Large Language Model Agents for Pixel-Level Cell Type Classification in Imaging Mass Cytometry

Campbell Kieran, Lunenfeld-Tanenbaum Research Institute, Supervisor

Sarah Asbury, University of Toronto, Mentor

Raya Yahya Abu Aljamal, University of Jeddah, Project lead

Abstract

Imaging Mass Cytometry (IMC) enables the detection of high-dimensional protein markers, offering insight into the spatial and cellular heterogeneity of human tissues. While recent advances in 5-micron resolution allow high-throughput tissue analysis, traditional cell segmentation methods often fail on lower-resolution IMC images. To address this limitation, we propose a novel computational framework that implements pixel-level cell classification for segmentation-independent analysis. Our approach leverages unsupervised clustering of individual pixels based on multidimensional antibody expression. In prior work, we demonstrated that pixel clusters can distinguish between positive and negative marker expression across an antibody panel, enabling detection of binary expression signatures for each pixel. We aim to develop a context-aware AI agent system powered by a large language model (LLM) to classify each pixel as a specific cell type from the CELLxGENE library, a mixture of cell types, or background noise. The antibody expression profiles from classified pure-cell pixels will be used to deconvolve the composition of mixed pixels and filter out background noise. Integrating context-aware agent systems offers a scalable and interpretable approach for the biological annotation of complex, high-dimensional pixel-level spatial proteomics data.

Project Objectives and Goals

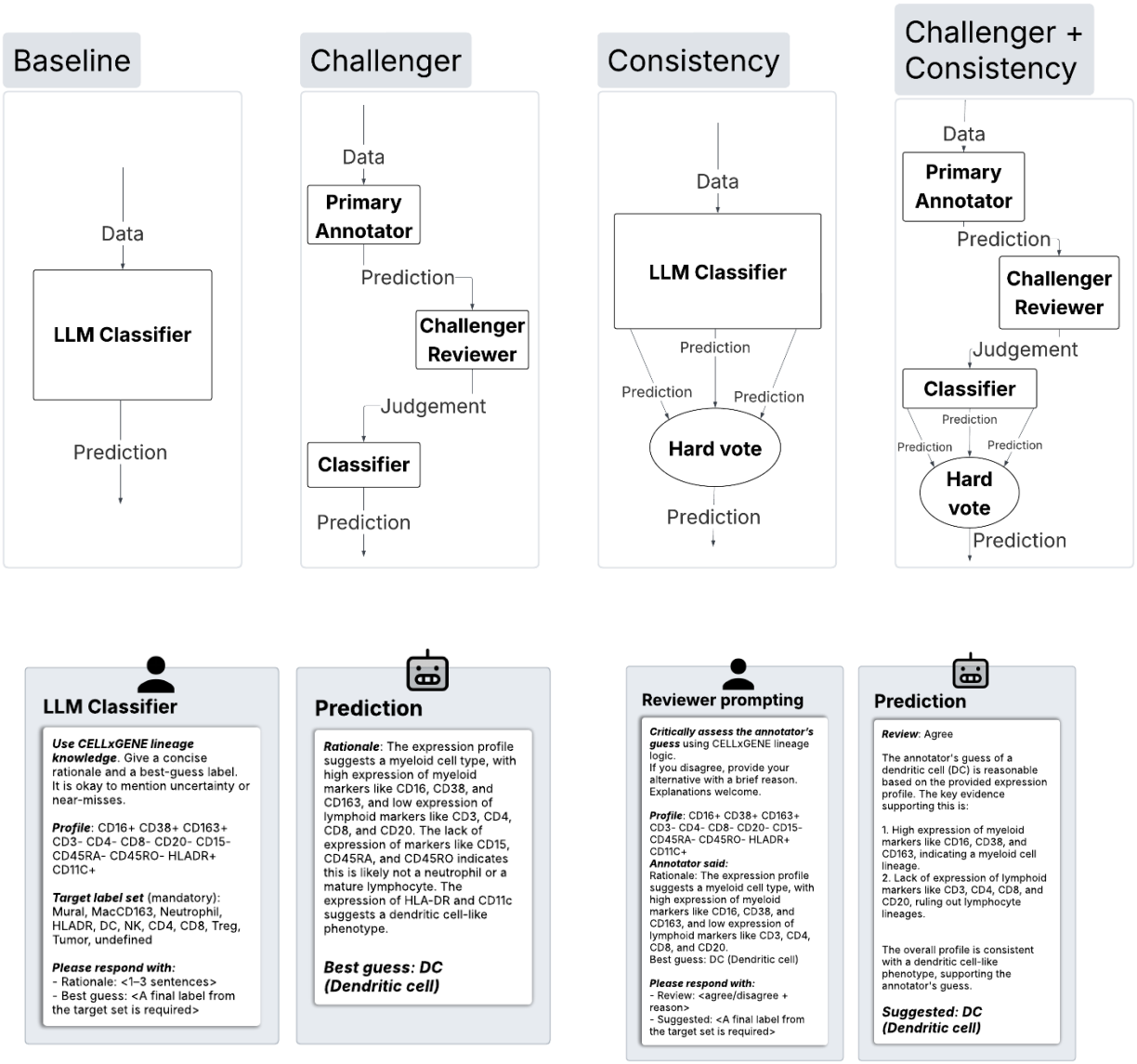
Imaging Mass Cytometry is a powerful technology for profiling the expression of dozens of protein markers on tissue sections. Conventional analysis requires accurate cell segmentation, but segmentation is error-prone because IMC images have low spatial resolution, ambiguous cell boundaries and densely packed cells[1]. These segmentation errors lead to merged or split cells and signal contamination from neighbouring cells, compromising downstream analyses[1]. Our project aims to circumvent these issues by:

1. Building a segmentation-free workflow that classifies individual pixels directly from marker intensities without delineating cells.
2. Developing a multi-agent LLM system that assigns each pixel to one of the following categories:
3. Known immune or tumour cell types based on reference atlases (e.g., CellxGene).
4. Mixtures of cell types where marker signals overlap.
5. Background noise and artifacts.
6. Using “pure-pixel” profiles from manually annotated cells to deconvolve mixed pixels.
7. Implementing diverse prompting architectures: baseline, challenger, majority vote and challenger-consistency, to improve robustness and spatial coherence.
8. Creating an evaluation pipeline to benchmark accuracy, precision, recall and F1 score at different spatial resolutions (1 μm and 5 μm).

Current Progress and Milestones Achieved

- Data preparation: curated IMC slices from the IMMUcan database, including tumour, CD4, regulatory T cells, natural killer (NK) cells and macrophages. Extracted both 1 μm and 5 μm pixel grids.
- Implemented four prompting architectures for LLM-based classification:
- Baseline architecture: a single LLM receives a description of the pixel’s marker profile along with the allowed labels and returns a predicted cell type in a pre-specified output format.

- Challenger pipeline: the baseline prediction is sent to a second LLM that reviews the reasoning and provides its own opinion. A third LLM then arbitrates between the two outputs to produce a final verdict.
- Majority-vote architecture: the baseline prompt is issued multiple times independently to mitigate randomness; the most frequent label among the runs is selected as the prediction.
- Challenger-consistency hybrid: combines the challenger pipeline with a spatial consistency check. After the challenger's assessment, the predicted label is concatenated with information from neighbouring pixels to encourage spatial coherence.



- The earlier marker-selection agent has been folded into future-work plans; rather than being a standalone architecture, marker selection will be implemented via retrieval-augmented generation.
- Balanced class representation in the training set to reduce bias towards abundant cell types.
- Designed an evaluation pipeline to compute accuracy, precision, recall, F1 score and confusion matrices for each architecture.
- Visualised results by generating spatial overlay maps comparing predicted and ground-truth cell types.
- Presented preliminary results at the KAUST AI Academy Summer School and received feedback from mentors and peers.

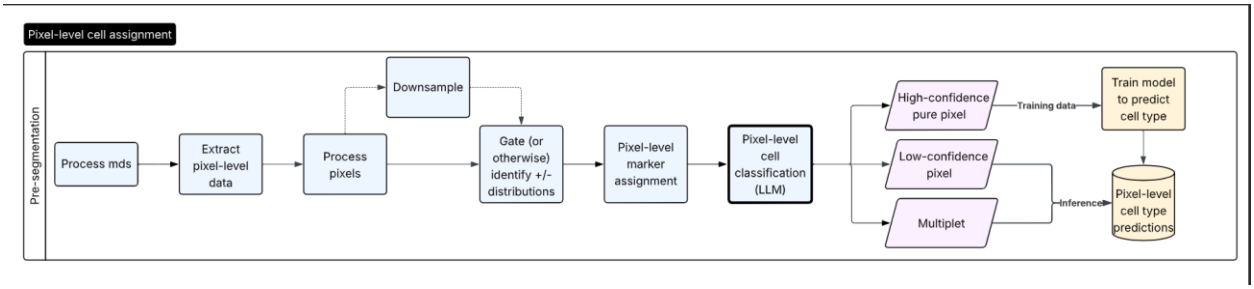
Methodology and Approach

1. Pre-processing and pixel-level marker assignment – Raw IMC data (MCD files) are processed to extract pixel-level marker intensities. Pixel grids are generated at both 1 μm and 5 μm resolutions. Images can be downsampled to 5 μm to cover larger tissue areas. For each marker we estimate positive/negative distributions (gating) and assign each pixel a binary marker state, producing a pixel-level marker assignment.
2. Pixel embedding and prompt design – Using the binary marker assignments, each pixel is encoded as a feature vector and converted into a natural-language prompt describing marker presence or absence. These prompts are passed to a base LLM (e.g., GPT-4 or Claude-3) that outputs a likely cell type along with a justification.
3. Prompting architectures and classification categories – To improve reliability we experimented with four prompting strategies:
4. Baseline: a single-agent prediction from the marker profile.
5. Challenger pipeline: the baseline prediction is reviewed by a second LLM, and a third LLM arbitrates between them, emulating peer-review.
6. Majority vote: multiple independent baseline prompts are issued and the most frequent label is chosen to reduce randomness.

7. Challenger-consistency hybrid: the challenger pipeline is augmented with a spatial consistency check that concatenates neighbouring pixel information to encourage coherent predictions.

The LLM's output assigns each pixel to one of three categories: high-confidence pure pixel, low-confidence pixel, or multiplet (pixels containing mixed cell signals). High-confidence pure pixels are used as training data for a conventional supervised model that learns to predict cell type directly from the marker assignments, enabling efficient inference on large datasets.

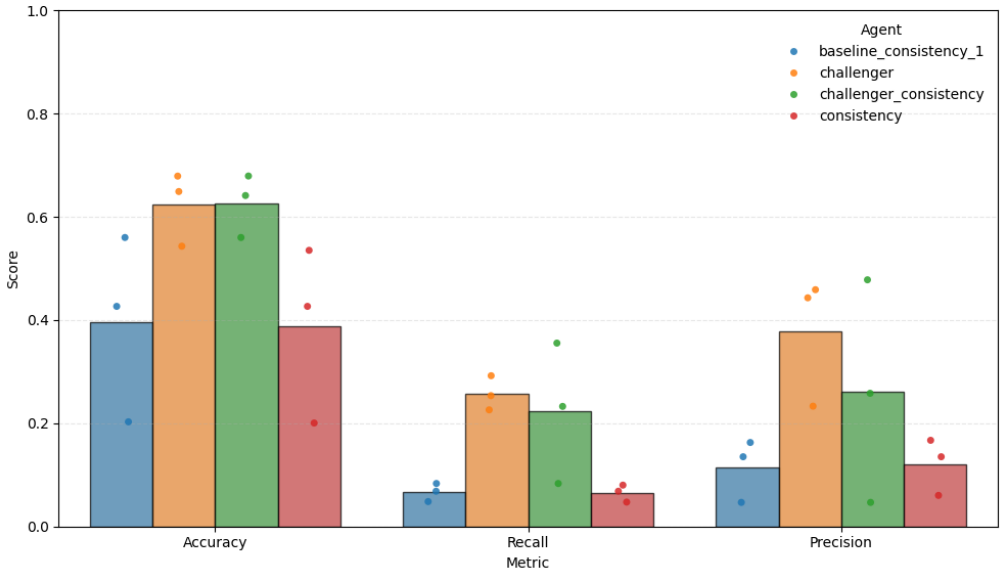
1. Debate and refinement – In the challenger architectures, agents provide reasons for their predictions. When disagreements occur, a debate ensues, forcing the models to justify their reasoning. This reduces hallucination and encourages consensus.
2. Evaluation and inference – We evaluate each architecture and the subsequent trained classifier using accuracy, precision, recall and F1 score on held-out IMC slices. Confusion matrices and spatial overlays are computed to analyse misclassifications at both 1 μm and 5 μm resolutions. The trained model is then used in inference mode to produce pixel-level cell type predictions across the entire tissue.



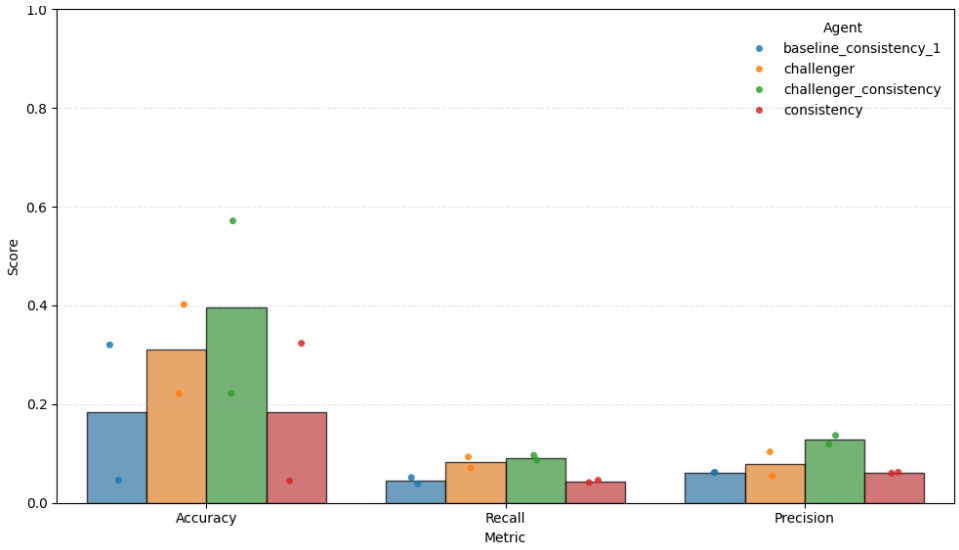
Preliminary Results and Findings

- The baseline architecture provides a useful starting point but suffers on ambiguous pixels where marker expression overlaps.
- The challenger pipeline and challenger-consistency hybrid achieve higher accuracy, precision and recall. Reviewing the baseline decision and enforcing spatial coherence reduce misclassifications.
- The majority-vote architecture helps stabilize predictions by mitigating random fluctuations but generally underperforms the challenger approaches.

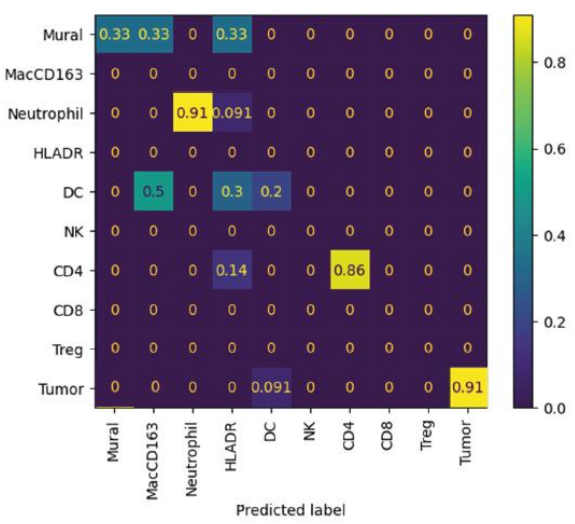
- Adding natural-language explanations to the prompts leads to better accuracy compared with requesting only the label, suggesting that forcing the model to justify its answer improves internal reasoning.

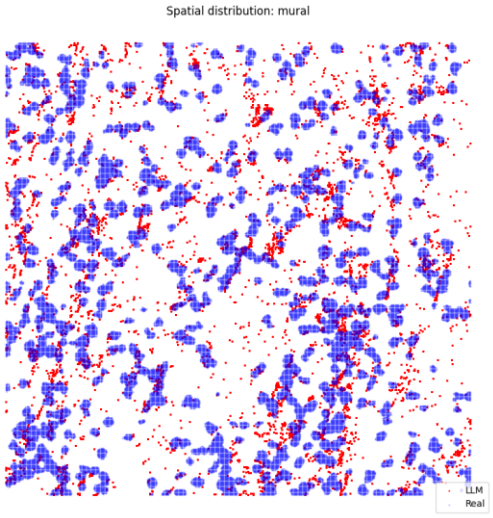


- The multi-agent architectures perform particularly well on 5 µm resolution data, compensating for the lower spatial resolution and making large-area scans viable for automated annotation.



- Class balancing improves detection of minority cell types such as regulatory T cells.
- Spatial overlay maps show that predictions become smoother and more consistent after applying the challenger-consistency architecture.





Challenges and Obstacles

- Marker ambiguity – Several cell types share similar markers (e.g., CD4 is expressed on tumour cells, regulatory T cells and helper T cells), making classification challenging.
- LLM hallucination – Base LLMs occasionally generate plausible but incorrect explanations. The debate framework mitigates this by forcing agents to justify their reasoning.
- Noisy markers – Some antibody channels contain background noise or low signal, leading to spurious classifications; determining optimal thresholds is ongoing.
- Scalability – Running multiple agents on millions of pixels incurs computational overhead. We mitigate this by caching intermediate results and batching prompts.

Next Steps and Future Work

- Iterative learning from true positives – Fine-tune the annotator using correctly classified pixels to improve accuracy on ambiguous cases.
- Orchestrating agents with a workflow engine – Implement a LangChain-style orchestrator that coordinates the baseline, challenger and consistency agents efficiently.
- Retrieval-augmented generation (RAG) – Integrate RAG to dynamically select the most informative marker panel based on the tissue context. This will replace the previous marker-selection agent.

- Atlas-level annotation – Extend the pipeline to large-scale tissue atlases, enabling comprehensive cell-type mapping across multiple organs and diseases.

Resources and Tools Used

- Datasets – IMMUcan database of manually annotated IMC slices with both 1 μm and 5 μm resolution.
- Large language models – Claude-3 and GPT-4 accessed via API calls for classification and debate.
- Programming languages and libraries – Python, Pandas and Scikit-learn for data processing and evaluation; Matplotlib for generating confusion matrices and overlay plots.
- Infrastructure – Local workstation for preprocessing and cloud-based compute for running LLMs.

Acknowledgements

We thank Dr. Kieran Campbell and Sarah Asbury for their mentorship, and the Campbell Lab and Jackson Lab for providing IMC datasets and technical advice. We also acknowledge support from KAUST Academy, KAUST NTPD and the University of Toronto Data Sciences Institute.