Language Models in Spatial Omics: Mitigating Batch Effects and Advancing Immunotherapy Response Prediction in NSCLC through Geometric Deep Learning

Shay Shimonov
Department of Applied Math
CPSC 588
Yale University
shay.shimonov@yale.edu

1 Introduction

Spatially resolved transcriptomics and proteomics are pivotal in unveiling the intricacies of both physiological and pathological processes. These processes, ranging from embryonic development to cancer pathologies, are intricately linked to the spatial arrangement of tissues and cells. However, extrapolating universal conclusions from specific or disparate experiments is challenging, primarily due to batch effects. These effects, stemming from variations in experimental conditions, sample handling, and technical methodologies, can significantly skew data in spatial omics, masking genuine biological signals. Such variability complicates the comparison and integration of datasets from different studies, thereby obstructing a unified understanding of cellular behavior and interactions.

In single cell RNA sequencing, recent strides in computational biology have been directed towards mitigating batch effects through robust analytical frameworks. A key approach involves the use of machine learning models, particularly foundation models inspired by successes in domains such as natural language processing. With the increase in single cell RNA sequencing datasets, these methods become more relevant. However, replicating such foundation models for genes, which have no inherent order in cells, presents unique challenges. Some strategies, like scBert, GeneFormer, and UniCell [1, 2, 3], focus on leveraging biological knowledge and heuristics, while others integrate natural language processing to enhance data interpretation. For instance, Chen and Zou [4] utilized GPT-3.5 to generate gene embeddings based on textual gene descriptions, subsequently creating single-cell embeddings by averaging these gene embeddings, weighted according to gene expression levels. Leviene et al. [5] proposed embedding cells as sentences, using sequences of gene names ordered by expression level, processed by GPT-2. These methods have shown to outperform traditional approaches in certain scenarios and even reduce batch effects in some cases.

To my knowledge, these methodologies have not been explored in proteomics or spatial data contexts. Furthermore, the examination of their transferability has been limited, primarily to cell typing tasks, which are relatively basic. This study focuses on classifying tissue samples from lung cancer patients based on their response to immunotherapy. Contemporary methods have incorporated Graph Neural Networks (GNN) to predict phenotypes from tissue samples [6, 7]. In this research, I employ SORBET, a geometric deep learning framework, to deduce emergent phenotypes, such as immunotherapy response, from spatial molecular profiling data. Tissues are modeled as graphs of adjacent cells, utilizing graph convolutional networks (GCN) [8] for classification. This is applied to two cohorts from non-small cell lung cancer (NSCLC) studies [9], profiled using Imaging Mass Cytometry (IMC) [10]. Despite similarities in phenotype, technology, and tissue type, these cohorts were derived from distinct labs under varying conditions, exhibiting significant batch effects. Previous

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

attempts to develop a unified model for outcome prediction using these cohorts were unsuccessful [9].

This study compares three methods for representing cell embedding in the graph-based tissue model: original raw data, protein summary averaged embedding, and cell sentences. These embeddings serve as the foundation for the SORBET GCN classifier. After optimizing hyperparameters on one cohort, I train the model on the entire cohort and test it on the other to evaluate the transferability of the three methodologies in the context of immunotherapy response prediction.

2 Related Work

SORBET: SORBET is a geometric deep learning framework used for classifying spatially resolved single-cell data. It categorizes data into two groups based on binarized outcomes from shared endpoints. The input includes endpoint labels, spatial coordinates of cells, and their molecular expressions. SORBET represents tissues as graphs, with cells as nodes linked by their physical contact, capturing cell-specific features and tissue structure. It focuses on tissue subregions, identified by an algorithm based on prior knowledge of the phenotype, enabling detailed analysis of cellular environments. This approach increases sample size for machine learning and offers localized biological insights, balancing computational efficiency with biological relevance.

Cell2Sentence [5]: is a method that converts single-cell gene expression profiles into "cell sentences." This process involves ranking the expression of gene names and forming sentences based on this order. Additionally, these cell sentences can be annotated with biological metadata like cell type, tissue, or disease. LLMs are then fine-tuned using these cell sentences. For inference, the model can generate new cells through autoregressive cell completion, create cells from textual input, or produce text from cell data. The cell sentences generated through this process can subsequently be transformed back into gene expression profiles.

GenePT:[4] GenePT employs ChatGPT embeddings to represent genes and cells. It utilizes gene descriptions from the NCBI database, processed with GPT-3.5, to create gene embeddings. GenePT generates single-cell embeddings by averaging gene embeddings, weighted by gene expression levels, or by creating sentence embeddings from ordered gene names. Despite its efficiency and ease of use, GenePT has limitations in capturing specific gene functionalities and may lack specificity for certain tissue types, given its reliance on existing gene summaries and the capabilities of GPT-3.5.

All-mpnet-base-v2: All-mpnet-base-v2 is a Bert like model designed for encoding sentences and short paragraphs, transforming input text into a vector that encapsulates its semantic content. This encoding is particularly useful for applications in information retrieval, clustering, and sentence similarity tasks. The model is a fine-tuned version of the pre-trained Microsoft MPNet-base, leveraging the strengths of the original model with additional refinements. It stands out as the highest quality model within the SBERT all family. The training of all-mpnet-base-v2 involves a large-scale dataset of 1 billion sentence pairs, using a self-supervised contrastive learning objective. This approach requires the model to identify the correct sentence pairings from a set of randomly sampled sentences, thus enhancing its ability to understand and represent textual meaning effectively.

Graph of Graphs: This technique involves a two-step tissue classification process. Initially, a random walk on the data-derived graph is performed until a steady-state distribution is reached. Then, this distribution is embedded, and the embedding space is used for data classification. Despite its application to the same cohorts as SORBET, this method did not yield a unified model due to batch effects.

3 Data

This study utilizes patient-derived non-small cell lung cancer (NSCLC) tissues, as part of a previously described dataset. Protein abundances in these tissues were quantified using imaging mass cytometry (IMC). The dataset is divided into two clinically distinct cohorts: Cohort 1 (C1) consisting of 29 patients (11 responders) and Cohort 2 (C2) with 41 patients (20 responders).

Anchor cells in the dataset were identified using the pan-cytokeratin (PanCK) tumor marker. Subgraphs representative of cellular interactions were extracted using the designated subgraph extraction

algorithm. For each patient, a number of subgraphs were derived. These subgraphs, with pre-annotated cell contacts, form the basis of our analysis of tissues as graphs.

The primary objective is to classify patients' responses to PD-L1 therapy, categorized into two groups: 'responders' (those with a durable clinical response) and 'non-responders' (those without a durable response). To evaluate the performance of our model, we employed five-fold cross-validation. In Cohort 1, this resulted in an average distribution of 23 patients for training, 3 for testing, and 3 for validation. Similarly, in Cohort 2, the distribution was 33 patients for training, 4 for testing, and 4 for validation. Each patient was uniquely assigned to a single test set, ensuring an unbiased evaluation of the model's predictive capabilities.

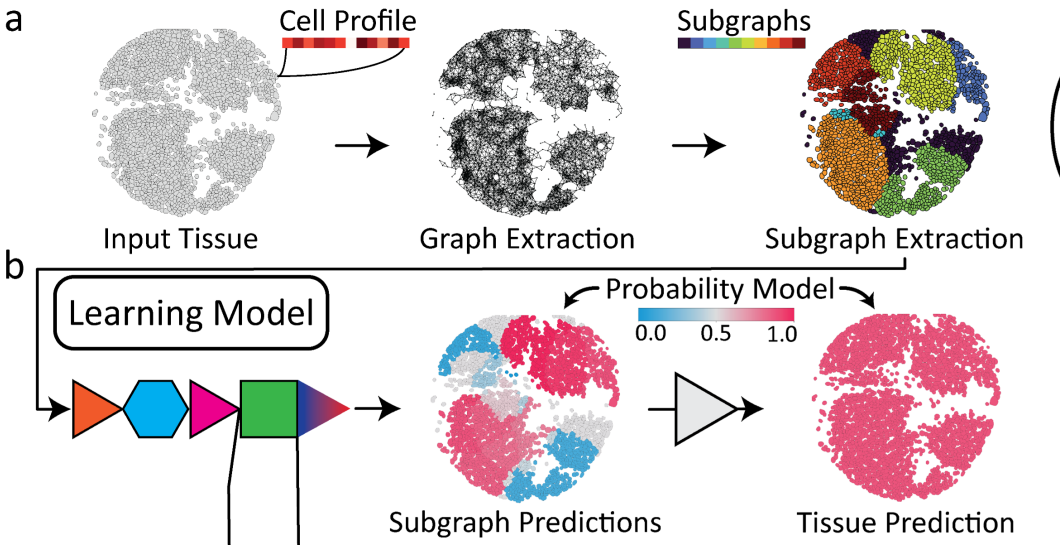

Figure 1: **SORBET Method Overview (a-b)** SORBET Phases: Graph Modeling, Learning. **(a)** Graph Modeling phase. Input tissues (left), which are defined by spatially-resolved single cell profiles, are modeled as graphs (middle). From the tissue graph, numerous subgraphs are extracted using an automated subgraph extraction algorithm. **(b)** Learning phase. A graph convolutional network (GCN)-based classifier classifies subgraphs ('Subgraph Prediction', left) and, from these subgraph predictions (middle), a composite tissue prediction is made (right).

4 Method

This study employs three distinct methods to generate the underlying graph features for tissue classification based on immunotherapy response. These methods are:

1. **Raw Features:** As a baseline, the raw expression levels of markers form the underlying graph. After omitting four control markers, the counts are log-normalized and then range normalized between 0 and 1. This normalization strategy aims to mitigate batch effects. The resultant node feature matrix is denoted as $\mathbf{X}_{\text{raw}} \in \mathcal{R}^{N_c, N_p}$, where N_p represents the number of proteins, and N_c is the number of cells.
2. **Protein Summaries Average Embedding:** The UniProt database is utilized for its extensive protein sequence and annotation data. For each marker, a brief summary describing its function is extracted from UniProt. These protein summaries are then embedded using the All-mpnet-base-v2 sentence transformer, resulting in an embedding dimension of $d = 768$. This process produces a matrix $\boldsymbol{E}_p \in \mathcal{R}^{N_p \times d}$, representing the embedded summaries for each protein. The final node feature matrix for each cell, $\mathbf{X}_{\text{summary}}$, is obtained by averaging the protein embeddings in accordance with their expression levels, thus forming a matrix $\mathbf{X}_{\text{summary}} = \mathbf{X}_{\text{raw}} \boldsymbol{E}_p \in \mathcal{R}^{N_c, d}$.

3. **Cell Sentences:** For each cell, a list of expressed markers is compiled and sorted by expression levels to create 'cell sentences.' These sentences are then encoded using the All-mpnet-base-v2 transformer, generating a node feature matrix $\mathbf{X}_{\text{cell sentence}} \in \mathcal{R}^{N_c,d}$ for each cell.

Following the creation of these underlying graphs, they are integrated into the SORBET pipeline. The primary objective is to use these graph representations to accurately classify tissues based on their response to immunotherapy.

5 Results

The experimental setup was designed to evaluate the performance of three different feature representations: original expression, protein summary, and cell sentences. The procedure for each feature type was as follows:

1. **Hyperparameter Optimization:** Utilizing Ray and Optuna, a hyperparameter search was conducted in a 5-fold cross-validation setup. This was performed within an optimal hyperparameter space, determined from prior experience. A total of forty different models were evaluated on Cohort 1.
2. **Model Selection and Retraining:** The best-performing model from the hyperparameter search was selected. This model, with the optimized hyperparameters, was then retrained on the entirety of Cohort 1.
3. **Testing for Transferability:** The retrained model's performance was tested on Cohort 2 to assess the transferability between cohorts. This process was then repeated by training on Cohort 2 and testing on Cohort 1.

Performance metrics included ROC-AUC score, calibration, conditional predictions, precision-recall curve, and predicted distribution analysis. The ROC-AUC results are summarized in Table 1.

Method	Cohort 1 (C1)		Cohort 2 (C2)	
	Tuning (C1)	Test on C2	Tuning (C2)	Test on C1
Original Expression	0.817	0.333	0.620	0.576
Protein Summary	0.733	0.417	0.665	0.217
Cell Sentence	0.761	0.387	0.744	0.281

Table 1: **ROC-AUC Results for Transferability of Different Methods**

The general trend observed, as depicted in Figure 2, was a lack of generalizability to out-of-distribution data. The models struggled with batch effects, as indicated by the significantly reduced performance in cross-cohort testing.

6 Discussion

The concept of employing natural language models to interpret and enrich biological data presents an intriguing yet complex challenge. Previous studies have indicated that using natural language modeling for cell representation may aid in reducing batch effects. These findings, however, were based on simplified tasks within specific contexts. In our study, we observed that these methods did not effectively address batch effects or improve out-of-distribution generalization, pointing to limitations in the modeling approaches employed.

One significant limitation pertains to the 'Cell Sentences' method. This approach sacrifices much of the expressive potential of the data by discarding exact expression levels and relying solely on rank order. Additionally, it incorporates only the top 20 markers, further constraining its expressive capacity. Another critical aspect is that the sentence transformer used was not trained on biological data, potentially limiting its effectiveness in this specific domain.

A similar challenge arises with the 'Protein Summaries Average Embedding' method. This approach assumes that the embedding space of protein summaries is structured such that a weighted average of

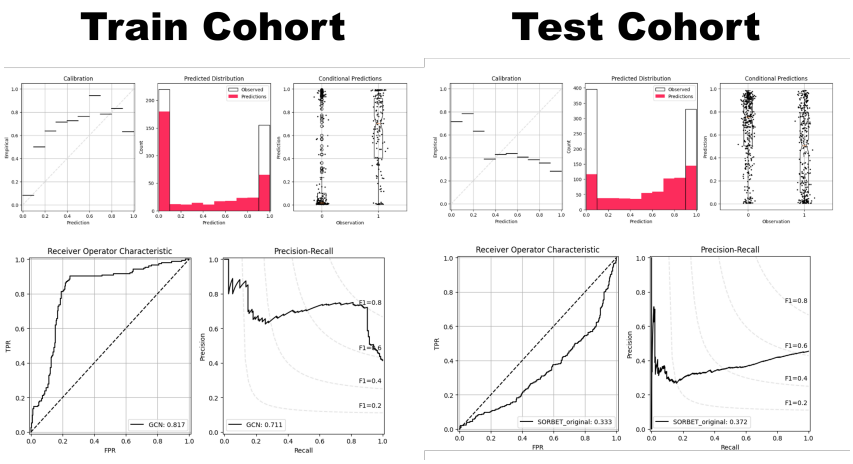

Figure 2: General Out-of-Distribution Behavior of the Models

these embeddings yields meaningful results. However, this is not typically the case, especially when the model has not been fine-tuned on related tasks. Moreover the method confines us to proteins that have been extensively researched and documented. Furthermore, summaries may vary in accuracy and detail, depending on the protein and the source.

Despite these challenges, pursuing this direction remains promising and holds significant potential for impact. Key questions arise: What are the appropriate pretraining or fine-tuning tasks that could mitigate these shortcomings? How can we develop a 'universal language' for proteins that encompasses a meaningful embedding space? Successfully addressing these questions, as our study begins to show, has the potential to resolve batch effect issues in biological data analysis, marking a significant advancement in the field.

7 Code Availability

The source code developed and utilized for this study is publicly accessible in `https://github.com/shikok/CPSC588`.

8 Data Availability

The data used in this study comprises patient-derived information. In recognition of the sensitivity and confidentiality of patient data, access to this dataset will be granted upon reasonable request.

References

[1] Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866, 2022.

[2] Zhanbei Cui, Yu Liao, Tongda Xu, and Yan Wang. Geneformer: Learned gene compression using transformer-based context modeling. *arXiv preprint arXiv:2212.08379*, 2022.

[3] Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorcan, Tabula Sapiens Consortium, Stephen R Quake, and Jure Leskovec. Universal cell embeddings: A foundation model for cell biology. *bioRxiv*, pages 2023–11, 2023.

[4] Yiqun T Chen and James Zou. Genept: A simple but hard-to-beat foundation model for genes and cells built from chatgpt. *bioRxiv*, pages 2023–10, 2023.

[5] Daniel Levine, Syed Asad Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, Ruiming Wu, Insu Han, Zihe Zheng, Antonio Henrique de Oliveira Fonseca, Xingyu Chen, Sina Ghadermarzi,

et al. Cell2sentence: Teaching large language models the language of biology. *bioRxiv*, pages 2023–09, 2023.

[6] Zhenqin Wu, Alexandro E. Trevino, Eric Wu, Kyle Swanson, Honesty J. Kim, H. Blaize D'Angio, Ryan Preska, Gregory W. Charville, Piero D. Dalerba, Ann Marie Egloff, Ravindra Uppaluri, Umamaheswar Duvvuri, Aaron T. Mayer, and James Zou. Graph deep learning for the characterization of tumour microenvironments from spatial protein profiles in tissue specimens. *Nature Biomedical Engineering*, 6(12):1435–1448, December 2022.

[7] Yanan Wang, Yu Guang Wang, Changyuan Hu, Ming Li, Yanan Fan, Nina Otter, Ikuan Sam, Hongquan Gou, Yiqun Hu, Terry Kwok, John Zalcberg, Alex Boussioutas, Roger J. Daly, Guido Montúfar, Pietro Liò, Dakang Xu, Geoffrey I. Webb, and Jiangning Song. Cell graph neural networks enable the precise prediction of patient survival in gastric cancer. *npj Precision Oncology*, 6(1):1–12, June 2022.

[8] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

[9] Ya-Wei Eileen Lin, Tal Shnitzer, Ronen Talmon, Franz Villarroel-Espindola, Shruti Desai, Kurt Schalper, and Yuval Kluger. Graph of graphs analysis for multiplexed data with application to imaging mass cytometry. *PLoS computational biology*, 17(3):e1008741, 2021.

[10] Qing Chang, Olga I. Ornatsky, Iram Siddiqui, Alexander Loboda, Vladimir I. Baranov, and David W. Hedley. Imaging mass cytometry. *Cytometry Part A*, 91(2):160–169, 2017.