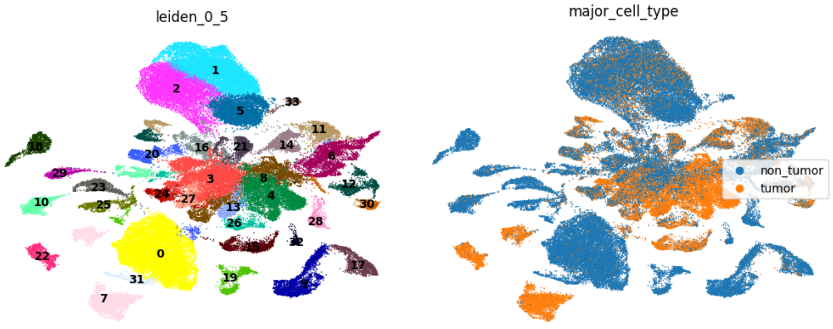
Mid Progress Report

Our project asks a focused question: does adding spatial context actually change how we infer cell–cell communication in glioblastoma, or do standard snRNA-seq–based methods already capture the key ligand–receptor interactions? To answer this, we are building a unified pipeline based on the data that we gathered from wang et al. [1]. Since it contains both single-cell RNA and spatial transcription data for cells from the same patient, it provided us with the opportunity to test our hypotheses. The plan is to run multiple communication algorithms on harmonized inputs, compare their outputs quantitatively, and finally test whether spatially informed interaction graphs better predict tumor state than graphs built from dissociated data alone.

The work completed so far corresponds to Steps 1–4 of our pipeline. We have completed the heavy lifting on data processing and quality control. Starting from Step 1, the GSE174554 raw matrices, we systematically parsed every matrix file, together with the matching feature and barcode files. We then filter any structurally broken libraries, like the SF8963 matrix, with malformed dimensions as an example. This leaves us with a curated snRNA-seq atlas covering essentially the intended ~80 GBM. For each valid library we computed global gene-wise counts and selected the top 5,000 most expressed genes as our working feature space. This provides a smaller file so that we can run it faster, and not have issues with the RAM (we could go back and increase the number if we saw that it is not enough later on with the project).

In Step 2, we merged all usable libraries into a single AnnData object with one row per cell and harmonized gene coordinates After basic QC filters (minimum total counts, removal of extreme outliers), we down-sampled to a working set of 80,000 cells to respect RAM constraints and to support repeated benchmarking. Importantly, this down sampling is done after filtering and stratified across samples, so we preserve representation of different tumors and conditions. On that object, we have applied library-size normalization, log-transformation, highly variable gene selection , PCA, KNN, and UMAP embedding.



A. .B

Figure 1, A. UMAP of the processed snRNA-seq (80k cells, 5k genes) colored by Leiden clusters shows clean, well-separated groups, indicating successful QC and capturing the expected heterogeneity of GBM. **B.** The same UMAP colored by tumor vs non-tumor labels shows tumor cells forming coherent regions and

By: Yaseen Arab

non-tumor cells mapping to distinct neighborhoods, supporting a credible sender/receiver setup for communication analysis.

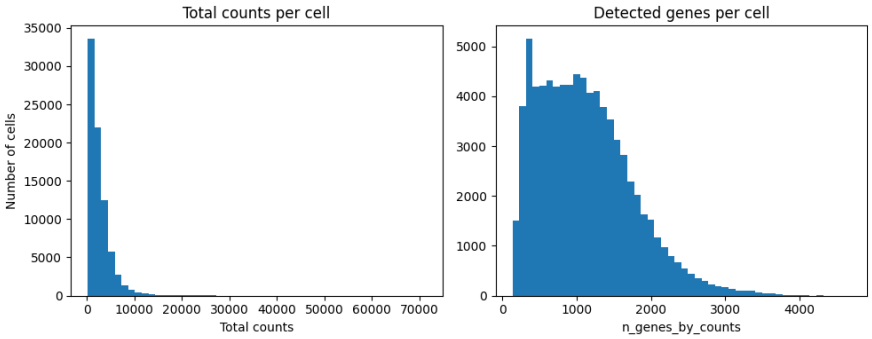


Figure 2, Distributions of total counts and detected genes per cell confirm that low-quality cells were removed while preserving high-quality profiles, justifying our thresholds and the 80k-cell working set.

In Step 3 for the spatial modality, we constructed a matched AnnData object from the processed. We mapped each spot id to the sample if using the meta data that the paper have offered, We selected the top expressed spatial genes, removed extremely low-coverage spots, and stored the spatial object. The result is a smaller spatial panel than the snRNA-seq atlas, but with high-confidence links to specific tumors and local neighborhoods. Each spot summarizes the microenvironment around tumor regions, making this dataset ideal for testing whether spatially constrained ligand–receptor patterns add information beyond dissociated profiles.

In Step 4, we turn our cleaned data into something every method can use in a consistent way. We start from the author's Tumor vs Normal labels and, for each group (tumor and non-tumor) and each gene, we calculate the average expression. Given a list of known ligand–receptor pairs, we then assign a simple score to each possible interaction: how strongly a ligand is expressed in the sender group multiplied by how strongly its receptor is expressed in the receiver group. This gives us an easy-to-interpret communication graph for both the snRNA-seq data and the spatial data, built with the same rule. These graphs are our common input: they are clean enough to feed into the next step tools (CellPhoneDB, CellChat, NicheNet and others), and structured enough to serve as the input for our own graph model ,GNN, that will test whether spatial information actually helps distinguish primary from recurrent tumors.

By: Yaseen Arab

Average ligand expression in sender group:

$$\overline{L}_S = \frac{1}{|S|} \sum_{i \in S} X_{i,L}$$

Average receptor expression in receiver group:

$$\overline{R}_T = \frac{1}{|T|} \sum_{j \in T} X_{j,R}$$

Communication score:

$$score(S \to T, L{-}R) = \overline{L}_S \cdot \overline{R}_T$$

Given: X (snRNA cells × genes), Y (spatial spots × genes), cell types c(i), tumor label t(i) and ligand–receptor pairs L,R). The score simply multiplies average ligand expression in the sender group by average receptor expression in the receiver group: higher values mean stronger, more signaling between those groups, lower values mean weaker support. This gives us a clean, comparable way to turn both snRNA and spatial data into weighted communication graphs for all later analyses.

With preprocessing done and both snRNA-seq and spatial data in standardized AnnData formats, the remaining of our pipeline focuses on communication inference, benchmarking, and modeling, yet to be done.

Step 5 will involve Runing CCC methods. We will apply a common panel of cell–cell communication tools CellPhoneDB, CellChat, NicheNet, CytoTalk, scTensor, iTALK, ICELLNET, SingleCellSignalR, and Scriabin on the same input that was created in step 4. For snRNA-seq, each method will see our 80k cells / 5k genes, shared major cell-type labels, and a consistent ligand–receptor reference. For spatial data, we will only run the tools that support spot-level or spatially constrained analysis using our curated spatial AnnData. This ensures differences between methods reflect modeling choices, not inconsistent preprocessing.

Step 6 will involve Benchmark CCC outputs, For each method and modality (snRNA vs spatial), we will **1.** compare overlap of top-ranked ligand–receptor interactions (Jaccard / rank correlation), **2.** check recovery of GBM-relevant pathways (EGFR, VEGF, TGF-β, immune checkpoints, myeloid–tumor crosstalk), and **3.** summarize network structure: which cell types act as hubs, how tumor–immune edges change between primary and recurrent, and how stable these patterns are across tools. This gives us a quantitative and biological view of when methods agree, when spatial information changes the story, and which readouts are trustworthy.

In the final step, we will build GNN models directly on these communication graphs. For each patient, we will construct two graphs one from snRNA-only interactions, one incorporating spatial constraints where nodes represent major cell types and edges encode aggregated ligand–receptor scores. Each graph is labeled as primary or recurrent, and we will train a GNN with standard binary cross-entropy loss and evaluate using ROC–AUC and F1. The comparison between snRNA-only and spatial GNNs, together with edge-level importance, will directly test our core hypothesis:

By: Yaseen Arab

whether spatially informed communication captures predictive and biologically meaningful structure beyond what dissociated transcriptomes alone can provide.

Taken together, this puts us in a strong position for the second half of the project: the data are cleaned, aligned, and encoded in exactly the way our proposal envisioned, and the remaining steps are straightforward applications of the plan rather than infrastructure problems. By contrasting snRNA-only and spatially informed communication graphs across multiple CCC methods, and then testing their predictive value with a shared GNN framework, we will be able to give a clear, interpretable answer to our core question: not just how glioblastoma cells communicate, but when spatial transcriptomics truly changes that picture and when conventional single-cell data are already enough.

[1] Diaz, Aaron, et al. “A Single Cell Atlas of Human Glioma under Therapy.” *Gene Expression Omnibus*, NCBI, 5 July 2022, www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE174554.

By: Yaseen Arab