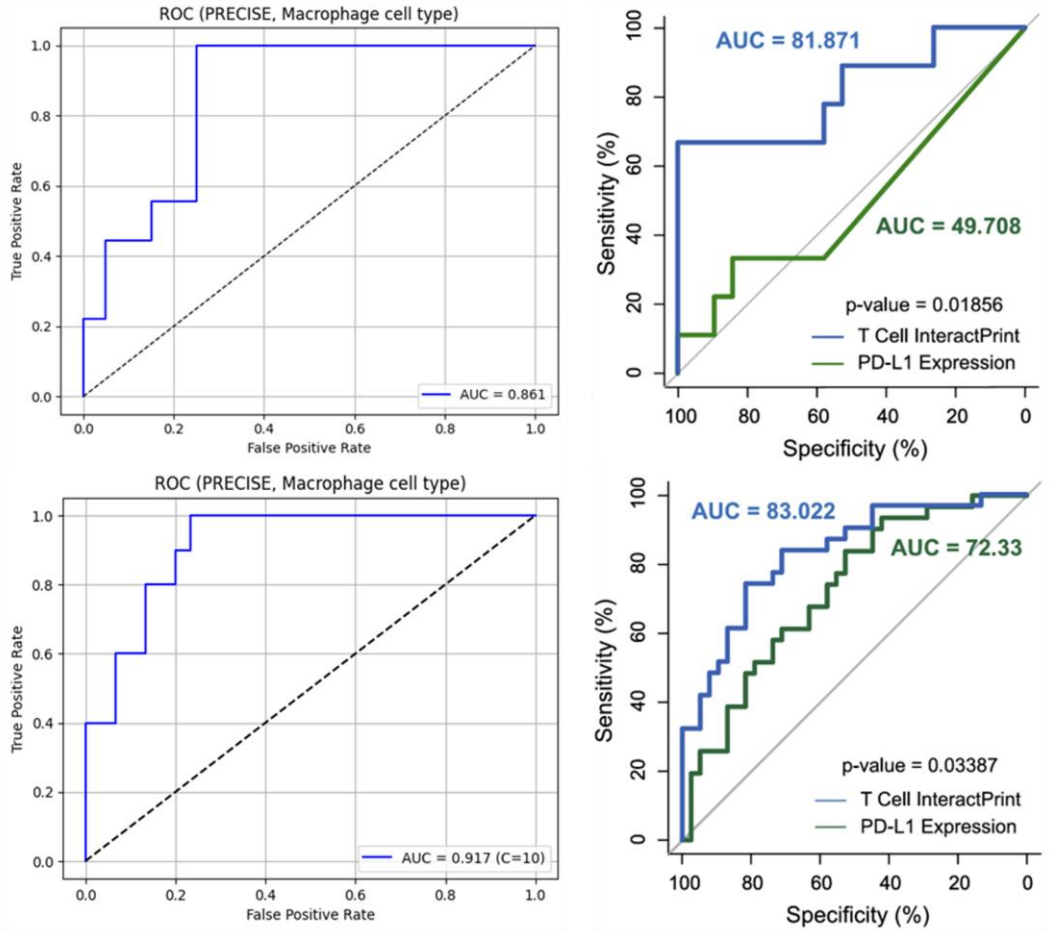


Figure 4: Comparison of PRECISE Macrophage-Based Prediction with Existing Published Models (**Left**: ROC curves from PRECISE using macrophage marker genes on the Bassez et al. dataset (top) and the I-SPY2 trial dataset (bottom). **Right**: ROC curves adapted from Xu et al., Cell Reports Medicine (2024) [https://doi.org/10.1016/j.xcrm.2024.101511]

Uncertainty Quantification of Treatment Response Prediction

Conformal prediction:

- **Motivation**: Beyond point prediction, we aim to quantify **uncertainty** in model outputs

Conformal histogram (right, e.g.):

- Majority of predictions are **high-confidence**
- **Only a few uncertain samples** fall near the decision threshold

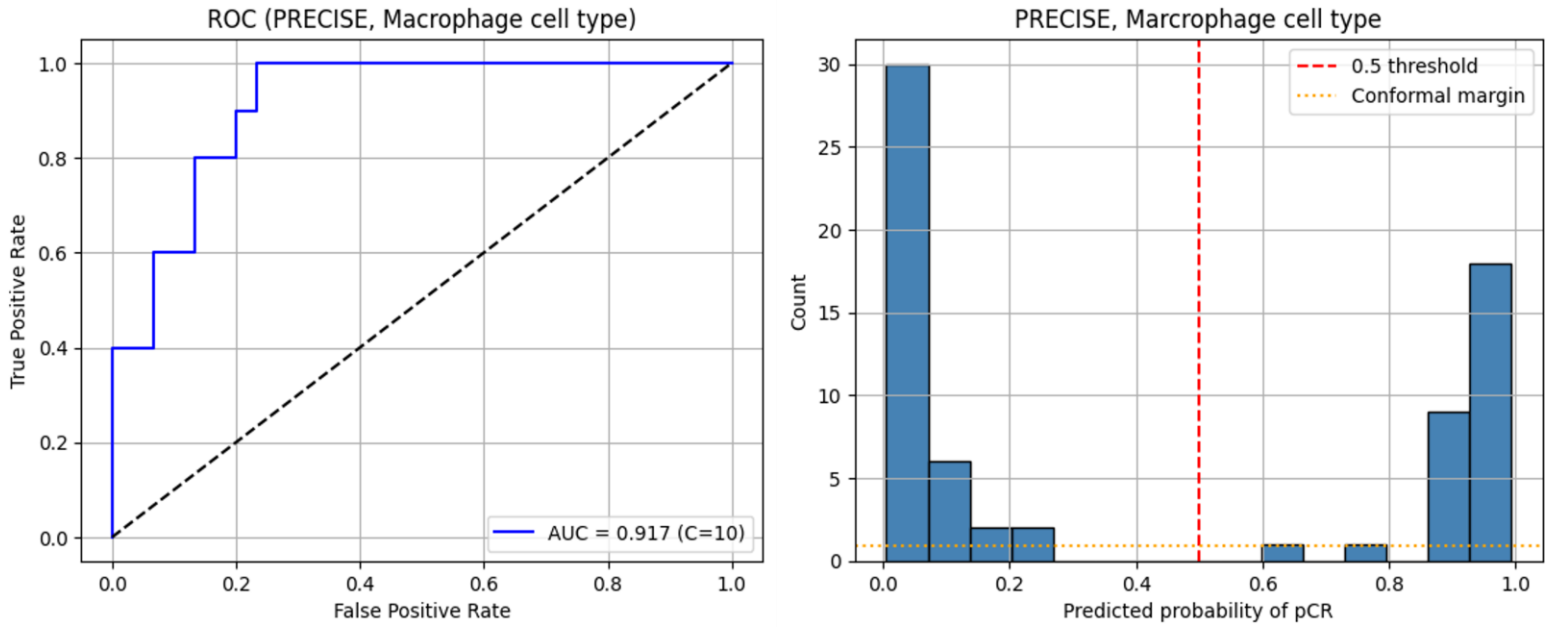


Figure 5: Quantifying prediction uncertainty with conformal prediction: macrophage markers in I-SPY2 trial
Left: *ROC curve (AUC = 0.917) using PRECISE-identified macrophage markers*
Right: *Conformal prediction at 90% confidence (Actual coverage: 0.97, Confident predictions: 67 / 69) illustrating model uncertainty*

Conclusion

1. **PRECISE's framework outperforms existing models**
   - Achieved higher AUCs than published and clinically used approaches across datasets
2. **Consistently better performance than Seurat pipeline**
   - Embedding-based features from foundation models generalize well in both discovery and validation cohorts
3. **Explored uncertainty quantification via conformal prediction**
   - Initial results demonstrate promise in flagging low-confidence predictions

Future Work

1. **Advance uncertainty quantification with conformal prediction**
   - Apply more rigorous conformal methods and validate confidence intervals across datasets
2. **Implement and benchmark published baselines**
   - Reproduce PD-L1 Expression and InteractPrint pipelines for formal statistical comparison (e.g., DeLong's test)
3. **Investigate clinical interpretability of key features**
   - Analyze top marker genes driving predictions for biological and clinical relevance

Reference

Bassez, A., Vos, H., Van Dyck, L., Floris, G., Arijs, I., Desmedt, C., Boeckx, B., Vanden Bempt, M., Nevelsteen, I., Lambein, K., Punie, K., Neven, P., Garg, A. D., Wildiers, H., Qian, J., Smeets, A., & Lambrechts, D. (2021). A single-cell map of intratumoral changes during anti-PD1 treatment of patients with breast cancer. Nature medicine, 27(5), 820–832. https://doi.org/10.1038/s41591-021-01323-8

Chen, Y., & Zou, J. (2024). GenePT: A Simple But Effective Foundation Model for Genes and Cells Built From ChatGPT. bioRxiv : the preprint server for biology, 2023.10.16.562533. https://doi.org/10.1101/2023.10.16.562533

DeSantis, C. E., Bray, F., Ferlay, J., Lortet-Tieulent, J., Anderson, B. O., & Jemal, A. (2015). International Variation in Female Breast Cancer Incidence and Mortality Rates. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 24(10), 1495–1506. https://doi.org/10.1158/1055-9965.EPI-15-0535

Li, X., & Li, J. (2023). Angle-optimized text embeddings. arXiv preprint arXiv:2309.12871

Wolf, D. M., Yau, C., Wulfkuhle, J., Brown-Swigart, L., Gallagher, R. I., Lee, P. R. E., Zhu, Z., Magbanua, M. J., Sayaman, R., O'Grady, N., Basu, A., Delson, A., Coppé, J. P., Lu, R., Braun, J., I-SPY2 Investigators, Asare, S. M., Sit, L., Matthews, J. B., Perlmutter, J., … van 't Veer, L. J. (2022). Redefining breast cancer subtypes to guide treatment prioritization and maximize response: Predictive biomarkers across 10 cancer therapies. Cancer cell, 40(6), 609–623.e6. https://doi.org/10.1016/j.ccell.2022.05.005

Xu, L., Saunders, K., Huang, S. P., Knutsdottir, H., Martinez-Algarin, K., Terrazas, I., Chen, K., McArthur, H. M., Maués, J., Hodgdon, C., Reddy, S. M., Roussos Torres, E. T., Xu, L., & Chan, I. S. (2024). A comprehensive single-cell breast tumor atlas defines epithelial and immune heterogeneity and interactions predicting anti-PD-1 therapy response. Cell reports. Medicine, 5(5), 101511. https://doi.org/10.1016/j.xcrm.2024.101511

Acknowledgments

Special thanks to Victoria Truong, Aoqi Xie, and Yu Shi for their support.

Digital Research Alliance of Canada

UNIVERSITY OF TORONTO
DALLA LANA SCHOOL OF PUBLIC HEALTH

Western
UNIVERSITY · CANADA

Dalla Lana
School of Public Health

Thanks

Any Questions?