🕵 TACTIC: An Explainable Multi-Agent Architecture for Classification & Interpretable Reasoning in Spatial Transcriptomics

Abdel Rahman Alsabbagh [*1,2] **Mahmoud Zahran** [*1] **Ali Balubaid** [2]
Sumeer Ahmad Khan [2,3] **Robert Lehmann** [2] **Xabier Martínez de Morentin** [2] **Vincenzo Lagani** [2,3,4]
Narsis A. Kiani [5] **David Gomez-Cabrero** [2] **Jesper Tegnér** [1,2,6,7]

Abstract

While spatial transcriptomics captures the spatial localization of transcriptional signatures, its analysis demands significant bioinformatics expertise. We investigate whether a collaborative multi-agent system can automate this complex workflow. We present **T**ranscriptomic **A**gents for **C**lassification via **T**houghtful **I**nference and **C**oordination (**TACTIC**), for interpretable cell type annotation in spatial transcriptomics. TACTIC integrates graph autoencoders and large language models in a chain-of-thought architecture featuring specialized agents, a junior and a senior bioinformatician, engaging in structured dialogue to deliver accurate, human-interpretable annotations. Evaluated on MERFISH, MIBI-TOF, and Drosophila Stereo-seq datasets, TACTIC achieves F1-scores of 0.80, 0.94, and 0.46, respectively, without task-specific fine-tuning across diverse platforms. Ablation studies show that agent collaboration enhances interpretability, reinforcing the value of structured reasoning. These results position TACTIC as a generalizable and explainable AI framework for spatial omics, requiring no task-specific fine-tuning.

[*]Equal contribution [1]Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia [2]Biological and Environmental Sciences and Engineering (BESE) Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia [3]SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence [4]Institute of Chemical Biology, Ilia State University, Tbilisi, Georgia [5]Algorithmic Dynamics Lab, Center of Molecular Medicine Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden [6]Unit of Computational Medicine, Center of Molecular Medicine & Department of Medicine, Karolinska Institutet, Stockholm, Sweden [7]Science for Life Laboratory, Tomtebodavägen, Solna, Sweden. Correspondence to: Jesper Tegnér <jesper.tegner@kaust.edu.sa>.

Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

1. Introduction

Spatial transcriptomics (Ståhl et al., 2016) enables the in situ analysis of cellular interactions, offering critical insights into processes such as tissue development and cancer progression. Central to the interpretation of spatial omics data is accurate cell type annotation, which requires the integration of gene expression profiles with spatial localization cues (Shen et al., 2022; Shi et al., 2023; Khan et al., 2025; Lu et al., 2025). However, existing methods often operate as black boxes, offering limited interpretability, an important shortcoming in biomedical contexts where transparency is essential for trust, reproducibility, and clinical validation.

The emergence of large language models (LLMs) presents a promising avenue for addressing this limitation. LLMs have demonstrated an ability to extract latent structure and logic from natural language (Zenil et al., 2023; Didolkar et al., 2024; Yang et al., 2024; Zhang et al., 2025), and their utility has been further expanded through the development of chain of thought (CoT) prompting and agentic frameworks. These systems transform LLMs from passive responders into autonomous agents capable of planning, reasoning, and task execution (Wei et al., 2022; Wu et al., 2025).

To harness these capabilities in spatial omics, we introduce TACTIC, a novel multi-agent framework designed specifically for cell type annotation in spatial transcriptomics. TACTIC employs multiple specialized agents that collaborate to deliver accurate annotations alongside human-readable justifications. This multi-agent approach enhances both the reliability and transparency of the annotation process.

TACTIC integrates Graph Autoencoder (GAEs) to generate spatial embeddings from gene expression data and spatial relationships, which are then processed through a multi-agent Chain-of-Thought (CoT) workflow. In this workflow, agents engage in dialogue, reconcile diverse perspectives, and produce robust, interpretable conclusions. Our contributions are:

- We present TACTIC, a collaborative multi-agent framework that models a principal–assistant bioinformatics

workflow to annotate cell types in spatial transcriptomics data accurately.

- We leverage GAE-generated spatial context embeddings, gene expression levels, and marker genes, using CoT prompting for LLMs reasoning to generate transparent, biologically grounded predictions.
- Through extensive ablation studies, we identify the key factors shaping the reliable annotation of the cell types in the different spatial transcriptomic datasets, by removing or randomizing spatial embeddings, modifying the agent collaboration and delegation protocols, and evaluating alternative model configurations.

2. Methods

2.1. Spatial Transcriptomics Datasets & Preprocessing

We evaluated TACTIC - a multi-agent system on three spatial omics datasets, MERFISH (Moffitt et al., 2018), MIBI-TOF (Hartmann et al., 2021), and Drosophila Stereo-seq (Qiu et al., 2024), spanning diverse organisms and measurement technologies. For each cell, we extracted spatial coordinates, cell type labels, and top marker genes identified via Wilcoxon analysis (top 20 for MERFISH and Stereo-seq, top 10 for MIBI-TOF). These features serve as the primary inputs to our multi-agent annotation pipeline. Detailed dataset statistics and the exact marker-selection procedures are described in the Appendix A and Appendix B.1.

2.2. Spatial Context Embedding Generation

To capture each cell's local microenvironment, we first built a 30-nearest-neighbor graph based on spatial coordinates. We then reduced each cell's high-dimensional gene-expression vector to its first 30 principal components via PCA. These components and the graph structure were combined into a single data object for input to a graph autoencoder with two GCN layers. The resulting 30-dimensional embeddings capture spatial-transcriptomic structure. We further summarize each cell's context by computing the L2 norm of its embedding; this scalar, together with marker-gene evidence, is passed to our language models. Figure A.4 shows the generated embeddings. Appendix B.2 presents a detailed description of the spatial context encoding step.

2.3. Multi-Agent Chain-of-Thought Workflow

We designed a two-agent system following a principal–assistant architecture, implemented using the `crewai` package[1] and Gemini 2.5 Pro (Gemini, 2025) as base LLMs for both agents. The first agent assumes the role of a junior bioinformatics researcher; it combines marker-gene profiles and spatial embeddings to propose a cell-type label, along with a confidence score reflecting data support. Complementing this role, the second agent is a senior quality assurance bioinformatician with domain expertise in transcriptomics. This Agent critically evaluates the junior's annotations, verifying their biological plausibility by cross-referencing known gene markers and contemporary literature. In doing so, it refines the predicted labels and adjusts confidence scores to enhance the robustness and accuracy of the overall system. When necessary, the senior Agent can delegate the task back to the junior Agent, prompting further analysis or context gathering. This iterative loop continues until sufficient biological evidence is collected to support a confident and well-justified cell type classification.

While both agents can optionally query the web via the `Serper` API[2], we found that disabling external lookups yields faster inference with no loss in accuracy, so the operational TACTIC runs without web access.

Together, the agents emulate a human-in-the-loop annotation pipeline that blends exploratory reasoning with expert validation. We provide a detailed breakdown of each Agent's responsibilities and objectives in Appendix B.4. Finally, we measure annotation performance using standard classification metrics: accuracy and F1-score.

3. Results & Discussion

3.1. Benchmarking TACTIC across different spatial technologies

We first benchmarked TACTIC on three spatial-omics datasets: MIBI-TOF, MERFISH, and Drosophila Stereo-seq. On the MIBI-TOF dataset, TACTIC showed consistently strong performance as reported in Table 1, even when detecting rare cell types (Figure A.2). Notably, Fibroblasts were associated with higher classification error rates. Gene expression embeddings revealed a clear separation between major cell types, while spatial embeddings, although broadly aligned, exhibited signs of overfitting, suggesting potential limitations in spatial resolution or model generalizability.

Using the MERFISH dataset, which features higher transcriptomic and spatial complexity, agents exhibited a modest drop in overall performance (Figure A.1). Here rare cell types were more often mislabeled, reflecting the combined challenges of class imbalance and greater tissue heterogeneity. While gene expression embeddings successfully captured distinct cell identities, the spatial embeddings failed to delineate smaller cell populations, possibly reflecting limitations from tissue-level variability. The Drosophila Stereo-seq dataset, marked by complex cellular heterogeneity, revealed similar trends,again driven down by misclassi-

[1] https://www.crewai.com

[2] https://serper.dev

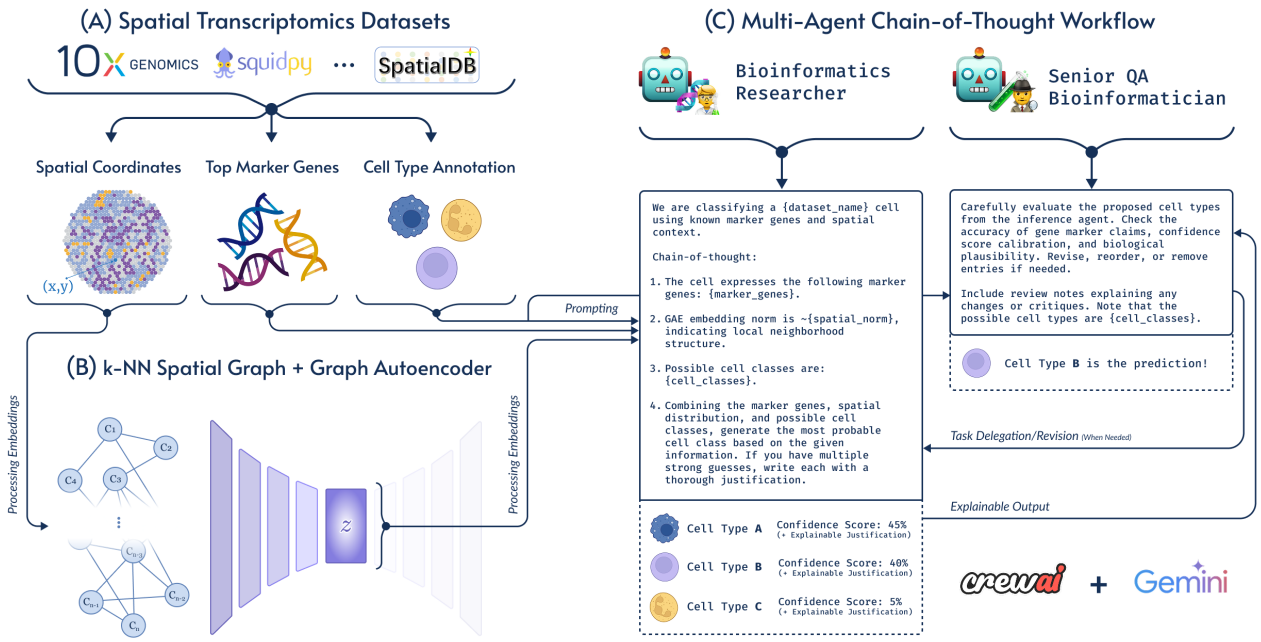


Figure 1. An overview of the TACTIC workflow for cell type classification: Spatial transcriptomic datasets from multiple databases and platforms are A) pre-processed to extract cell centroid coordinates and identify the top marker genes per class via differential expression;(B) a k-nearest-neighbor graph (k=30) is built on those centroids and, together with PCA-reduced expression (30 PCs), fed into a Graph Autoencoder (GAE) to generate low-dimensional spatial context embeddings; (C), the resulting embeddings, combined with marker-gene summaries, are assembled into Chain-of-Thought (CoT) prompts for the downstream multi-agent cell-type classification. Full details of the framework can be seen in Appendix B.4.

fications in rare cell groups.

Table 1. TACTIC Performance metrics across different datasets.

Dataset	Accuracy	F1 Score
MERFISH	0.83	0.80
MIBI-TOF	0.95	0.94
Drosophila Stereo-seq	0.46	0.46

In general, misclassifications across all datasets stem primarily from class imbalance and high transcriptomic similarity among certain rare cell types, as observed in (Alsabbagh et al., 2023), even in agents leveraging merely linguistic LMMs. This pattern is particularly evident in the MERFISH dataset, as shown in Figure C.5 and further supported by the class distribution in Figure A.1c. Rare cell populations such as Endothelial, Ependymal, Microglia, and Pericytes exhibit either substantial misclassification or complete absence from the predictions. Notably, OD Immature, a rare cell type, is consistently and accurately classified. We hypothesize that this exception is due to its well-defined transcriptomic signature, characterized by distinct marker genes, in contrast to the substantial overlap observed among the other misclassified rare cell types. The same phenomenon is depicted in the MIBI-TOF and Drosophila stereo-seq datasets, in Figures C.6 (A.2c) and C.7 (A.3c), respectively.

3.2. Ablation Studies

To understand which components of our pipeline (i.e., spatial context, agent architecture, and model selection) contribute most to these outcomes, we ran a series of ablation experiments on the MERFISH data. To this end we removed or randomized spatial embeddings, altered the agent collaboration and delegation structure, and evaluated different model pairings for inference and QA. For each experiment, we evaluated both agents using the same underlying language models, either Google 2.0 Flash (referred to interchangeably as the *Flash* LLM) or Google 2.5 Pro (referred to as the *Pro* LLM). The *Flash* LLM is optimized for low latency, efficient reasoning, and real-time interaction, making it well-suited for speed-critical tasks. In contrast, the *Pro* LLM offers enhanced reasoning capabilities, multimodal comprehension, and advanced coding proficiency, making it better equipped for more complex, context-rich scenarios. Due to computational limitations, all ablations were performed on the MERFISH dataset, which exhibits the highest degree of spatial stochasticity among the datasets in our evaluation.

When the spatial norm is removed from the CoT prompt

provided to the junior agent (Table 2), performance remains comparable to that of the control experiment. This contrasts with the findings reported in (Khan et al., 2025), where spatial norm was shown to have a significant impact. We attribute this discrepancy to the nature of our multi-agent setup, where agents engage in collaborative dialogue that tends to compensate for the absence of this specific contextual cue implicitly. The same phenomenon can be seen when completely randomizing the spatial norm in Table 3. Multi-agent justifications offer insight into the underlying reasoning processes that may explain this behavior. As illustrated in Listings in Appendix C, the agent's conversations often overlook certain pieces of contextual information, such as the spatial normalization, suggesting that key details can be diluted or lost during collaborative reasoning.

Table 2. Spatial context ablation results on MERFISH

Model	Accuracy	F1 Score
Gemini 2.0 Flash	0.680	0.657
Gemini 2.5 Pro	0.840	0.804

Table 3. Randomized spatial control results on MERFISH

Model	Accuracy	F1 Score
Gemini 2.0 Flash	0.670	0.630
Gemini 2.5 Pro	0.830	0.786

Next, we evaluated the role of the senior QA agent(Bioinformatician Agent) by (a) removing it entirely and (b) revoking its ability to delegate back to the junior agent. As shown in Table 4, neither scenario led to a significant quantitative change in performance. However, from a qualitative standpoint, the QA agent's authority proves essential. Its ability to demand more thorough and interpretable justifications from the junior agent enhances the transparency and explainability of the final classifications, an important factor for downstream validation and scientific trust.

Table 4. QA Agent ablation results on the MERFISH

Ablation	Model	Accuracy	F1 Score
No QA Agent	Gemini 2.0 Flash	0.750	0.708
No QA Agent	Gemini 2.5 Pro	0.830	0.791
No Delegation	Gemini 2.0 Flash	0.710	0.664
No Delegation	Gemini 2.5 Pro	0.840	0.814

Finally, we evaluated four LLM configurations by varying the assignment of *Flash* and *Pro* models to the Inference and QA agents. Our goal was to identify the most effective pairing of models for overall performance. As anticipated, the configuration with both agents running on the *Pro* model yielded the highest accuracy (Table 5). Interestingly, assigning *Pro* to the Inference agent and *Flash* to the QA agent outperformed the inverse setup. This finding aligns with our earlier results in Table 4, reinforcing that the QA agent contributes less to quantitative performance metrics, despite its qualitative importance to ensure explainability and oversight.

Table 5. Agent architecture model combinations on the MERFISH dataset

Configuration	Accuracy	F1 Score
Flash Inference + *Flash* QA	0.710	0.681
Pro Inference + *Flash* QA	0.790	0.771
Flash Inference + *Pro* QA	0.760	0.725
Pro Inference + *Pro* QA	0.820	0.779

4. Conclusion & Future Directions

We introduced TACTIC, a novel multi-agent system for explainable and collaborative cell type classification in spatial transcriptomics.

TACTIC integrates gene expression and spatial context through a team of specialized agents, enabling accurate cell type annotation across diverse platforms. Beyond performance, it generates transparent, human-readable justifications, tracing the reasoning behind each prediction. This interpretability makes TACTIC particularly well-suited for complex, multimodal spatial data, where multiple perspectives and conflicting evidence must be reconciled.

As spatial technologies evolve to capture information across molecular, cellular, and physiological scales, the need for systems that can synthesize such data and explain how it becomes critical. TACTIC's agentic design is uniquely positioned to address this surfacing tensions in data and constructing a rational path toward resolution.

Looking ahead, TACTIC offers a foundation for broader applications, including cell–cell communication inference and spatial neighborhood enrichment, where agent specialization and coordination can be further leveraged.

In summary, TACTIC showcases the power of multi-agent systems to improve classification accuracy and to deliver interpretable, modular, and extensible solutions for the next generation of spatial and single-cell transcriptomic challenges.

References

Alsabbagh, A. R., de Infante, A. M. R., Gomez-Cabrero, D., Kiani, N. A., Khan, S. A., and Tegnér, J. N. Foundation models meet imbalanced single-cell data when learning cell type annotations. *bioRxiv*, pp. 2023–10, 2023.

Didolkar, A., Goyal, A., Ke, N. R., Guo, S., Valko, M., Lillicrap, T., Rezende, D., Bengio, Y., Mozer, M., and Arora, S. Metacognitive capabilities of llms: An exploration in mathematical problem solving, 2024. URL https://arxiv.org/abs/2405.12205.

Gemini, T. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.

Hartmann, F. J., Mrdjen, D., McCaffrey, E., Glass, D. R., Greenwald, N. F., Bharadwaj, A., Khair, Z., Verberk, S. G., Baranski, A., Baskar, R., et al. Single-cell metabolic profiling of human cytotoxic t cells. *Nature biotechnology*, 39(2):186–197, 2021.

Khan, S. A., de Morentin, X. M., Lagani, V., Lehmann, R., Alsabbagh, A. R., Zahran, M., Kiani, N. A., Gomez-Cabrero, D., and Tegnér, J. Spell: Spatial prompting with chain-of-thought for zero-shot learning in spatial transcriptomics. In *ICLR 2025 Workshop on Machine Learning for Genomics Explorations*, 2025.

Lu, Q., Ding, J., Li, L., and Chang, Y. Graph contrastive learning of subcellular-resolution spatial transcriptomics improves cell type annotation and reveals critical molecular pathways. *Briefings in Bioinformatics*, 26(1):bbaf020, 2025.

Moffitt, J. R., Bambah-Mukku, D., Eichhorn, S. W., Vaughn, E., Shekhar, K., Perez, J. D., Rubinstein, N. D., Hao, J., Regev, A., Dulac, C., and Zhuang, X. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*, 362(6416):eaau5324, 2018.

Palla, G., Spitzer, H., Klein, M., Fischer, D., Schaar, A. C., Kuemmerle, L. B., Rybakov, S., Ibarra, I. L., Holmberg, O., Virshup, I., et al. Squidpy: a scalable framework for spatial omics analysis. *Nature methods*, 19(2):171–178, 2022.

Qiu, X., Zhu, D. Y., Lu, Y., Yao, J., Jing, Z., Min, K. H., Cheng, M., Pan, H., Zuo, L., King, S., et al. Spatiotemporal modeling of molecular holograms. *Cell*, 187(26): 7351–7373, 2024.

Shen, R., Liu, L., Wu, Z., Zhang, Y., Yuan, Z., Guo, J., Yang, F., Zhang, C., Chen, B., Feng, W., Liu, C., Guo, J., Fan, G., Zhang, Y., Li, Y., Xu, X., and Yao, J. Spatial-id: a cell typing method for spatially resolved transcriptomics via transfer learning and spatial embedding. *Nature Communications*, 13(1), 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-35288-0. URL https://www.nature.com/articles/s41467-022-35288-0.

Shi, X., Yang, Y., Ma, X., Zhou, Y., Guo, Z., Wang, C., and Liu, J. Probabilistic cell/domain-type assignment of spatial transcriptomics data with spatialanno. *Nucleic acids research*, 51(22):e115–e115, 2023.

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., and Lundeberg, J. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

Wu, J., Zhu, J., and Liu, Y. Agentic reasoning: Reasoning llms with tools for the deep research, 2025. URL https://arxiv.org/abs/2502.04644.

Yang, S., Gribovskaya, E., Kassner, N., Geva, M., and Riedel, S. Do large language models latently perform multi-hop reasoning?, 2024. URL https://arxiv.org/abs/2402.16837.

Zenil, H., Tegnér, J., Abrahão, F. S., Lavin, A., Kumar, V., Frey, J. G., Weller, A., Soldatova, L., Bundy, A. R., Jennings, N. R., Takahashi, K., Hunter, L., Dzeroski, S., Briggs, A., Gregory, F. D., Gomes, C. P., Rowe, J., Evans, J., Kitano, H., and King, R. The future of fundamental science led by generative closed-loop artificial intelligence, 2023. URL https://arxiv.org/abs/2307.07522.

Zhang, Y., Khan, S. A., Mahmud, A., Yang, H., Lavin, A., Levin, M., Frey, J., Dunnmon, J., Evans, J., Bundy, A., Dzeroski, S., Tegner, J., and Zenil, H. Advancing the scientific method with large language models: From hypothesis to discovery, 2025. URL https://arxiv.org/abs/2505.16477.

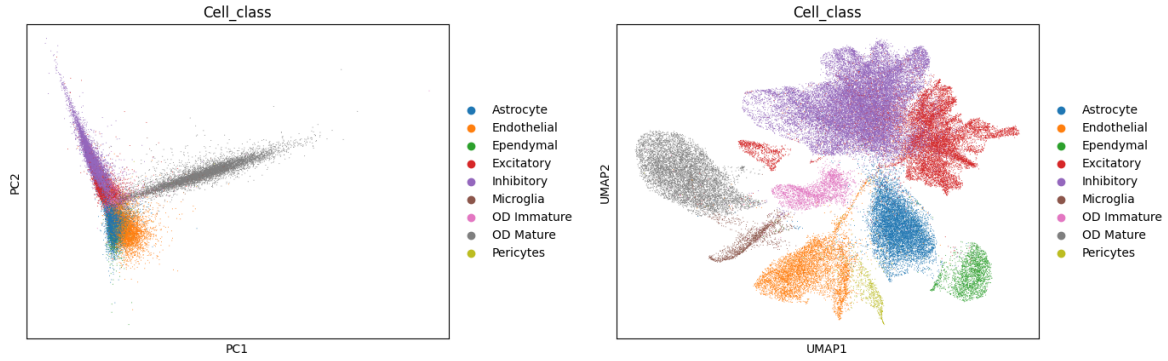
Appendix

A. Supplementary Dataset Information

To evaluate the TACTIC multi-agent system, we employed three diverse and well-established spatial omics datasets: MERFISH (Moffitt et al., 2018), MIBI-TOF (Hartmann et al., 2021), and Drosophila Stereo-seq (Qiu et al., 2024). These datasets span a range of organisms, tissue types, molecular modalities, and spatial complexities, allowing us to assess model generalizability and robustness across different biological contexts comprehensively.

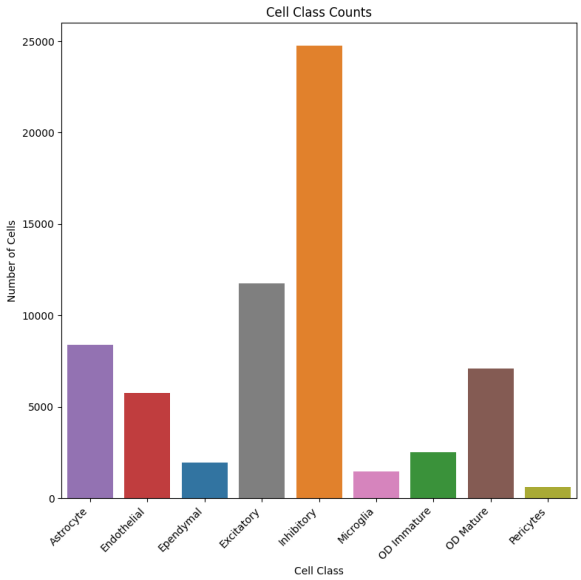
A.1. MERFISH Dataset

The MERFISH (Multiplexed Error-Robust Fluorescence In Situ Hybridization) dataset (Figure A.1) profiles single-cell spatial transcriptomics in the hypothalamic preoptic region of the mouse brain. It includes 73,655 spatially resolved spots, each described by 161 transcriptomic features. The original cell type annotations were aggregated into nine macro-labels to account for biologically meaningful groupings and reduce noise from rare subpopulations. MERFISH serves as a representative of large-scale, high-resolution transcriptomic spatial data with complex cell-type heterogeneity and spatial structure.



(a) PCA Profiles

(b) UMAP Profiles

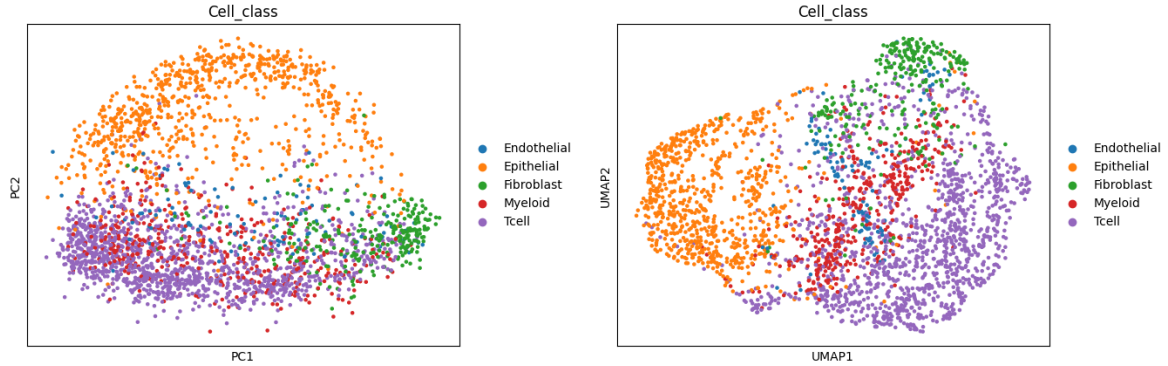


(c) Cell Counts

Figure A.1. Visualization of the MERFISH dataset via A) Principle Component analysis, B) UMAP, and C) the fraction of cell types according to the annotation provided in the original publication.

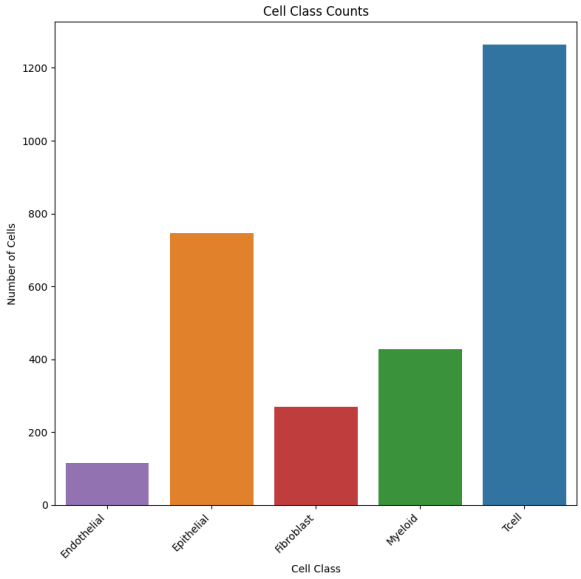
A.2. MIBI-TOF Dataset

MIBI-TOF (Multiplexed Ion Beam Imaging by Time-of-Flight) captures spatial proteomic data from the immune microenvironment at the invasive margin of human colorectal carcinoma. The dataset (Figure A.2) consists of 3,309 spots, each characterized by 36 protein expression features. Cell types were grouped into five macro-labels to reflect higher-order immune and stromal cell categories. This dataset presents unique challenges in spatial immune profiling due to high cell-type similarity and the complex tumor-immune interface.



(a) PCA Profiles

(b) UMAP Profiles

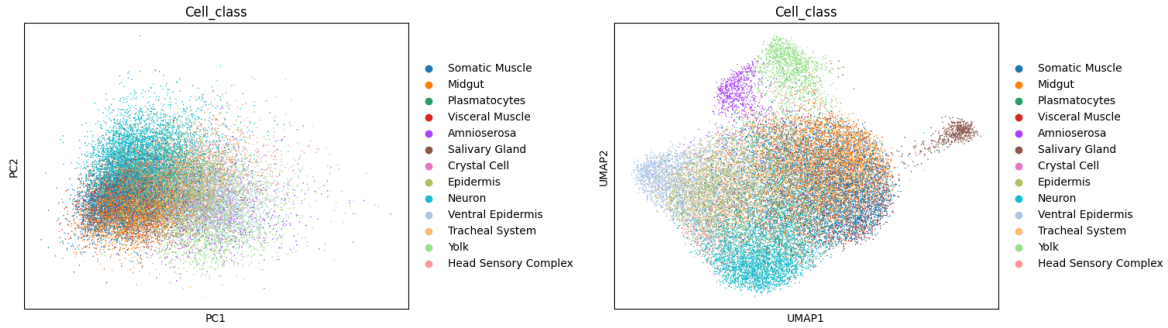


(c) Cell Counts

Figure A.2. MIBI-TOF dataset: PCA, UMAP, embeddings, and cell counts.

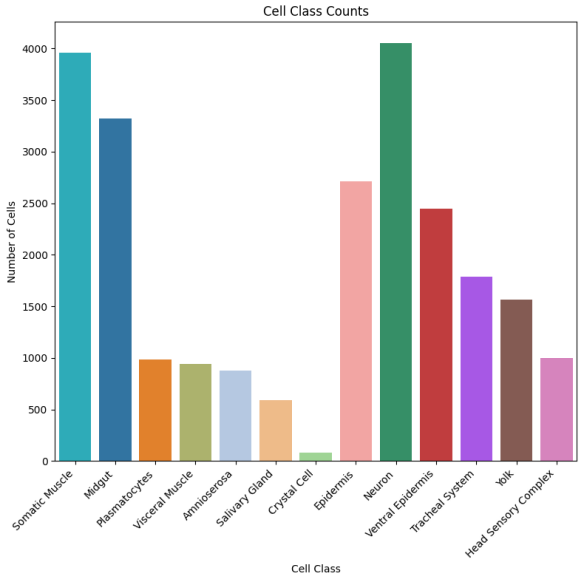
A.3. Drosophila Stereo-seq Dataset

The Drosophila Stereo-seq dataset (Figure A.3) captures the spatial transcriptomic landscape of Drosophila melanogaster embryos during developmental stages E9–10 h. It contains 24,327 spatial spots, each annotated with 8,484 gene expression features. Cell type annotations were aggregated into 13 macro-labels based on developmental lineages and spatial domains. With its high dimensionality and dynamic spatial organization, this dataset presents a unique opportunity to test model performance in early developmental biology and non-mammalian systems.



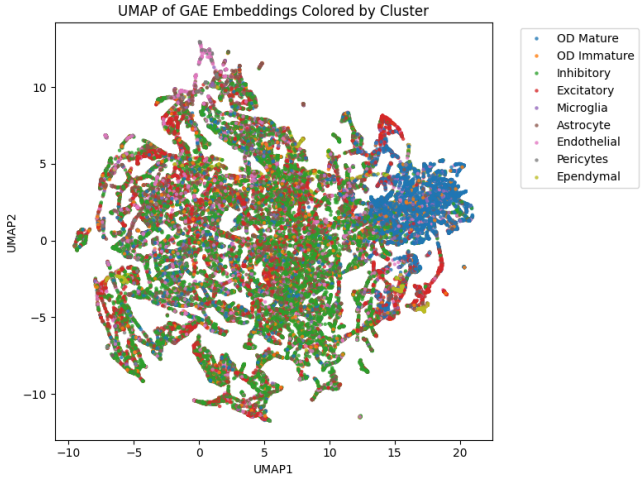
(a) PCA Profiles

(b) UMAP Profiles

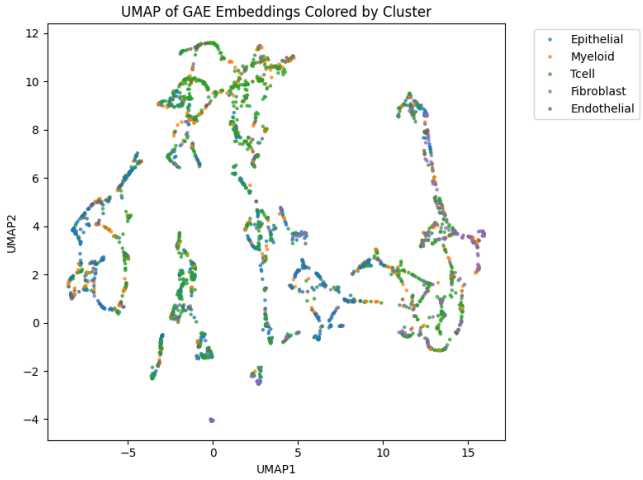


(c) Cell Counts

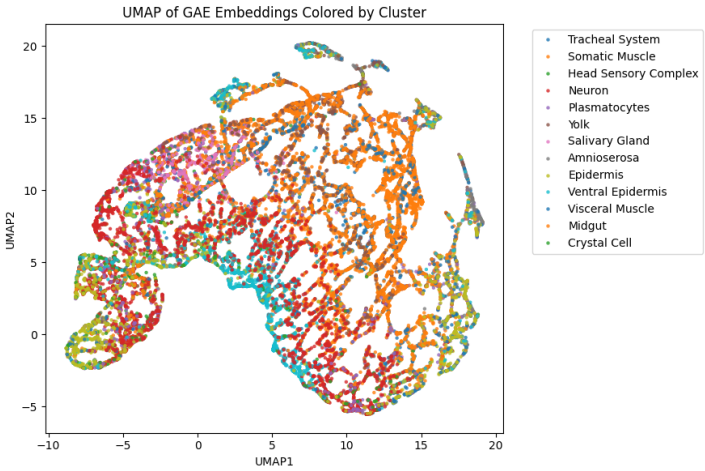
Figure A.3. Stereo-seq dataset: PCA, UMAP, embeddings, and cell counts.



(a) MERFISH



(b) MIBI-TOF



(c) Stereo-seq

Figure A.4. UMAP of Spatial Embeddings for Different Spatial Transcriptomics Datasets

B. Multi-Agent System Details

The multi-agent system inputs a spatial transcriptomic dataset and outputs a predicted cell type label for each cell iteratively. For each cell, a CoT prompt is generated with cell markers and spatial encoding sc res. Markers and spatial encoding are generated beforehand.

In the following section, we expand on each step and how it was done, including: DE Analysis and Marker identification, Spatial graph encoding, CoT prompting, Agent assembly and task assignment, and Outputs.

B.1. Spatial Transcriptomics Datasets & Preprocessing

To comprehensively assess the classification capabilities of multi-agent system with TACTIC framework, we employed three spatial omics datasets: MERFISH (Moffitt et al., 2018), MIBI-TOF (Hartmann et al., 2021) as curated by (Palla et al., 2022), and the Drosophila Stereo-seq dataset (Qiu et al., 2024). Each dataset offers distinct biological contexts and spatial complexities, enabling a rigorous evaluation across varied modalities and organ sms. More information about these datasets can be seen in Appendix A.

From each dataset, we extracted three key pieces of information per cell: spatial coordinates, cell type annotation, and transcriptomic markers. To address the inherent sparsity of single-cell transcriptomic profiles, we performed differential expression analysis using the Wilcoxon signed-rank test and identified top marker genes for each cell type. Specifically, we selected the top 20 marker genes for the MERFISH and Drosophila Stereo-seq datasets, and the top 10 for MIBI-TOF, which contains a considerably smaller gene set.

B.2. Spatial Context Embedding Generation

To incorporate spatial context into our classification framework, we begin by constructing an undirected k-nearest neighbor (k-NN) graph based on the spatial coordinates of individual cells, linking each cell to its 30 nearest neighbors. This graph structure effectively captures local spatial topology and is encoded as an adjacency matrix representing proximal cellular relationships.

Simultaneously, we apply Principal Component Analysis (PCA) to the gene expression profiles, preserving the top 30 components to reduce dimensionality while retaining key variance. These reduced expression features, together with the spatial graph, are integrated into a `PyTorch Geometric` object.

We then employ a Graph Autoencoder (GAE) architecture, composed of two Graph Convolutional Network (GCN) layers, to learn low-dimensional representations that encapsulate the combined structure of gene expression and spatial organization. The GAE is optimized to reconstruct the graph's edge structure, ensuring that the learned embeddings faithfully represent each cell's local neighborhood. The resulting 30-dimensional latent vectors serve as compact encodings of each cell's spatial and transcriptomic context.

B.3. Multi-Agent Chain-of-Thought Workflow

To enhance interpretability within the language model-driven classification pipeline, we compute the norm of each embedding vector, using it as a scalar summary of a cell's spatial environment, such as local density or isolation. This scalar is incorporated into the CoT prompt provided to the Agent, enriching the input with spatial cues that augment the raw gene expression data. By translating complex spatial relationships into a form that can be naturally integrated into textual reasoning, this method effectively bridges the domains of spatial transcriptomics and natural language processing.

The customized CoT prompt is described as follows:

Listing B.1. CoT Prompt Template iteratively modified with results of Cell Markers and Spatial Embedding score for each cell.

```
"We are classifying a {dataset_name} cell using known marker genes and
    spatial context.\n"
    "Chain-of-thought:\n"
    f"1. The cell expresses the following marker genes: {marker_genes}.\n"
    f"2. GAE embedding norm is ~{spatial_norm:.2f}, indicating local
        neighborhood structure.\n"
    f"3. Possible cell classes are: {str(cell_classes)}\n"
```

```
        "4. Combining the marker genes, spatial distribution, and possible cell
            classes, search the web if needed, and generate the most probable
            cell class based on the given information. If you have multiple
            strong guesses, write each with a thorough justification.\n"
    )
```

With that in hand, the three components (1) the marker list, (2) the spatial embeddings, and (3) the generated prompt, are then used as foundational inputs for the rest of our multi-agent system, TACTIC.

B.4. Multi-Agent Assignments

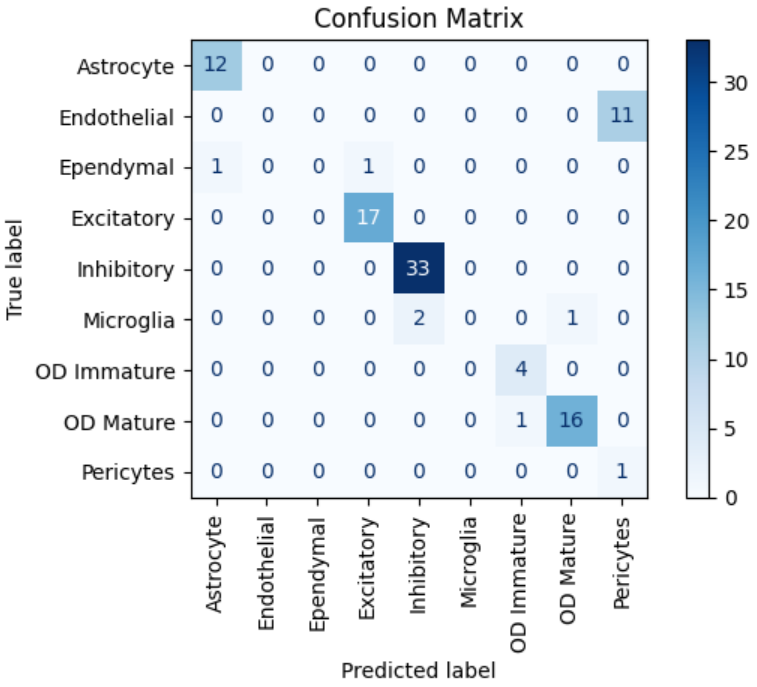
TACTIC consists of a two-agent system following a principal–assistant architecture. In our architecture, we name the principal "Senior Quality Assurance Bioinformatician" and the assistant "Bioinformatics Researcher". We outline the tasks assigned from each, their goals, backstory, whether they can delegate, the under-the-hood LLM, and the description of their outputs:

- **Bioinformatics Researcher**
  - **Goal:** Generate the most probable cell type(s) for a given dataset cell using a combination of known marker genes, spatial context, and reference databases. Additionally, provide a justification for each proposed label with clear biological reasoning.
  - **Backstory:** A junior computational biologist at a spatial genomics lab. Your focus is on decoding cellular identity. You are trained in marker gene databases, cell ontology, and spatial neighborhood anal sis. You are precise and cautious, always providing thorough justifications. When uncertain, you prefer to suggest multiple plausible options with confidence scores. You know how to search databases like NCBI, PanglaoDB, CellMarker, and more to support your findings.
  - **Can Delegate?** No.
  - **LLM Used:** Gemini-2.5-pro-preview-05-06.
  - **Output Description:** Given the predefined prompt described in section B.3, produce a dictionary of possible cell type predictions, each with a confidence score and justification based on marker gene expression and spatial con ext. If multiple hypotheses are plausible, include all and explain the ambiguity. All results are to be stored in a prediction.json file.
- **Senior Quality Assurance Bioinformatician**
  - **Goal:** Review and refine the output of the junior Agent. Ensure all predicted cell types are biologically plausible, appropriately justified, and consistent with known gene markers and spatial distributions.
  - **Backstory:** You are a senior scientist and quality assurance expert with 15+ years of experience in transcriptomics, including single-cell and spatial omics technologies. You scrutinize every prediction for accuracy, check against the latest gene marker studies, and adjust both labels and confidence scores where needed. Your mission is scientific precision and trustworthy results.
  - **Can Delegate?** Yes.
  - **LLM Used:** Gemini-2.5-pro-preview-05-06.
  - **Output Description:** Carefully evaluate the proposed cell types from the Bioinformatics Researcher agent. Check the accuracy of gene marker claims, confidence score calibration, and biological plausibility. Revise, reorder, or remove entries if needed. Include review notes explaining any changes or criticisms. Finally, present a single word, just the predicted class label.

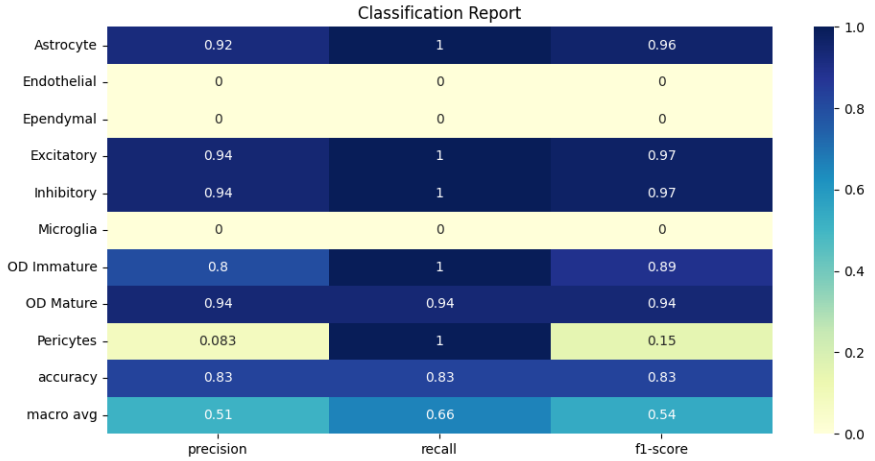
B.5. Evaluate the Results

Finally, for each predicted cell label, the accuracy and F1-scores are computed, and the confusion matrix and classification report are generated and visualized. Additionally, we visualize the gene expression embeddings as well as the latent spatial embeddings.

C. Detailed Presentation of Results

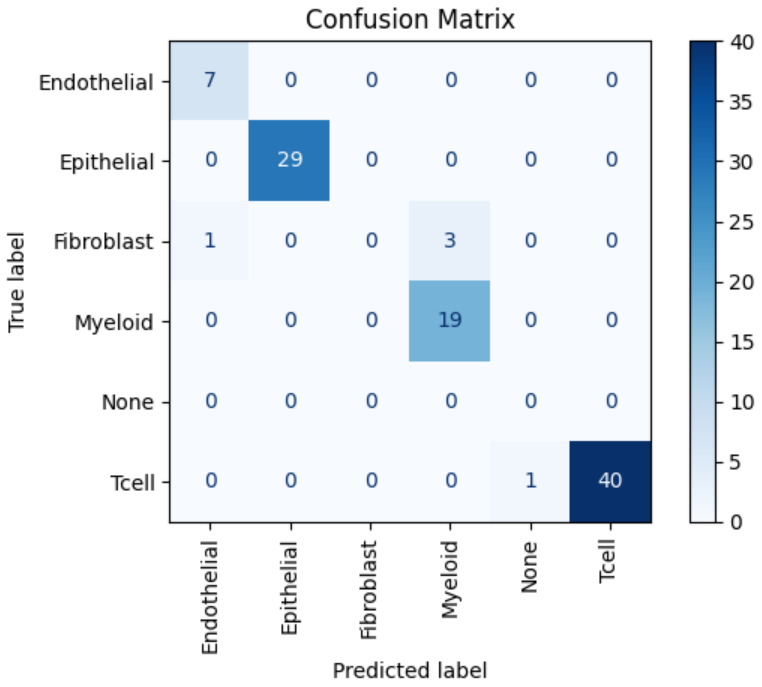


(a) Confusion Matrix



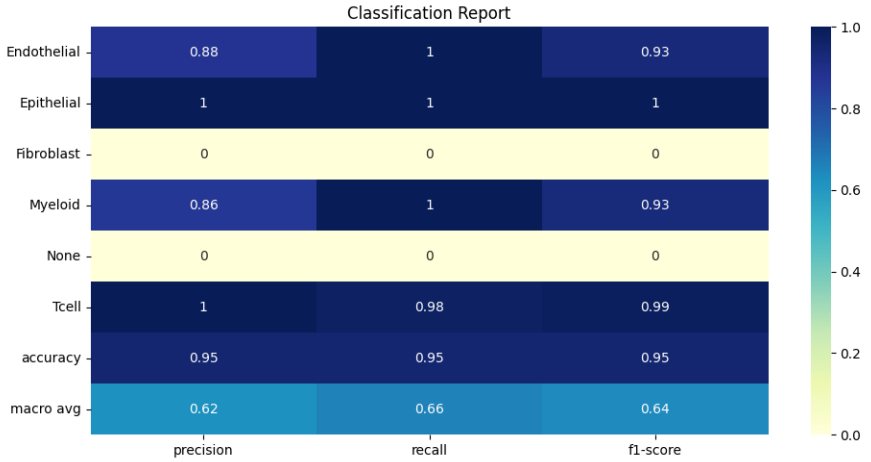
(b) Classification Report

Figure C.5. Classification performance of the MERFISH dataset



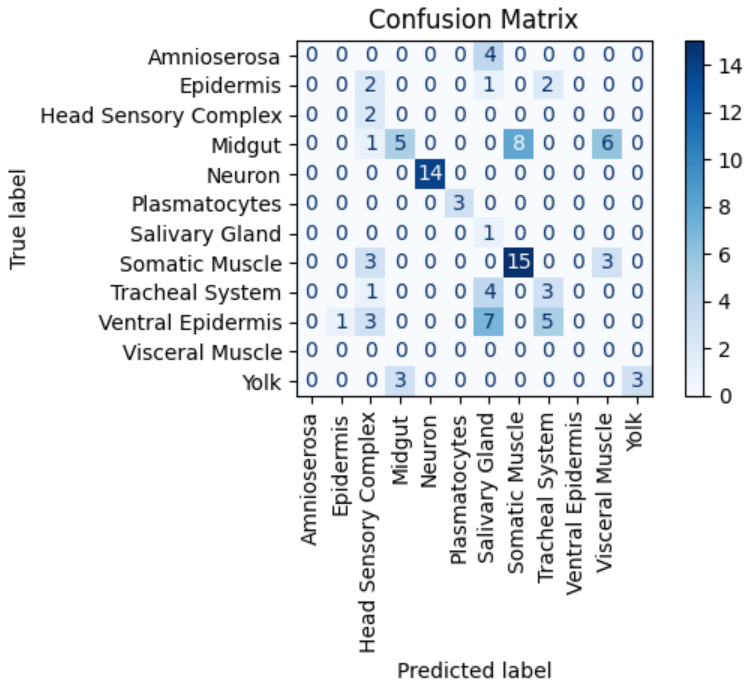
(a) Confusion Matrix

]

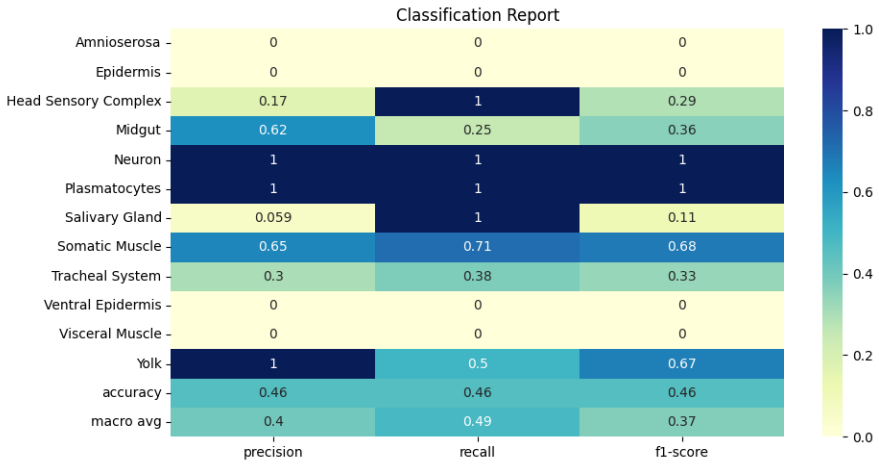


(b) Classification Report

Figure C.6. Classification performance of the MIBI-TOF dataset



(a) Confusion Matrix



(b) Classification Report

Figure C.7. Classification performance of the Drosophila Stereo-seq dataset

Listing C.2. Example of cell classification reasoning output on the bioinformatician researcher agent on the MERFISH dataset.

```
# Agent: Bioinformatics Researcher
## Final Answer:
{
  "predictions": [
    {
      "cell_type": "OD Mature (Mature Oligodendrocyte)",
      "confidence": 0.55,
      "justification": "This cell expresses multiple canonical and highly
          specific markers for mature oligodendrocytes: Mbp (Myelin Basic Protein
          ), Ermn (Ermin), Gjc3 (Connexin 29), and Opalin (Oligodendrocytic
          Myelin Paranodal And Nodal Axon ensheathing Protein). These genes are
          fundamental to oligodendrocyte identity and their primary function of
          myelination. Additional markers like Ndrg1 (N-Myc Downstream Regulated
          1) and Sox8 (SRY-Box Transcription Factor 8) can also be associated
          with the oligodendrocyte lineage (Sox8 more with development, Ndrg1
          with myelination). The GAE embedding norm of ~14.43, indicating local
          neighborhood structure, is consistent with the organized nature of
          oligodendrocytes forming myelin sheaths around axons within brain
          tissue. \n\nAmbiguity/Challenges: The primary challenge to this
          classification is the concurrent expression of a significant number of
          potent neuronal markers: Cck (Cholecystokinin), Trh (Thyrotropin
          Releasing Hormone), Gnrh1 (Gonadotropin Releasing Hormone 1), Sema4d (
          Semaphorin 4D, though also in ODs), Ucn3 (Urocortin 3), Omp (Olfactory
          Marker Protein), Sst (Somatostatin), Etv1 (ETS Variant Transcription
          Factor 1), and Adcyap1 (Adenylate Cyclase Activating Polypeptide 1).
          The co-expression of such a diverse suite of neuronal genes,
          particularly neuropeptides and the highly specific Omp, is highly
          atypical for a canonical mature oligodendrocyte. This unusual gene
          signature could suggest a specific reactive state of the
          oligodendrocyte, a rare or uncharacterized subtype, or potential
          contamination/doublet if the input data preparation allowed for it (
          though we assume a single cell as per the task). The Omp marker is
          especially puzzling if the tissue sample is not from an olfactory-
          related brain region, as it's a hallmark of olfactory sensory neurons."
    },
    {
      "cell_type": "Neuron (Excitatory or Inhibitory)",
      "confidence": 0.40,
      "justification": "This cell expresses a broad array of genes
          characteristically associated with neuronal identity and function.
          These include several neuropeptides such as Cck, Sst, Trh, Gnrh1, Ucn3,
           and Adcyap1, which are synthesized and released by neurons.
          Furthermore, it expresses Etv1, a transcription factor often found in
          specific excitatory neuron populations, and Omp (Olfactory Marker
          Protein), a highly specific marker for olfactory sensory neurons (its
          relevance is high if the tissue is from an olfactory area). Sema4d is
          also expressed in neurons. The GAE embedding norm of ~14.43, suggesting
           local neighborhood structure, is consistent with the organization of
          neurons into circuits, layers, or nuclei within the brain. The neuronal
           markers suggest a metabolically active neuron, possibly peptidergic.
          The mix of markers like Etv1 (often excitatory) and Sst (typically
          inhibitory) makes a more precise Excitatory/Inhibitory sub-
          classification difficult without further information.\n\nAmbiguity/
```

```
            Challenges: The most significant challenge to this classification is
            the strong expression of canonical oligodendrocyte markers: Mbp, Ermn,
            Gjc3, and Opalin. These genes are integral to myelin formation and are
            defining features of oligodendrocytes, not typically expressed by
            neurons. For this cell to be a neuron, it would represent a highly
            unusual or aberrant state characterized by the ectopic expression of
            these myelin-associated genes. Such a profile is not consistent with
            known neuronal subtypes."
        }
    ],
    "reasoning_for_exclusion_of_other_types": "Other cell types from the provided
        list ('Astrocyte', 'Endothelial', 'Ependymal', 'Microglia', 'OD Immature',
        'Pericytes') are considered significantly less probable due to the lack of
        strong positive marker gene expression for these lineages. For instance:\n-
         Astrocyte: No clear expression of canonical markers like Gfap, Aldh1l1, or
         Slc1a2.\n- Endothelial: No clear expression of markers like Cldn5, Pecam1,
         or Flt1.\n- Ependymal: No clear expression of markers like Foxj1 or
        Ccdc153.\n- Microglia: While Sema4d can be found in microglia, core
        microglial markers like Aif1 (Iba1), Tmem119, or Cx3cr1 are absent.\n- OD
        Immature (Oligodendrocyte Precursor Cell): While Lpar1 and Sox8 are
        expressed (consistent with OPCs/immature ODs), the strong expression of
        late-stage maturation markers like Mbp and particularly Opalin points more
        towards a mature oligodendrocyte rather than an OPC. OPCs typically do not
        express high levels of Opalin.\n- Pericytes: No clear expression of markers
         like Pdgfrb, Rgs5, or Cspg4.\n\nThe gene expression profile is
        overwhelmingly dominated by markers strongly associated with either mature
        oligodendrocytes or neurons, making these two categories the most plausible
        , albeit conflicting, hypotheses.",
    "overall_assessment": "The gene expression profile for this cell is highly
        ambiguous, presenting strong evidence for both mature oligodendrocyte and
        neuronal identities. This could be indicative of a rare cell state, a cell
        undergoing a highly unusual transcriptional program, or potentially a
        technical artifact such as a doublet (signal from two closely apposed cells
        , one oligodendrocyte and one neuron). Given the task to classify as a
        single cell from the provided list, both 'OD Mature' and 'Neuron (
        Excitatory or Inhibitory)' are presented as possibilities, with
        justifications highlighting the supporting and conflicting evidence for
        each. The GAE embedding norm indicates a structured local environment but
        does not strongly differentiate between these two possibilities without
        further dataset-specific context."
}
```

Listing C.3. Example of cell classification reasoning output on the senior quality assurance bioinformatician agent on the MERFISH dataset.

```
# Agent: Senior Quality Assurance Bioinformatician
## Final Answer:
OD Mature

True=OD Mature, Pred=OD Mature
```

Listing C.4. Example of cell classification reasoning output on the bioinformatics researcher agent on the MIBI-TOF dataset.

```
# Agent: Bioinformatics Researcher
Final Answer:
{
  "Tcell": {
    "confidence": 0.90,
    "justification": "The cell is most probably a T-cell based on the strong
       expression of the pan-leukocyte marker CD45 and the definitive T-cell
       lineage marker CD3. Further supporting a T-cell identity is the
       expression of CD4 and CD8. While mature peripheral T-cells are typically
       CD4+ (helper) or CD8+ (cytotoxic), the co-expression could indicate a CD4
       +CD8+ double-positive (DP) T-cell (which can be found peripherally, often
        highly activated or in specific pathological states), or it may signify
       that one of these markers is dominantly expressed while the other is
       present at a lower, yet detectable, l vel. The expression of PD1 (
       Programmed Death-1) and CD57 points towards an activated, terminally
       differentiated, and potentially exhausted or senescent T-cell phenotype.
       This state is common in chronic antigenic stimulation, such as in tumor
       microenvironments or chronic infections. The metabolic markers HK1 (
       Hexokinase 1) and PKM2 (Pyruvate Kinase M2) are consistent with a high
       metabolic activity, characteristic of activated T-cells. Vimentin
       expression, while classically mesenchymal, is also found in activated
       lymphocytes, including T-cells. The GAE embedding norm of ~1.30 suggests
       the cell is integrated within a local neighborhood structure, indicative
       of its presence in an organized tissue microenvironment where cell-cell
       interactions occur. The primary confounding marker is CD14, which is a
       canonical marker for monocytes and macrophages. However, its presence on
       a cell with such a strong T-cell signature can be explained by atypical
       expression. Literature suggests that T-cells can acquire CD14 from
       myeloid cells via trogocytosis (antigen-presenting cells in close contact
       , plausible given the GAE norm) or, more rarely, express CD14 de novo
       under specific inflammatory conditions. Given the overwhelming evidence
       for a T-cell lineage and activation state, this atypical CD14 expression
       is considered the most parsimonious explanation. Therefore, the cell is
       classified as a T-cell, likely in an activated/exhausted state, with an
       unusual CD14 expression profile potentially influenced by its
       microenvironment."
  },
  "Myeloid": {
    "confidence": 0.08,
    "justification": "This cell expresses CD45 (pan-leukocyte) and CD14 (monocyte
       /macrophage marker), which are characteristic of the myeloid lineage.
       Additionally, CD4 can be expressed by some myeloid cells (e.g.,
       macrophages, some dendritic cells), and PD1, vimentin, HK1, and PKM2 can
       also be found on activated myeloid cells. However, the strong expression
       of CD3, a definitive T-cell co-receptor, is highly uncharacteristic of
       myeloid cells. Furthermore, the presence of CD8 (typically cytotoxic T-
       cells/NK cells) and CD57 (marker of terminal differentiation/senescence
       on T-cells/NK cells) is also not typical for conventional myeloid
       populations like monocytes or macrophages. For this cell to be of myeloid
        origin, it would require the aberrant expression of multiple key T-cell
       lineage and differentiation markers (CD3, CD8, C57). This scenario is
       considered less probable than a T-cell atypically expressing a single
       myeloid marker (CD14), especially in a complex tissue microenvironment
```

```
        suggested by the GAE embedding norm."
  },
  "Endothelial": {
    "confidence": 0.005,
    "justification": "Endothelial cells are typically CD45 negative. The strong
       expression of CD45, along with T-cell specific markers like CD3, CD4, and
        CD8, makes an endothelial cell identity highly unlikely."
  },
  "Epithelial": {
    "confidence": 0.005,
    "justification": "Epithelial cells are typically CD45 negative and express
       cytokertins. The strong expression of CD45, along with T-cell specific
       markers like CD3, CD4, and CD8, makes an epithelial cell identity highly
       unlikely."
  },
  "Fibroblast": {
    "confidence": 0.01,
    "justification": "Fibroblasts are mesenchymal cells, typically CD45 negative,
        and express markers like vimentin (which this cell has) but not
       hematopoietic lineage markers. The strong expression of CD45, along with
       T-cell specific markers like CD3, CD4, and CD8, makes a fibroblast
       identity highly unlikely, despite vimentin expression, which can be
       shared."
  }
}
```

Listing C.5. Example of cell classification reasoning output on the senior quality assurance bioinformatician agent on the MIBI-TOF dataset.

```
# Agent: Senior Quality Assurance Bioinformatician
## Final Answer: Tcell

True=Tcell, Pred=Tcell
```

Listing C.6. Example of cell classification reasoning output on the bioinformatician researcher agent on the Drosophila Stereo-seq dataset.


```
# Agent: Bioinformatics Researcher
## Final Answer:
{
  "predictions": [
    {
      "cell_type": "Plasmatocytes",
      "confidence_score": 0.90,
      "justification": "The gene expression profile strongly supports the
          identification of this cell as a Plasmatocyte, a type of Drosophila
          hemocyte. \n\n**Key Supporting Marker Genes:**\n1. **NimC4:** This
          gene encodes a Nimrod C family protein, which is a well-established
          marker for plasmatocytes and is involved in phagocytosis. Its
          expression is a strong indicator of plasmatocyte identity.\n2. **Ppn (
          Pericardin):** Encodes a collagen IV-like extracellular matrix (ECM)
          protein. Plasmatocytes are known to synthesize and secrete ECM
          components, and Pericardin is associated with hemocytes and basement
          membranes. \n3. **CG6310 (Glutactin):** Another ECM protein known to
          be expressed by hemocytes, including plasmatocytes, and the fat body.\n
          \n**Additional Supporting Evidence:**\n*   **High Secretory Activity:**
           The expression of multiple genes associated with the endoplasmic
          reticulum and protein folding/secretion (Pdi, Calr, CG43799 (PDI-like 1
          A), Fkbp14, ERp60) suggests high activity in protein synthesis and
          secretion. This is consistent with plasmatocytes producing and
          secreting ECM proteins, antimicrobial peptides, and other immune
          effector molecules.\n*   **Phagocytic/Lysosomal Activity:** Expression
          of **Sap-r** (Sphingolipid activator protein receptor), involved in
          lysosomal function, aligns with the phagocytic role of plasmatocytes.\n
          *   **Dynamic Cytoskeleton:** Expression of **chic** (Profilin) and **
          sn** (fascin), both involved in actin cytoskeleton organization,
          supports cell motility, morphological changes, and phagocytosis, which
          are characteristic functions of plasmatocytes.\n*   **General Actin
          Expression:** **Act5C** and **Act42A** are abundant actin isoforms,
          providing the structural components for these cytoskeletal activities.\
          n*   **scf (stem cell factor):** While its role in Drosophila is less
          defined than in vertebrates, its vertebrate homolog is involved in
          hematopoiesis, which could be relevant for hemocyte maintenance or
          function.\n\n**Spatial Context (GAE embedding norm ~ 1.41):**\nThis
          value indicates that the cell possesses a defined local neighborhood
          structure within the GAE embedding. This means the cell is likely
          surrounded by cells with similar transcriptional profiles (e.g., other
          plasmatocytes in a cluster) or cells with which it is functionally
          interacting (e.g., at a site of inflammation, injury, or tissue
          surveillance). This context is consistent with the known behavior of
          plasmatocytes.\n\n**Discussion of *fax* (failed axon connections)
          Expression:**\nThe expression of *fax*, typically a neural-specific
          marker involved in cholesterol esterification, is an interesting
          finding in a cell otherwise strongly identified as a plasmatocyte.
          Several possibilities could explain this:\n1. **Neuro-immune
          Interaction:** The plasmatocyte might be interacting with neural tissue
           (neurons or glia). Drosophila hemocytes are known to interact with the
           nervous system, especially during development, injury, or infection.
          The GAE norm supports such localized interactions.\n2. **Specific
          Plasmatocyte Subtype/State:** This could represent a specific subtype
```

```
        of plasmatocyte or a particular activation state that involves *fax*
        expression, possibly related to functions at the interface with the
        nervous system or specific metabolic requirements.\n3. **Undocumented
        Expression:** *fax* might have broader expression patterns or roles
        than currently documented, including low-level or context-dependent
        expression in certain hemocyte populations.\n\nDespite the *fax*
        expression, the collective weight of evidence from multiple canonical
        plasmatocyte markers (NimC4, Ppn, CG6310) and functional pathway
        indicators (secretion, phagocytosis, motility) makes 'Plasmatocytes'
        the most probable cell type. The *fax* expression warrants further
        investigation as it might highlight a specific functional context or
        subtype of plasmatocyte."
    },
    {
      "cell_type": "Neuron",
      "confidence_score": 0.15,
      "justification": "The expression of **fax (failed axon connections)**, a
         known neural-specific marker involved in cholesterol esterification and
          axon guidance, provides some evidence for a neuronal identity. **
        Akap200** is also expressed in the nervous system, among other tissues.
         \n\n**However, this hypothesis is significantly weaker due to:**\n1.
        **Presence of Strong Non-Neural Markers:** The cell strongly expresses
        **NimC4**, a canonical plasmatocyte (hemocyte) marker. It also
        expresses **Ppn (Pericardin)** and **CG6310 (Glutactin)**, ECM proteins
         typically produced by hemocytes and fat body, not primarily neurons.\
        n2. **High Secretory and Phagocytic Profile:** The suite of genes
        indicating high ER activity (Pdi, Calr, ERp60, etc.) and potential
        phagocytic/lysosomal activity (Sap-r) is more characteristic of
        secretory immune cells like plasmatocytes than typical neurons,
        although some neuronal subtypes can be secretory.\n3. **Lack of Other
        Core Neuronal Markers:** While *fax* is present, a broader signature of
         neuronal identity (e.g., specific neurotransmitter pathway genes, ion
        channels, other pan-neuronal markers like *elav* or *nSyb*) is not
        evident from the provided list, which would be expected for a
        definitive neuronal classification.\n\n**Spatial Context (GAE embedding
         norm ~ 1.41):**\nIf this cell were a neuron, its neighborhood might
        consist of other neurons or glial cells. The *fax* expression could be
        consistent with this. \n\n**Conclusion:** While *fax* is a notable
        neural marker, its presence is overshadowed by a much stronger and
        broader signature for plasmatocytes. The neuronal hypothesis is less
        likely unless *fax* is considered in isolation, which would ignore
        substantial contradictory evidence."
    }
  ]
}
```


Listing C.7. Example of cell classification reasoning output on the senior quality assurance bioinformatician agent on the Drosophila Stereo-seq dataset.

```
# Agent: Senior Quality Assurance Bioinformatician
## Final Answer:
Plasmatocytes
```

True=Plasmatocytes , Pred=Plasmatocytes