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

**Abstract:**  
Breast cancer heterogeneity complicates the prediction of Pathologic Complete Response (pCR) to neoadjuvant therapy. Current models rely on bulk RNA-seq or predefined cell-cell interactions, limiting their ability to capture cellular diversity. This study integrates single-cell RNA sequencing (scRNA-seq) with embeddings from large language models (LLMs) and foundation models (e.g., scGPT) to identify cell type-specific biomarkers to predict pCR in breast cancer. Using scRNA-seq data from Xu et al. (2024) study, we generated cell-level embeddings and applied unsupervised clustering to the embedding data to uncover novel biomarkers. These biomarkers were used to build a cell-type specific classifier to predict the pCR of breast cancer in I-SPY2-990 cohorts. We achieved a significantly better performance than baseline models including InteractPrint and RPS-5. Our results demonstrate the potential of LLM-driven approaches in enhancing biomarker discovery for personalized breast cancer treatment.

**Title:** Amélioration de la prédiction de la réponse au traitement du cancer du sein grâce au séquençage de l’ARN unicellulaire et aux grands modèles de langage

**Abstract:**  
L'hétérogénéité du cancer du sein complique la prédiction de la réponse complète pathologique (pCR) à la thérapie néoadjuvante. Les modèles actuels reposent sur le séquençage de l’ARN en vrac (bulk RNA-seq) ou sur des interactions cellulaires prédéfinies, ce qui limite leur capacité à capturer la diversité cellulaire. Cette étude intègre le séquençage de l'ARN unicellulaire (scRNA-seq) avec des embeddings issus de grands modèles de langage (LLMs) et de modèles fondamentaux (par exemple, scGPT) afin d’identifier des biomarqueurs spécifiques aux types cellulaires pour prédire la pCR dans le cancer du sein. À partir des données de scRNA-seq de l’étude de Xu et al. (2024), nous avons généré des embeddings au niveau cellulaire et appliqué un regroupement non supervisé sur ces embeddings pour découvrir de nouveaux biomarqueurs. Ces biomarqueurs ont ensuite été utilisés pour construire un classificateur spécifique aux types cellulaires permettant de prédire la pCR du cancer du sein dans les cohortes I-SPY2-990. Nous avons obtenu des performances significativement meilleures que les modèles de référence, notamment InteractPrint et RPS-5. Nos résultats démontrent le potentiel des approches basées sur les LLMs pour améliorer la découverte de biomarqueurs et la personnalisation du traitement du cancer du sein.