Position Paper

Towards Automated Biological Discovery: LLM-based Multi-Agent AI System for Single-Cell Disease Marker Identification and Discovery

Jieli Zhou[*], **Luting Zhou**[*], **Hongyi Xin**
Shanghai Jiao Tong University, Shanghai, China
[*]These authors contributed equally to this work.

The advent of single-cell RNA sequencing has revolutionized our understanding of cellular heterogeneity, yet identifying reliable disease markers remains a major challenge. Current approaches are bottlenecked by the complex interplay of data analysis, time-intensive literature validation, and the limitations of human domain knowledge. Inspired by recent breakthroughs in multi-agent systems (Swanson et al. 2024)(Lu et al. 2024), we present MarkerAgent, a LLM-based multi-agent AI system that automates and accelerates disease marker discovery from single-cell data.

MarkerAgent orchestrates five specialized AI agents designed to work in concert for marker discovery, as shown in Figure 1. The Data Analysis Agent processes scRNA sequencing data to identify differentially expressed genes and cell-type-specific patterns. The Literature Analysis Agent leverages graph retrieval-augmented techniques to conduct a comprehensive literature review, examining disease associations and pathway involvement. The Hypothesis Agent employs MarkerAnalogyPrompting (MAP), a novel knowledge-based prompting mechanism for discovering potential markers through cross-disease analogies. The Validation Agent evaluates the candidates using a systematic scoring framework that considers the importance of differential expression, support from the literature, and relevance of the pathway. The Report Agent synthesizes the findings into actionable reports detailing the evidence chains and validation strategies.

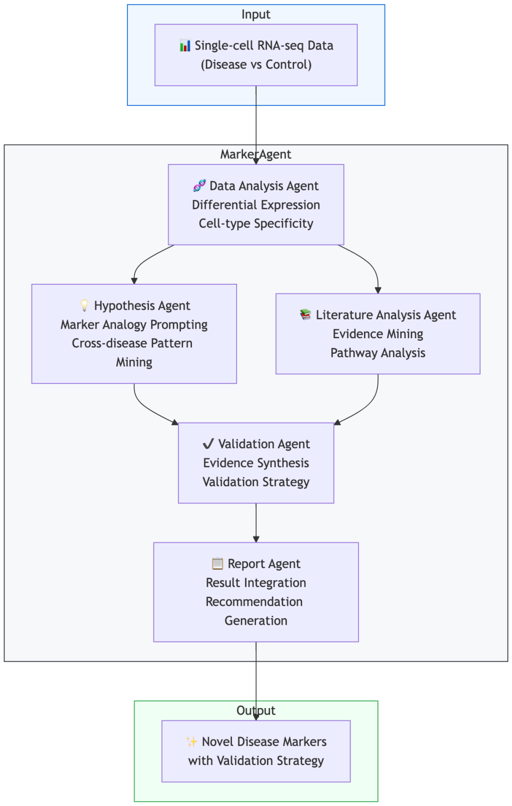

Figure 1: MarkerAgent architecture. Each agent is based on a LLM. The overall architecture is developed with CrewAI.

We demonstrate our framework through a case study of cholesteatoma marker discovery, based on recently published single-cell sequencing data (Shimizu et al. 2023). The Data Analysis Agent identifies INHBA as a significantly upregulated gene in cholesteatoma-specific fibroblasts, alongside established markers COL3A1, COL10A1, and COL11A1. The Literature Analysis Agent reveals important connections between INHBA and osteoclastogenesis through activin A signaling, aligning with the bone destruction phenotype documented in the original study. The Hypothesis Agent generates a novel hypothesis connecting INHBA's role in cholesteatoma to inflammatory pathways in rheumatoid arthritis. The Validation Agent evaluates this candidate by scoring its differential expression significance, literature evidence strength, and pathway relevance. Finally, the Report Agent synthesizes these findings into an evidence chain, prioritizing INHBA for experimental validation and demonstrating MarkerAgent's potential to accelerate biological discovery through automated, systematic analysis.

MarkerAgent represents a paradigm shift in biological discovery. Rather than manualy identifying differentially expressed genes and searching for literature support and come up with hypotheses, MarkerAgent demonstrates one potential way to automate the entire scientific discovery process. Looking forward, we envision MarkerAgent as a stepping stone towards fully automated scientific discovery.

References

Lu, C.; Lu, C.; Lange, R. T.; Foerster, J.; Clune, J.; and Ha, D. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Shimizu, K.; Kikuta, J.; Ohta, Y.; Uchida, Y.; Miyamoto, Y.; Morimoto, A.; Yari, S.; Sato, T.; Kamakura, T.; Oshima, K.; et al. 2023. Single-cell transcriptomics of human cholesteatoma identifies an activin a-producing osteoclastogenic fibroblast subset inducing bone destruction. *Nature Communications* 14(1):4417.

Swanson, K.; Wu, W.; Bulaong, N. L.; Pak, J. E.; and Zou, J. 2024. The virtual lab: Ai agents design new sars-cov-2 nanobodies with experimental validation. *bioRxiv* 2024–11.