Analysis of Therapeutic Response in Hepatocellular Carcinoma with Graph-based Machine Learning Methods on Spatial Transcriptomics Data

Hannah Thomas
Nashua High School South

Aarthi Venkat
Eric and Wendy Schmidt Center at the
Broad Institute of MIT and Harvard

Abstract

Spatial transcriptomics (ST) enables the investigation of tissue organization by preserving spatial context while profiling gene expression. However, understanding how gene-gene interactions differ between tissue phenotypes remains challenging due to sparsity and noise in gene measurements and patient-specific differences. In this study, we analyze Visium spatial gene expression data from hepatocellular carcinoma (HCC) tumor microenvironments (TME) to compare spatial gene-gene networks between responder and non-responder patient groups. Using recently developed methods, spARC and GSPA, we derived gene embeddings, which represent functional similarity between genes in terms of both transcriptional and spatial concordance, and computed localization scores for each gene, which capture the specificity of enrichment patterns. We found that pathways associated with structural remodeling were enriched among genes uniquely localized in responders, whereas pathways associated with extracellular matrix (ECM)-driven resistance and epithelial-mesenchymal transition (EMT) were enriched in non-responders. We additionally created gene-gene similarity graphs utilizing embedding values and analyzed the network structure differences between responder and non-responder samples using the Jaccard Similarity Index, Uniform Manifold Approximation and Projection (UMAP), and random forest classification of gene-centered ego-subgraphs. Furthermore, we introduced a novel metric, the differential response score based on mean Euclidean distance, to further quantify differences between the two groups. Collectively, these analyses indicate that responder and non-responder tissues have distinct molecular programs, particularly involving ECM interactions, immune regulation, and cell migration. Our findings demonstrate that integrating spatial context into network-based analyses can uncover phenotype-specific molecular organization in the HCC TME, uncovering spatial features that shape therapeutic efficacy.

1 Introduction

Biological systems in multicellular organisms are highly complex, with tissues made up of a variety of specialized cells collaborating to perform essential biological processes. The position of cells within a tissue and their interactions with neighboring cells and the surrounding microenvironment play a critical role in regulating their behavior [1]. Spatial context is particularly important for understanding processes such as immune responses, tissue regeneration, and the progression of complex diseases, including cancer [2]. Understanding the tumor microenvironment (TME) is especially challenging because tumors are composed of diverse cell types whose interactions with one another evolve continuously. These interactions are highly heterogeneous between patients and even in different regions of the TME, making it difficult to capture the full complexity with conventional bulk or single-cell approaches [3]. Without spatial context, we cannot fully understand how cells communicate in different biological states. Spatial Transcriptomics (ST) addresses this issue by capturing gene expression while preserving the spatial coordinates of each measurement across a tissue. This allows researchers to map where specific genes are being expressed and differentiate spatial domains [1].

A previously published study conducted by Zhang et al [4]. has demonstrated the utility of spatial transcriptomics for investigating how the tumor architecture influences therapeutic response in hepatocellular carcinoma (HCC). In particular, Zhang et al. characterized distinct molecular and cellular features of the TME that distinguished patients who responded to neoadjuvant cabozantinib and nivolumab therapy from those who did not, highlighting differences in immune cell activity, extracellular matrix (ECM) structure, and expression of cancer stem cell markers [4]. In this study, we reanalyze this publicly available Visium ST dataset [4] to extend these findings by examining how spatial gene-gene interaction networks differ between responders and non-responders. Since gene networks are still poorly understood in cancer, and network-level approaches can capture higher-order relationships between genes while diminishing noise that can affect individual gene measurements, our work offers insights beyond those obtainable from standard ST profiling, which may reflect important pathways involved in cancer immunotherapy response.

Revealing gene networks from spatial data remains challenging, so to achieve this we are leveraging two recently developed methods: spARC (spatial Affinity-graph Recovery of Counts) [5] and GSPA (Gene Signal Pattern Analysis) [6]. spARC is a graph-based method that denoises and refines ST data by diffusing gene expression signals between cells with similar spatial location and expression profiles, which can reveal gene co-expression and simulate communication between cells. GSPA is a method that represents gene measurements as graph signals and uses graph signal processing to create rich gene embeddings given a gene expression matrix. The resulting embeddings characterize the gene expression patterns and spatial locations.

While spARC and GSPA have both shown early success in capturing gene-gene relationships in various contexts, their applications remain limited, and their potential in diverse biological contexts has yet to be explored. By applying and adapting these methods, we gained insights into how spatial gene expression patterns underlie cancerous tissues. In particular, pathways related to extracellular structural organization, epithelial–mesenchymal transition (EMT), tumor cell and ECM signaling, and the MHC protein complex were found to have biological significance in HCC therapy response. This work contributes to a growing effort to develop spatially aware analysis techniques that move beyond cataloging gene expression levels and begin to uncover how the spatial architecture of tissues contributes to overall function.

2 Methods

To explore how gene interaction patterns differ between responder and non-responder tissues, we designed a computational workflow that integrates data preprocessing, clustering, network analysis, dimensionality reduction, and machine learning. Starting with ST data from seven HCC tissue samples (4 responders, 3 non-responders to treatment), we identified highly variable genes and ran two algorithms, spARC and GSPA, to perform manifold learning and obtain gene embeddings, which are numerical vectors capturing functional similarities between genes. With the embeddings, we derived gene-gene networks, from which we extracted subgraphs centered at each gene to represent the local interaction landscape of the gene. We then visualized these subgraph embeddings using UMAP to examine whether responders and non-responders showed distinct neighborhood structures, and measured network similarity using the Jaccard Similarity Index and mean Euclidean distance. Finally, we trained a Random Forest Classifier to test whether responder status could be predicted from a gene's ego subgraph, and used pathway enrichment analysis on the most informative genes to highlight biological processes potentially linked to treatment response. The overall methodology is summarized in Figure 1.

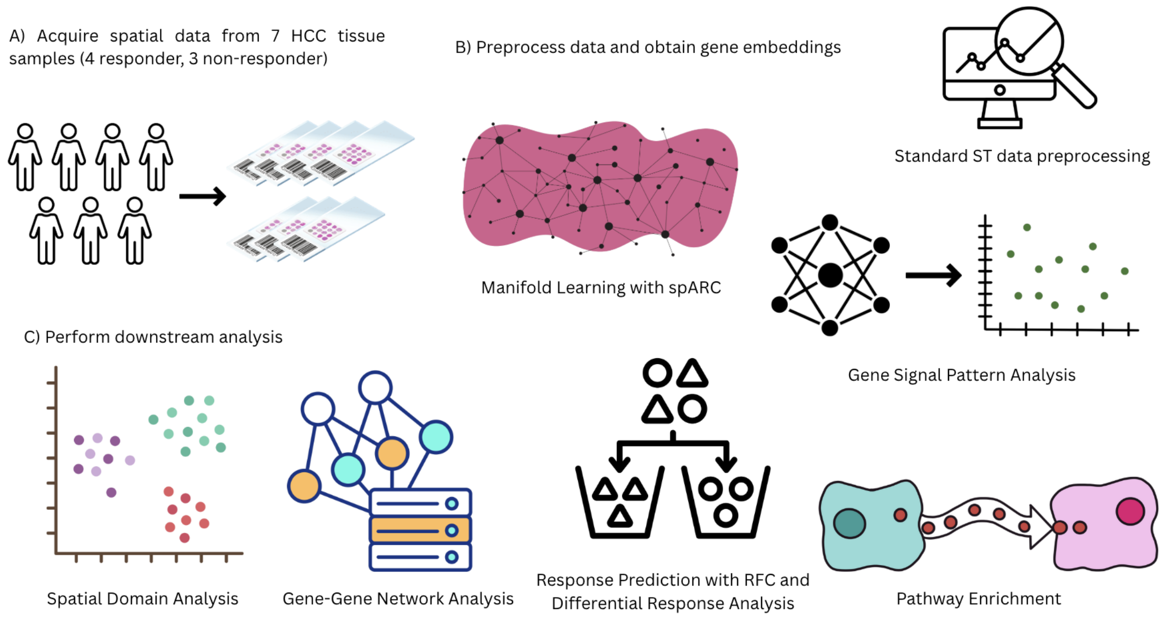
A) Acquire spatial data from 7 HCC tissue samples (4 responder, 3 non-responder)
B) Preprocess data and obtain gene embeddings
Standard ST data preprocessing
Manifold Learning with spARC
Gene Signal Pattern Analysis
C) Perform downstream analysis
Spatial Domain Analysis
Gene-Gene Network Analysis
Response Prediction with RFC and Differential Response Analysis
Pathway Enrichment

Figure 1: An outline of the methodology. 1. Acquire 10x Visium data of HCC TME from 4 responders and 3 non-responders to treatment 2. Preprocess data using standard filtration and normalization techniques. Utilize spARC and GSPA to perform manifold learning and obtain rich gene embeddings. 3. Downstream analyses include performing Leiden clustering, comparing responder and non-responder gene-gene networks, classifying ego subgraphs as from a responder or non-responder network using an RFC, and conducting enrichment analysis on discriminative genes across graphs.

2.1 Acquiring the Datasets

To compare responder and nonresponder HCC microenvironments, we collected ST data from a study conducted by Zhang et al. that investigated the effects of two cancer treatment drugs, cabozantinib and nivolumab, in neoadjuvant therapy on advanced HCC [7]. Specifically, we obtained 10x Visium data from seven patient samples (4 responder, 3 nonresponder) that were acquired through R0 resection. The process by which Zhang et al. procured the samples received approval from the institutional review board of Johns Hopkins University, and the ST data of all seven surgical HCC specimens passed standard quality control parameters [4].

2.2 Preprocessing the Data

All 10x Visium datasets share a standardized structure that supports integrated ST analysis. Each dataset includes:

- A gene expression count matrix, in which genes represent rows and barcoded spatial spots represent columns.
- A histology image of the stained tissue section.
- A spatial metadata file that maps each spot to its (x, y) coordinates on the histology image and indicates whether the spot falls within the tissue.

These components enable the precise overlay of gene expression data on tissue morphology, allowing for downstream spatial analyses such as clustering and cell-type deconvolution.

Prior to analysis, we applied a series of preprocessing steps to ensure data quality and reliability. Mitochondrial genes were first annotated to help identify potentially low-quality spots. This is because when cells are stressed, dying, or ruptured during tissue processing, cytoplasmic and nuclear RNA degrades faster than mitochondrial RNA [8]. Thus, having a high percentage of mitochondrial RNA is an indicator that the spot contains dying cells, broken membranes, or degraded RNA, so it might not reflect the tissue's true transcriptomic state. This was followed by the calculation of standard quality control (QC) metrics using the scanpy method `calculate_qc_metrics`, which returns various observation-level metrics and variable-level metrics to add to the dataset [9]. Spots with less than 200 detected genes, often representing empty or degraded tissue regions, were filtered out. We then applied cell-depth normalization and log transformation to correct for technical variability in sequencing, which is the standard single-cell RNA sequencing (scRNA-seq) normalization approach. This technique works by dividing the observed transcript count by the total number of RNAs detected in that cell, multiplying by a scaling factor (10,000), and performing a log transformation to decrease the influence of highly expressed genes [10]. However, unlike in scRNA-seq datasets, each spot in our dataset captures transcripts from a multicellular region, so total transcript counts may have significant biological meaning, such as tissue density or metabolic activity. As a result, the optimal approach for normalizing spatial transcriptomic data is still an open area of research, and all normalization methods must be applied with caution [11]. Finally, we conducted principal component analysis (PCA) to reduce the data dimensionality and eliminate noise prior to manifold learning. PCA identifies principal components (PCs), which are the variables that capture the most variance in the dataset, and projects the high-dimensional gene expression matrix onto these PCs. This denoises the data while retaining meaningful information from the dataset [12].

2.3 Leiden Clustering and Selecting SVGs

Leiden clustering was applied to outline spatial domains, which are regions within the tissue that exhibit similar gene expression profiles. While Leiden clustering is commonly used in the ST community to group spots or cells based solely on transcriptional similarity, our procedure incorporated both spatial coordinates and gene expression data. This allowed the identification of domains that are not only transcriptionally distinct but also spatially localized within the tissue architecture [11].

PCs derived from the gene expression matrix were integrated with spatial coordinates, and both were standardized to ensure equal contribution from both the gene expression profiles and spatial context in the Leiden clustering process. The resulting clusters were projected onto the tissue image to delineate spatial domains of interest. Then, spatially variable genes (SVGs) were identified to determine which genes exhibited expression patterns that varied significantly across spatial locations within the tissue. This was accomplished by calculating Moran's I statistic, which assesses spatial autocorrelation in gene expression patterns. The steps performed were consistent with the methods described by Merchan [11].

2.4 Manifold Learning with spARC (spatial Affinity-graph Recovery of Counts)

In order to find key differences between the responder and non-responder tissue data, we decided to implement diffusion-based manifold learning. Manifold learning is a type of unsupervised representation learning that uncovers the underlying geometric structure, or manifold, of the data by creating lower-dimensional representations that preserve the essential geometric relationships present in the data [5]. Manifold learning relies on the assumption that high-dimensional data can be modeled as a collection of smoothly varying locally Euclidean neighborhoods of low dimensionality [13]. This assumption aligns with the high dimensionality of gene expression data and the fact that cells exist in spatially structured, continuous environments with gradually changing gene expression patterns, making manifold learning particularly effective for ST data.

Manifold learning was conducted via a data diffusion-based method, spARC [5]. In order to model signaling interactions in the ST data and reveal gene-gene interactions, spARC shares information between both spatially adjacent cells and transcriptionally similar cells in a process mathematically analogous to diffusing heat throughout the data. When spARC is applied to our patient datasets, the result is a denoised, spatially smoothed gene expression matrix in each sample's data that better reflects biological structure and spatial organization [5]. This method was conducted with the `spARC` package available on GitHub (https://github.com/KrishnaswamyLab/spARC).

There are four steps to representing a manifold from our processed datasets using spARC: creating a spot-spot similarity graph, defining the diffusion operator, powering the diffusion operator, and calculating diffusion wavelets [13].

2.4.1 Creating a Spot-Spot Similarity Graph

A spot-spot similarity graph based on gene expression profiles ($G_{\text{expression}}$) was constructed by first computing pairwise distances between cells and connecting each cell to its k nearest neighbors using a k-nearest neighbors (kNN) graph. A shared nearest neighbor (SNN) graph representing spatial similarity (G_{spatial}) was then formed by weighting edges based on the neighborhood overlap between each cell and its neighbors, determined by Jaccard Similarity.

2.4.2 Defining the Diffusion Operator

Since we were dealing with ST data, two diffusion operators were defined: an expression diffusion operator ($P_{\text{expression}}$) derived from $G_{\text{expression}}$ and a spatial diffusion operator (P_{spatial}) derived from G_{spatial}. To define each diffusion operator, the affinity matrix (A) was multiplied by the inverse of the degree matrix (D) [14]. The affinity matrix is a square $n_{\text{spots}} \times n_{\text{spots}}$ where each entry A_{ij} denotes how similar two spots i and j are in the gene expression space or the physical tissue. This matrix determines which spots influence each other in the data diffusion process and by how much. The degree matrix represents how connected each spot is to its neighborhood, and it row-normalizes the diffusion operator so that the transition probabilities sum to 1 for each node. This way, spots with stronger neighbors do not dominate the diffusion. The resulting diffusion operator (P) is an $n_{\text{spots}} \times n_{\text{spots}}$ Markov transition matrix, where each entry P_{xy} is the probability of moving from spot x to spot y and sharing information [14].

2.4.3 Powering the Diffusion Operator

Then, both $P_{\text{expression}}$ and P_{spatial} were raised to the power of some user-defined t_1 and t_2, respectively, to simulate a t-step random walk through the data, which reduces noise and uncovers the manifold structure [14].

These steps were performed by first initializing a `spARC` object that used all available CPU cores for parallel computation (`n_jobs` $= -1$) and setting a random state of 42. The `fit_transform` method then constructs two graphs: one based on transcriptional similarity (expression_X) and one based on physical proximity (spatial_X). The data diffusion–based filtering described above was subsequently applied to both graphs.

The resulting powered diffusion operators were multiplied to obtain the integrated diffusion operator, $P_{\text{integrated}}$, which combines information from both sources [5].

$$P_{\text{integrated}} = P_{\text{expression}}^{t_1} \cdot P_{\text{spatial}}^{t_2}$$

2.4.4 Calculating Diffusion Wavelets

To capture both the local and global manifold structure, diffusion wavelets were computed by taking the difference between successive dyadic powers of P [15].

$$\Psi_j = P^{2^j} - P^{2^{j-1}}, \quad 1 \leq j \leq J.$$

J is a user-defined hyperparameter. When j is small, the wavelets capture the structure of local microenvironments within the tissue, whereas for larger values of j, they capture broader spatial domains. The collection of results across all j provides a rich representation of the spot–spot graph, enabling analysis of spatial structure across both local and global scales for all samples.

2.5 Gene Signal Pattern Analysis

We then applied Gene Signal Pattern Analysis (GSPA) to compute each sample's gene embeddings, numerical vectors capturing semantic similarity between genes, informed by both expression and spatial context using the refined gene expression matrix together with the previously calculated diffusion wavelets. GSPA models gene expression measurements as signals defined over the nodes of a spot–spot graph [6]. In this framework, gene–gene distances are calculated based on the structure of the spot–spot graph, so that genes expressed in nearby or connected regions are considered close, whereas genes expressed in distant regions are considered far apart.

All computational methods for GSPA were performed using the gspa library (version 1.1.1) available in Google Colab. An instance of the GSPA class was initialized with $P_{\text{integrated}}$, taken from spARC, as the diffusion operator attribute. After obtaining the diffusion wavelets at scale j, these wavelets were combined into a large wavelet dictionary W of size $n \times Jn$ by calling the build_wavelet_dictionary() method on the GSPA instance.

Highly Variable Genes (HVGs), defined as genes whose expression levels vary substantially across spots, were then selected using the scprep library. We chose the 70th percentile as the threshold for HVGs, as this produced the best results after testing the 90th, 80th, 75th, and 70th percentiles (reference section 3.4 for details).

The gene matrix X, containing the gene signals of the HVGs, was multiplied by W to express each gene's expression profile in terms of the graph wavelet basis functions derived from $P_{\text{integrated}}$. This transformation projects raw gene expression data into a space that reflects the underlying graph structure, such that similarities and differences between genes are evaluated with respect to spatial relationships in the tissue, rather than expression alone. The resulting matrix XW is high-dimensional, so to improve computational efficiency in downstream analyses, it was passed through an autoencoder that learned a lower-dimensional representation. Autoencoders are neural networks consisting of an encoder, which produces a compressed latent representation, and a decoder, which reconstructs the original data from this lower-dimensional space [16].

The final gene embeddings provide insight into gene co-expression within the tissue, as well as which genes exhibit spatial localization and are therefore more informative for explaining cellular heterogeneity. Projecting the gene matrix onto the wavelet dictionary and learning the gene embeddings was implemented by calling the get_gene_embeddings() method on the GSPA instance.

These steps were performed for each sample dataset to get all patients' gene embeddings. Since the different embeddings included patient-specific HVGs, we first extracted the subset of each embedding corresponding to genes common across all embeddings prior to performing any analysis. We then used the gene embeddings to identify which genes are most informative about cell–cell variation. This was achieved by calculating the localization score of each gene, a technique that quantifies the extent to which a gene's expression signal is concentrated within specific regions of the graph topology [6].

2.6 Analyzing Highly Localized Genes

The top 25% most localized genes from each dataset were computed and grouped into responder- and non-responder-specific sets. To determine which genes were uniquely localized to one group, we separated genes present only in responders from those present only in non-responders. We then performed gene set enrichment analysis using the Enrichr tool, accessed through the enrichr method in the gseapy Python package. Enrichr analyzes a list of genes by comparing it to various curated gene sets to determine the biological processes, pathways, or molecular functions that are statistically overrepresented in that list [17, 18, 19]. For this analysis, enrichment was focused on the gene sets in the Gene Ontology Biological Processes 2025, Gene Ontology Cellular Component 2025, and MSigDB Hallmark 2020 databases, and only enrichment pathways with an adjusted p-value of ≤ 0.05 were retained. Adjusted p-value measures the probability of enrichment by chance while considering the number of gene sets tested and accounting for the increased chance of false positives, making it more reliable for selecting truly significant pathways.

2.7 Comparing Gene-Gene Networks

To compare differences in gene co-localization between responder and non-responder tissues, we converted the gene embeddings into graphs in order to capture the gene–gene interactions present in each sample. For each dataset, a k-nearest neighbors (KNN) graph with $k = 5$ was constructed using the embeddings, formatted as a pandas.DataFrame, to the kneighbors_graph method from the sklearn.neighbors module. In this graph, each gene is represented as a node connected to its k most similar genes, where similarity was determined by the Euclidean distance between their embedding values.

Once the graphs were created, we used four methods to evaluate the differences between the neighborhood structures of the responder and non-responder gene networks.

1. Edge Jaccard Similarity

2. Training a Binary Graph Classifier
3. Uniform Manifold Approximation and Projection
4. Response Differential Score

2.7.1 Edge Jaccard Similarity

Edge Jaccard similarity takes the ratio of the shared edges between two graphs to the total number of distinct edges across them [20]. Given two graphs G_1 and G_2, with edge sets E_1 and E_2, the edge Jaccard similarity can be represented by the following equation.

$$J(G_1, G_2) = \frac{|E_1 \cap E_2|}{|E_1 \cup E_2|}$$

Thus, a Jaccard similarity index of 0 indicates that the graphs have no edges in common, while an index of 1 indicates identical graphs. We computed the pairwise Jaccard similarity for all seven embeddings and displayed the results in a 7x7 heatmap.

2.7.2 Constructing Ego Subgraph Embeddings

In order to compare the gene-gene networks on a local scale in addition to looking at overall similarity, we analyzed subgraphs from different samples. For each of the seven gene networks, ego subgraphs containing all nodes within 2 hops of the center node were created for all genes shared across the different networks.

We then applied the `Graph2Vec` algorithm, implemented in the `karateclub` library, to compute numerical embeddings for all ego subgraphs.

2.7.3 Uniform Manifold Approximation and Projection

We utilized Uniform Manifold Approximation and Projection (UMAP) to visually analyze and compare the ego subgraphs of common genes across different samples. UMAP is a dimensionality reduction technique that embeds high-dimensional data into a lower-dimensional space (in this case, two dimensions) while preserving both local and global structural relationships [21]. In this embedding, genes with similar biological functions or spatial expression patterns will tend to cluster together, whereas unrelated genes are positioned farther apart, enabling identification of clusters and relationships within the networks.

The Graph2Vec embeddings were projected into two dimensions using the `UMAP fit_transform` method to represent all subgraphs. The resulting plot visualized the graph embeddings, with points color-coded by sample.

2.7.4 Training a Binary Graph Classifier

An approach to identify gene–gene interactions that most strongly distinguish responder from nonresponder tissue signaling patterns is to train a classifier to determine whether an ego subgraph embedding originates from a responder or nonresponder network. Subgraphs with the highest classification confidence were deemed the most discriminative between the two states.

The analysis pipeline proceeds as follows: for each node, an ego graph is constructed; each subgraph is embedded using Graph2Vec; a classifier is trained to predict whether the subgraph is derived from a responder or nonresponder network; and lastly, the most confidently classified subgraphs and their center nodes are examined.

Tests were conducted using a Random Forest Classifier (RFC). RFCs were selected for their robustness to noise, ability to model complex decision boundaries, and strong performance as a baseline method with minimal parameter tuning [22]. The model performance was assessed using the mean accuracy obtained from 5-fold cross-validation. These analyses were conducted utilizing the `RandomForestClassifier` class from the `sklearn.ensemble` module, in conjunction with the `sklearn.model_selection`.

The center genes of the most discriminative subgraphs were then input into Enrichr to determine the pathways they are most associated with. A drawback of using the classifier-based approach to identify response-specific pathways is that genes with the highest classification confidence may belong to modules that are unique to individual samples rather than reflective of broader response patterns. To address this, we developed an alternative strategy based on mean Euclidean distance, which highlights subgraphs that exhibit greater similarity within responder and non-responder groups while remaining more distinct between the groups.

2.7.5 Response Differential Score

The mean Euclidean distance is the average of all Euclidean distances computed between pairs of vectors in a dataset. The Euclidean distance of two vectors is the straight-line distance connecting them in an n-dimensional space. Given two vectors x and y, the Euclidean distance d can be represented with the following equation[23].

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},$$

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

When d = 0, the vectors are identical, and as d increases, the vectors become more different. To determine which gene interactions were the most consistent within responders and non-responders, but different between those groups, we computed the pairwise Euclidean distances of each individual gene's seven subgraph vector embeddings from the different sample networks.

Let

$$E_g^r = \{\mathbf{e}_1, \ldots, \mathbf{e}_m\} \quad \text{and} \quad E_g^{\mathrm{nr}} = \{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$$

be the embeddings of subgraphs centered at gene g for responder and non-responder patients. We computed the mean pairwise Euclidean distance for values within E_g^r, within E_g^{nr}, and between E_g^r and E_g^{nr} to get (S_r), (S_{nr}), and ($S_{r,\mathrm{nr}}$), respectively, using the `numpy.linalg.norm` method.

Using these three values, we defined a novel metric called the *response differential score*, which quantifies how distinct or similar the responder and non-responder groups are in terms of that gene's vector embeddings. The differential score is calculated as:

$$\text{Differential Score}_g = \frac{1}{2}\left(S_r + S_{\mathrm{nr}}\right) - S_{r,\mathrm{nr}}$$

A negative differential score indicates that the gene's subgraph is more consistent within each group than it is between the groups. As a result, we selected the genes with a negative differential score for further analysis. Since these genes appear to be involved in interactions indicative of sample response, they have the potential to reveal major pathways and cellular functions that play a role in whether a sample responds to the HCC neoadjuvant therapy.

3 Results

3.1 Leiden Clustering

Figure 2 depicts the spatial domains identified for each tissue sample using Leiden clustering. Each color represents a distinct spatial domain, with domains differing in both size and structure. For instance, in sample HCC1R, the purple cluster appears as a small, compact region located in the lower right portion of the tissue, whereas the red cluster is more diffuse, extending across the central region. These domains outline regions within the tissue that may serve distinct biological roles, contributing to the function of the tissue as a whole.

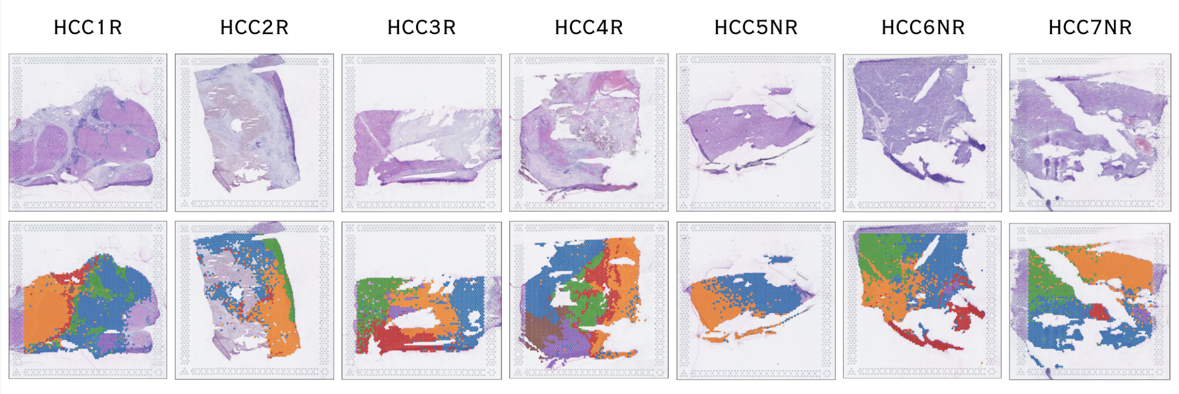

Figure 2: Each column contains two pictures from each sample: an H&E image of the tissue (Row 1) and an image showing the results of Leiden clustering (Row 2). The regions with different colors in the Leiden clustering figures represent different clusters.

3.2 Analyzing Top Enriched Processes

Of the top 25% most localized genes identified across all samples, 7 were uniquely localized in responder tissues, 50 were uniquely localized in non-responder tissues, and none were shared across all samples. When these gene sets were examined using Enrichr, distinct biological pathways were enriched that may reflect fundamental differences in the TMEs of responders and non-responders. Pathways with the smallest adjusted p-value in each gene set were chosen for further analysis. The following figures and tables present the pathway enrichment results for the genes uniquely localized in responder (Figure 3 and Table 1) and non-responder (Figure 4 and Table 2) samples.

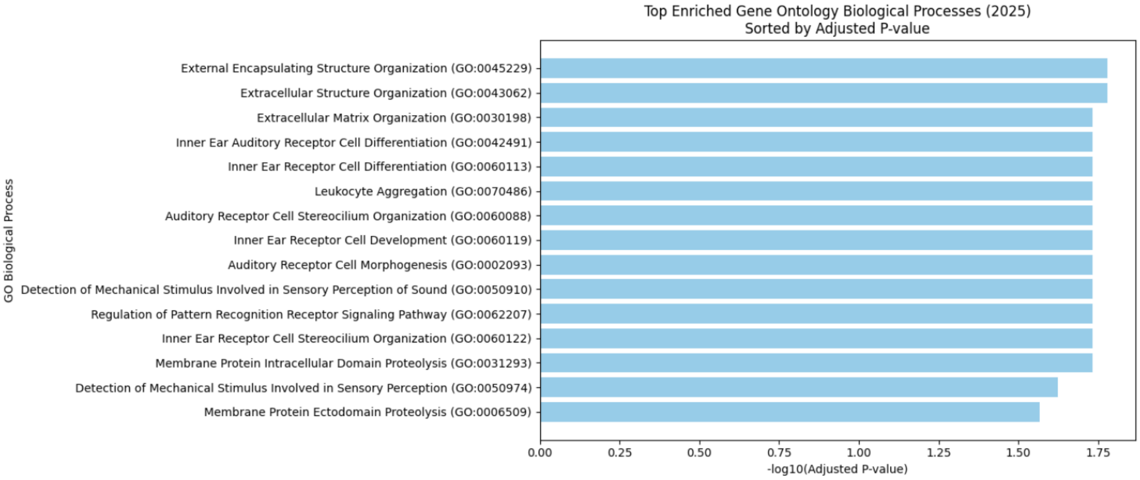

Figure 3: Each bar graph displays the top enriched processes (in a specific gene set) among the genes uniquely localized in responders. Gene sets with significant pathways were Gene Ontology Biological Processes 2025 (shown graph) and MSigDB Hallmark 2020 (graph not included because it contained only 1 pathway). The Gene Ontology Cellular Component 2025 set did not have any significantly enriched pathways. The processes on the y-axis are sorted based on -log10(Adjusted p-value). Processes with the highest -log10(Adjusted p-value) were chosen for downstream analysis.

Gene Set Name	**Pathway Name**	**Adjusted p-value**	**Genes Enriched**
GO Biological Process 2025	External Encapsulating Structure Organization	1.67×10^{-2}	MMP7; COL10A1
GO Biological Process 2025	Extracellular Structure Organization	1.67×10^{-2}	MMP7; COL10A1
GO Biological Process 2025	Extracellular Matrix Organization	1.85×10^{-2}	MMP7; COL10A1
MSigDB Hallmark 2020	Coagulation	4.73×10^{-2}	MMP7

Table 1: Top enriched pathways for genes uniquely localized in responder samples across gene set collections. Genes Enriched is the subset of input genes driving the enrichment of a particular pathway.

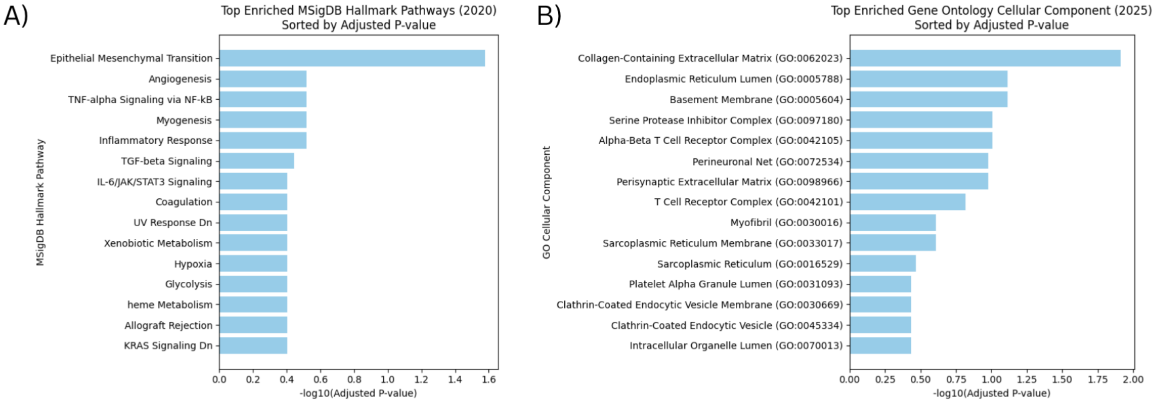

Figure 4: Each bar graph displays the top enriched processes (in a specific gene set) among the genes uniquely localized in non-responders. Gene sets with significant pathways were MSigDB Hallmark 2020 (A) and Gene Ontology Cellular Component 2025 (B). The Gene Ontology Biological Processes 2025 set did not have any significantly enriched pathways. The processes on the y-axis are sorted based on -log10(Adjusted p-value). Processes with the highest -log10(Adjusted p-value) were chosen for downstream analysis.

Gene Set Name	**Pathway Name**	**Adjusted p-value**	**Genes Enriched**
MSigDB Hallmark 2020	Epithelial Mesenchymal Transition	2.65×10^{-2}	VCAN; LAMA2; SERPINE1; GEM
GO Cellular Component 2025	Collagen-Containing Extracellular Matrix	1.22×10^{-2}	VCAN; LAMA2; SERPINE1; COL8A1; SULF1; ASPN

Table 2: Top enriched pathways for genes uniquely localized in non-responder samples across gene set collections. Genes Enriched is the subset of input genes driving the enrichment of a particular pathway.

Responder-specific genes were enriched in pathways linked to tissue remodeling and structural organization. Enrichment of the coagulation pathway (adj. p-value = 0.04732) reflects the well-established pro-coagulant state of HCC tumors, though in responders, this may indicate a TME that supports effective immune response despite platelet shielding of tumor cells [24]. Pathways associated with external encapsulating structure organization (adj. p-value = 0.01668) highlight the role of organized tissue barriers, which in HCC may contribute to the formation of an enclosure that restricts tumor spread [25]. Furthermore, enrichment of extracellular structure organization (adj. p-value =

0.01668) and ECM organization (adj. p-value = 0.01852) suggests active ECM remodeling in responder tissues, potentially reducing matrix stiffness and allowing immune infiltration and drug accessibility [26].

Additionally, several pathways associated with inner ear hair cell function, including Inner Ear Auditory Receptor Cell Differentiation (adjusted p-value = 0.01852) and Inner Ear Receptor Cell Differentiation (adjusted p-value = 0.01852) pathways, were enriched among the responder genes. Inner ear hair cells are specialized for mechanosensing, and many mechanosensory pathways are also active in HCC tissue samples [27, 28]. This convergence of mechanosensory signaling suggests that similar molecular mechanisms may underlie mechanosensing in both contexts, highlighting a new connection that could be explored in future studies.

Among the non-responder localized genes, enriched pathways included epithelial-to-mesenchymal transition (EMT) (adj. p-value = 0.02652), a process often upregulated in non-responders because it enables epithelial tumor cells to acquire mesenchymal traits that promote invasiveness and drug resistance [29]. In addition, enrichment of the collagen-containing ECM pathway (adj. p-value = 0.01221) suggests that excessive ECM deposition may act as a physical barrier, blocking immune cell infiltration and drug penetration [26]. These findings suggest a TME conducive to cancer progression and therapy resistance.

Altogether, these results indicate that responder and non-responder tumors are characterized by fundamentally different microenvironmental states: responders show evidence of structural remodeling that may enhance therapeutic efficacy, while non-responders exhibit features of EMT and ECM-driven resistance.

3.3 Jaccard Similarity

The heatmap below depicts the Jaccard similarity between the KNN networks of the seven sample datasets. As stated in the methods, the Jaccard index quantifies the similarity between two networks based on the proportion of shared edges, with values approaching 0 indicating minimal overlap and values approaching 1 indicating near-identical network structure. The results reveal no clear similarities within either the responder or non-responder groups, nor a significant difference between groups relative to within-group comparisons. This finding was unexpected, and we attribute it to the Jaccard index's focus on capturing global differences across entire networks. Although certain gene modules may be consistent within the responder and non-responder groups and distinct between them, these group-specific patterns are likely masked by the presence of other gene modules that are not associated with treatment response, leading to a diluted overall network comparison.

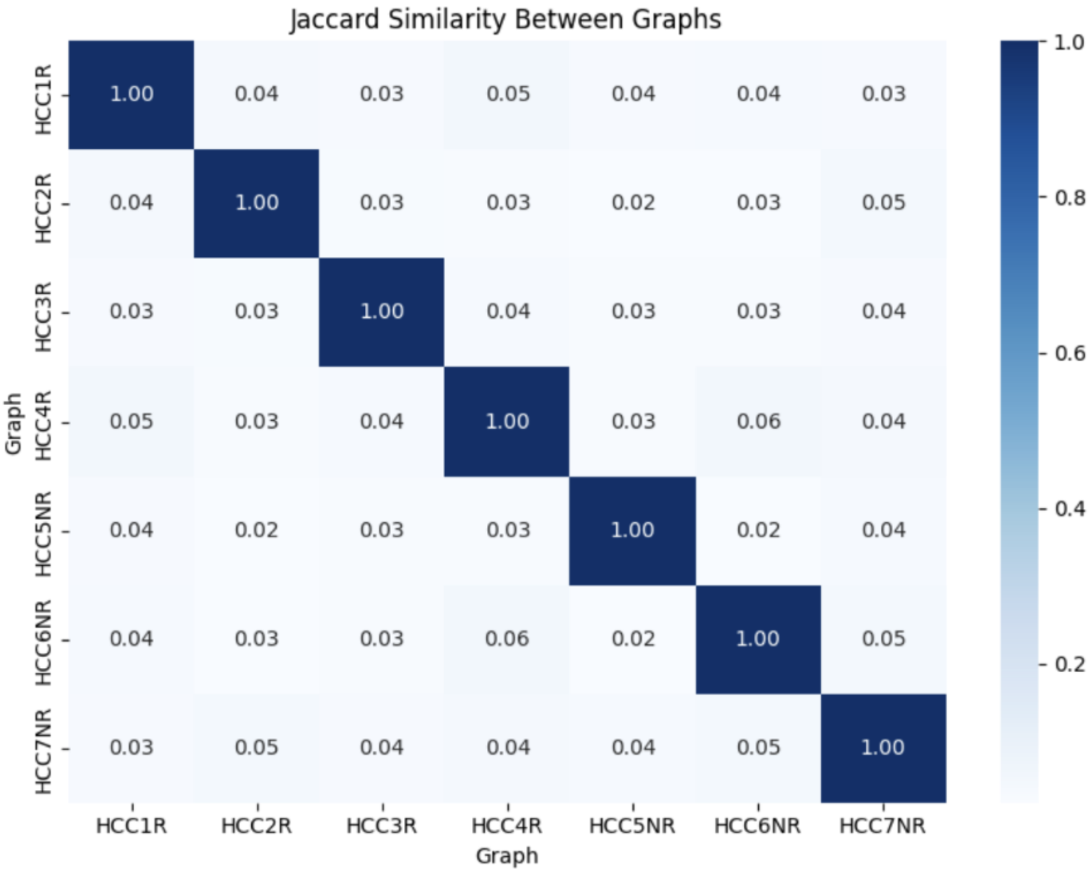

Figure 5: A 7x7 heatmap showing the results of Jaccard Similarity conducted between the gene-gene networks from each sample. Darker blue squares represent higher Jaccard similarity between the respective graphs.

3.4 UMAP

To compare individual subgraphs between responder and non-responder networks, we applied UMAP to embed the high-dimensional data into two dimensions while maintaining both local and global structural relationships. Figures 6.A and 6.C show points colored by the seven individual sample networks, while Figures 6.B and 6.D distinguish responders from non-responders. Our initial analysis only contained the top 10% of HVGs, which included 111 genes (Figures 6.A and 6.B). Surprisingly, every gene exhibited a similar neighborhood structure across all samples, indicating insufficient resolution.

To address this, we expanded our analysis to include the top 20%, 25%, and 30% of HVGs, while avoiding larger percentages to ensure that only genes with variable expression across the tissue were considered. These trials included 276, 342, and 360 genes. As more genes were incorporated, the overlap between responder and non-responder embeddings decreased, suggesting that the additional genes provided more informative differentiation. Based on these observations, as well as the results from the RFC, all downstream analyses were conducted using the results from trials that had the 70th percentile as the threshold for HVGs.

In the UMAP embedding for the 70th percentile HVGs (Figures 6.C and 6.D), sample embeddings display clear overlap within their respective responder and non-responder groups. For example, HCC3R and HCC4R overlap with each other among responders, while HCC6NR and HCC7NR overlap among non-responders, indicating that certain gene modules are likely response-specific. Outlier embeddings, such as those from HCC2R, HCC5NR, and HCC7NR, likely reflect sample-specific gene modules rather than response-related patterns.

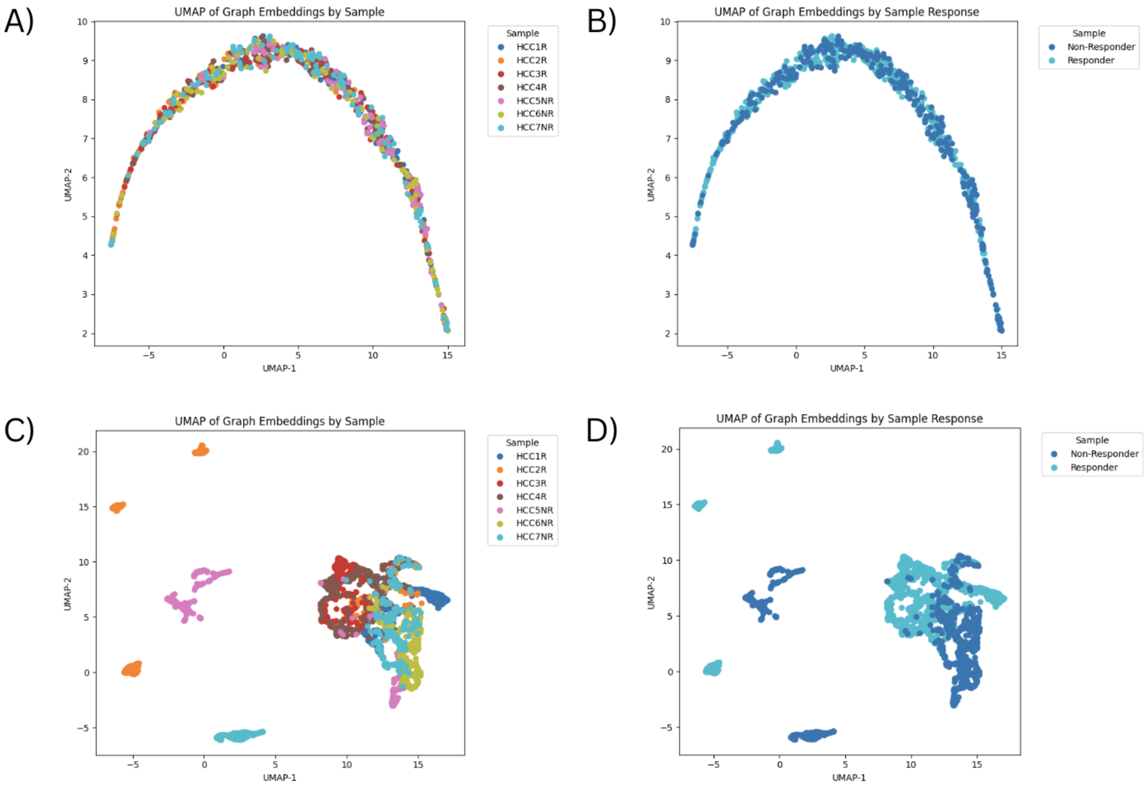

Figure 6: UMAP plot of ego subgraph embeddings from each sample. A) UMAP embedding for the 90th percentile HVGs. Points are colored differently based on which sample the gene embedding is from. B) UMAP embedding for the 90th percentile HVGs. Points are colored differently based on whether the gene embedding is from a responder or non-responder sample. C) UMAP embedding for the 70th percentile HVGs. Points are colored differently based on which sample the gene embedding is from. D) UMAP embedding for the 70th percentile HVGs. Points are colored differently based on whether the gene embedding is from a responder or non-responder sample.

3.5 Training an RFC

This approach involved training a Random Forest Classifier to predict whether an ego subgraph embedding originated from a responder or non-responder network. In this way, we could determine whether recognizable structural patterns distinguish responder networks from non-responder networks. Furthermore, genes at the center of subgraphs with the highest classification confidence were selected for pathway enrichment analysis, since these pathways may be associated with treatment response.

As with the UMAP analysis, the RFC was initially trained on subgraph embeddings derived from the top 10% of HVGs. However, the classifier had a low performance, with an average accuracy of approximately 62% after five-fold cross-validation. This reduced accuracy likely reflects the same limitation observed in the UMAP trial: that the 111 genes in the top 10% subset provided insufficient information to reliably distinguish the neighborhood structures of responder and non-responder networks.

To address this, we conducted additional experiments using the top 20%, 25%, and 30% of HVGs. Classifier performance improved with increasing gene set size, reaching its highest accuracy of approximately 95% when trained on the top 30% of HVGs. These findings demonstrate that ego subgraph embeddings contain sufficient information to accurately classify responder and non-responder samples, particularly when a larger set of variable genes is included.

Several pathways were enriched among the genes corresponding to ego subgraphs with the highest classification confidence, including the Regulation of Androgen Receptor Signaling Pathway and the Regulation of Protein Localization. However, as mentioned in the Methods, because the classifier was designed to distinguish whether a given gene subgraph originates from a responder or non-responder patient, it does not explicitly assess whether the subgraph structure is preserved and consistent across multiple patients within each group. As a result, these pathways may be different in one specific sample instead of across all responders or non-responders.

3.6 Response Differential Score

There were 119 genes with a negative response differential score, indicating that their gene–gene interactions were more similar within responder and non-responder groups than across groups. These genes were highly enriched in pathways related to cell adhesion, immune regulation, and cellular plasticity. These findings suggest that the processes that determine how cells interact with their surrounding microenvironment and immune system are highly response-specific in the outcomes of HCC therapy. Figure 7 and Table 3 display the pathway enrichment results in more detail.

Notably, enrichment of the focal adhesion (adj. p-value = 0.000004879) and cell-substrate junction (adj. p-value = 0.000004879) pathways highlights the importance of cell-ECM interactions in shaping distinct TMEs between responders and non-responders. Focal adhesions activate both attachment and signaling between tumor cells and the ECM, which promotes their proliferation, migration, and resistance to therapy [30]. Similarly, cell-substrate junctions improve tumor structure and regulate signaling cascades that control the effectiveness of immune cell infiltration into tumors [31].

The enrichment of the MHC protein complex pathway (adj. p-value = 0.000004879) emphasizes the importance of immune surveillance in HCC therapy response. MHC molecules play a vital role in antigen presentation, which is crucial to triggering adaptive immune responses. Differences in MHC-related interactions between responder and non-responder samples may determine whether tumor cells are able to evade immune detection, potentially influencing therapeutic outcomes [32].

In addition, pathways related to the regulation of cell migration (adj. p-value = 7.221e-7) and EMT (adj. p-value = 1.043e-7), both of which are central to tumor progression and metastasis, were also strongly enriched. Specifically, EMT allows tumor cells to invade surrounding tissue, resist apoptosis, and develop drug resistance, and increased tumor cell migration can contribute to HCC progression and therapy resistance [29].

These results suggest that responder and non-responder tissues have distinct biological mechanisms, particularly in ECM interactions, immune regulation, and cell migration. Such processes may significantly influence whether HCC tumors respond to treatment.

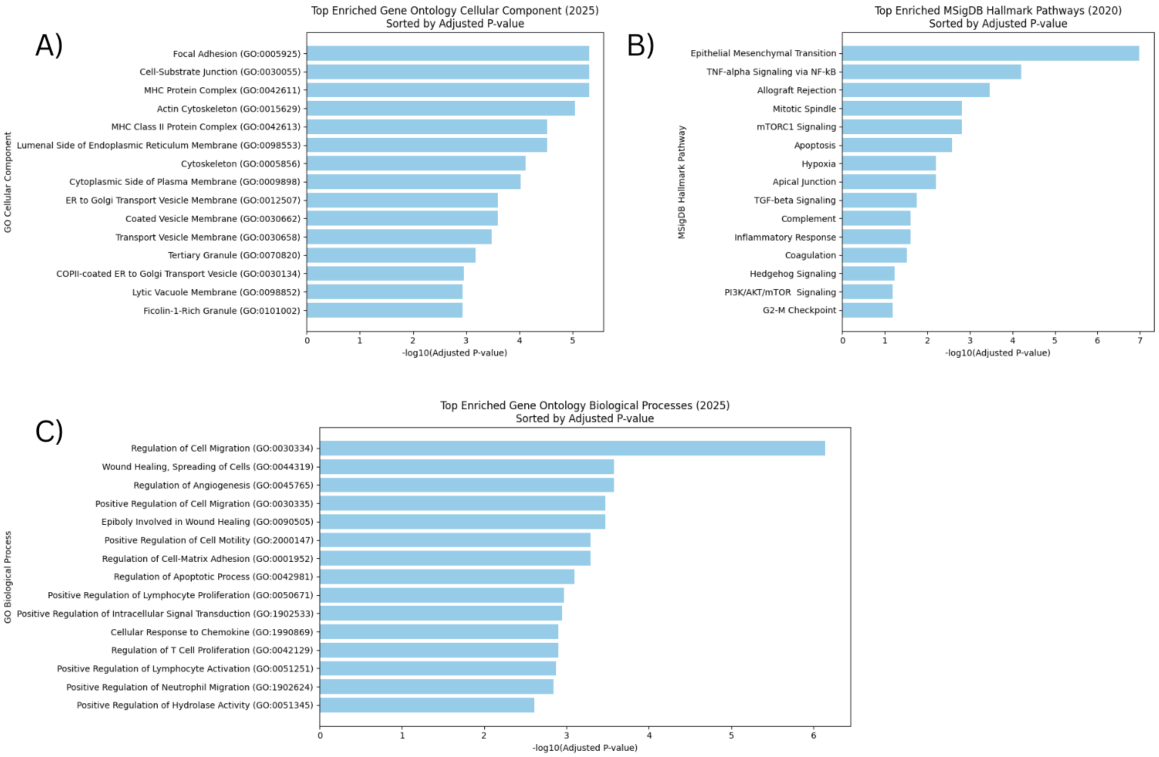

Figure 7: Each bar graph displays the top enriched processes (in a specific gene set) among the genes with a negative response differential score. Gene sets analyzed were Gene Ontology Cellular Component 2025 (A), MSigDB Hallmark 2020 (B), and Gene Ontology Biological Processes 2025 (C). The processes on the y-axis are sorted based on -log10(Adjusted p-value). Processes with the highest -log10(Adjusted p-value) were chosen for downstream analysis.

Gene Set Name	**Pathway Name**	**Adjusted p-value**	**Genes Enriched**
GO Cellular Component 2025	Focal Adhesion	4.88×10^{-6}	HSPA9; ITGB1; CD151; DST; AHNAK; ARPC1B; NEXN; YWHAZ; RHOA; MARCKS; MYH9; FLNA; CD59; ATP6V0C
GO Cellular Component 2025	Cell–Substrate Junction	4.88×10^{-6}	HSPA9; ITGB1; CD151; DST; AHNAK; ARPC1B; NEXN; YWHAZ; RHOA; MARCKS; MYH9; FLNA; CD59; ATP6V0C
GO Cellular Component 2025	MHC Protein Complex	4.88×10^{-6}	CD74; HLA-DRA; HLA-DOA; HLA-E; HLA-DPA1
GO Biological Process 2025	Regulation of Cell Migration	7.22×10^{-7}	ITGB1; CD74; GRN; CD151; ROCK1; LIMCH1; NEXN; RHOC; THBS1; RTN4; RHOA; CXCL12; GPNMB; CCL5; NF1; FLNA; CCL19; CCL18
MSigDB Hallmark 2020	Epithelial–Mesenchymal Transition	1.04×10^{-7}	ITGB1; CXCL12; DST; IGFBP4; CALD1; MGP; FLNA; CD59; EMP3; SAT1; THBS1; PDLIM4

Table 3: Top enriched pathways for genes with a negative response differential score across gene set collections. Genes Enriched is the subset of input genes driving the enrichment of a particular pathway.

4 Discussion

Fully understanding the processes underlying cancerous tissue is a difficult task due to the complexity and heterogeneity of the TME [3]. Traditional ST methods often treat gene measurements as isolated signals, leading to results heavily influenced by noise. Moreover, these approaches typically focus on identifying differentially expressed genes, providing limited insight into the coordinated activity of genes and the higher-order structures that drive cellular behavior. As a result, critical mechanisms that depend on gene-gene interactions may remain undetected. This highlights the need for computational methods that can combine both expression and spatial context to construct spatially informed gene networks, enabling a more comprehensive understanding of how genes function together within the TME and potentially uncovering novel determinants of therapeutic response.

Our study builds on the research conducted by Zhang et al [4], which showed that response to neoadjuvant cabozantinib and nivolumab therapy in HCC is associated with distinct immune activity, ECM remodeling, and EMT. In alignment with their work, we observed enrichment of pathways related to tumor cell-ECM interactions and EMT in discriminative genes between responder and non-responder samples. However, by leveraging spARC [5], GSPA [6], and our new response differential score, we extend these findings to identify additional response-specific pathways that reflect biological mechanisms involved in therapy resistance and tumor progression, including those involved in antigen presentation through the MHC complex, coagulation, and external encapsulating structure remodeling. We also detected pathways that showed strong associations but had less well-characterized links to HCC therapy response, suggesting that they may contribute to tissue response in ways that are not yet discovered. These findings demonstrate the ability of graph-based computational ST approaches to reveal differences in spatial gene-gene interaction patterns that go beyond traditional expression-level analyses.

5 Future Work

There are several directions for future work that could enhance and expand our analyses. Alternative normalization techniques specifically tailored to ST data should be explored, since conventional scRNA-seq normalization may disregard biologically meaningful variation in spot-level transcript counts. Additionally, further analyses of gene expression patterns within Leiden clusters could reveal spatially coherent modules that may contribute to therapy response. By examining the gene expression profiles within these domains, we can infer their potential functional niches and gain deeper insight into the differences between the molecular and spatial organization of responder and non-responder tissues. Furthermore, systematic testing of our method parameters, such as the number of neighbors in the kNN graph, may provide additional results.

Beyond these methodological improvements, it would be beneficial to expand the scope of analysis. Applying the workflow to larger and more balanced datasets of HCC responders and non-responders will help determine whether the response-associated pathways we identified are consistent across a more diverse patient set. Additionally, extending this framework to ST datasets from other diseases or treatment settings could provide insight into whether the methods generalize to different biological contexts and reveal shared or disease-specific spatial processes. Finally, the pathways enriched in our analysis should be further investigated by experimental biologists to determine whether their activity is truly altered between responder and non-responder tissues.

These future directions highlight how computational modeling and experimental validation can be combined to build a more complete understanding of treatment response in HCC and guide therapeutic strategies informed by the spatial organization of tumor microenvironments.

References

[1] Zahedi R et al. Deep learning in spatially resolved transcriptomics: a comprehensive technical view. *Briefings in Bioinformatics*, 2024.

[2] Williams C et al. An introduction to spatial transcriptomics for biomedical research. *Genome Medicine*, 2022.

[3] Ochoa Garrido I Ayuso JM. The importance of the tumor microenvironment to understand tumor origin, evolution, and treatment response. *Cancers*, 2022.

[4] Zhang S et al. Spatial transcriptomics analysis of neoadjuvant cabozantinib and nivolumab in advanced hepatocellular carcinoma identifies independent mechanisms of resistance and recurrence. *Genome Medicine*, 2023.

[5] Kuchroo M et al. sparc recovers human glioma spatial signaling networks with graph filtering. *bioRxiv*, 2022.

[6] Venkat A et al. Mapping the gene space at single-cell resolution with gene signal pattern analysis. *Nature Computational Science*, 2024.

[7] Gse238264 dataset, 2024.

[8] Cai JJ Osorio D. Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell rna-sequencing data quality control. *Bioinformatics*, 2020.

[9] *Scanpy: calculate_qc_metrics*, 2023.

[10] Single-cell rna-seq data normalization, 2023.

[11] Merchan. Python for visium spatial transcriptomics, 2022.

[12] Kavlakoglu E. Reducing dimensionality with principal component analysis with python, 2024.

[13] Venkat A et al. Multiscale geometric and topological analyses for characterizing and predicting immune responses from single cell data. *Trends in Immunology*, 2023.

[14] Dijk D et al. Recovering gene interactions from single-cell data using data diffusion. *Cell*, 2019.

[15] Maggioni M Coifman R. Diffusion wavelets. *Science Direct*, 2006.

[16] Koumakis L. Deep learning models in genomics; are we there yet? *Science Direct*, 2020.

[17] Chen Y et al. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC Bioinformatics*, 2013.

[18] Kuleshov M et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 2016.

[19] Xie Z et al. Gene set knowledge discovery with enrichr. *Current Protocols*, 2021.

[20] Karabiber F. Jaccard similarity.

[21] McInnes L et al. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv*, 2018.

[22] Donges N. Random forest: A complete guide for machine learning, 2023.

[23] Euclidean distance, 2025.

[24] Monteiro RQ Lima LG. Activation of blood coagulation in cancer: implications for tumour progression. *Bioscience Reports*, 2013.

[25] Go:0045230: Cell-substrate junction, 2023.

[26] Jiang Y He X, Lee B. Extracellular matrix in cancer progression and therapy. *Medical Review*, 2022.

[27] Zahler S Passi M. Mechano-signaling aspects of hepatocellular carcinoma. *J. Cancer*, 2021.

[28] Holt J Zheng W. The mechanosensory transduction machinery in inner ear hair cells. *Annu. Rev. Biophys.*, 2021.

[29] Weinberg RA Kalluri R. The basics of epithelial-mesenchymal transition. *The Journal of Clinical Investigation*, 2009.

[30] Alahari SK Maziveyi M. Cell matrix adhesions in cancer: The proteins that form the glue. *Oncotarget*, 2017.

[31] Harjunp H et al. Cell adhesion molecules and their roles and regulation in the immune and tumor microenvironment. *Frontiers in Immunology*, 2019.

[32] Rock K Dhatchinamoorthy K, Colbert J. Cancer immune evasion through loss of mhc class i antigen presentation. *Frontiers in Immunology*, 2021.