Bridging LLM-based planning and workflow languages for automated, validated, scalable exploration of scRNA-seq analyses

Yoshinori Hayakawa [1] **Haruka Ozaki** [1]

[1] *Laboratory for AI Biology, RIKEN Center for Biosystems Dynamics Research*. Correspondence to: Haruka Ozaki haruka.ozaki@riken.jp.

1. Introduction

Single-cell RNA sequencing (scRNA-seq) has become a standard technique in biomedical research, enabling the characterization of cellular diversity at single-cell resolution [1]. Despite its widespread adoption, scRNA-seq analysis is highly dependent on the choice of methods and parameters, and different analysis pipelines can lead to substantially different biological conclusions, making trial-and-error–based pipeline design common in practice and raising concerns about analyst bias and poor reproducibility [2]. Moreover, scRNA-seq datasets support multiple valid analysis tasks [3], such that comprehensive biological interpretation often requires exploring diverse analysis strategies in parallel rather than committing to a single pipeline. This complexity motivates the need for computational systems that can support flexible and scalable exploration of analysis workflows.

Workflow languages such as Nextflow and Snakemake enable stable and reproducible execution, but they are not well suited for rapid modification or exploratory analysis when used in isolation [4, 5]. In contrast, large language model (LLM) agents can flexibly generate analysis plans and code, yet they often suffer from unstable operation due to inappropriate plans, tool misuse, or invalid execution steps [6, 7].

In this work, we present an automated scRNA-seq analysis framework that bridges LLM-based planning and workflow languages, combining the flexibility of LLMs with the rigor and reproducibility of formal workflows. This framework enables validated and scalable exploration of diverse data analysis strategies and ultimately supports integrated interpretation and report generation from large numbers of analysis outputs.

2. System Overview

The proposed analysis framework consists of five complementary components (Fig. 1), designed to coordinate flexible LLM-based planning with reproducible workflow execution.

(1) Modularization of analysis steps: Major scRNA-seq analysis steps—such as quality control, normalization, clustering, cell type annotation, and downstream analyses—are redefined as computational modules with explicit input and output specifications. These modules are provided as MCP (Model Context Protocol) server endpoints, forming a unified tool catalog that can be safely invoked by LLM agents.

(2) Analysis request generation: Given a dataset and a high-level analysis instruction, an LLM-based agent generates multiple concrete analysis requests (e.g., cell–cell interaction analysis, cell type annotation, or pseudotime inference). This design avoids committing to a single pipeline and explicitly promotes diversity in analysis entry points.

(3) Dynamic planning and validation: For each analysis request, a planner composes an analysis plan by selecting and ordering MCP modules. The generated plan is then validated in terms of input–output consistency, dependency correctness, and request satisfaction. Incomplete or inappropriate plans are filtered out before execution.

(4) Workflow execution: Only validated plans are compiled into workflows expressed in a workflow language (Snakemake) and executed with standardized logging. This ensures reproducibility and operational stability while enabling many analysis plans to be executed in parallel.

(5) Integrated interpretation and report generation: Finally, outputs from multiple workflows are aggregated and interpreted to identify consistent patterns and condition-dependent differences. The framework supports not only execution but also integrated interpretation and report generation across large numbers of analyses.

Together, these components enable end-to-end automation from analysis request generation to integrated interpretation and report generation.

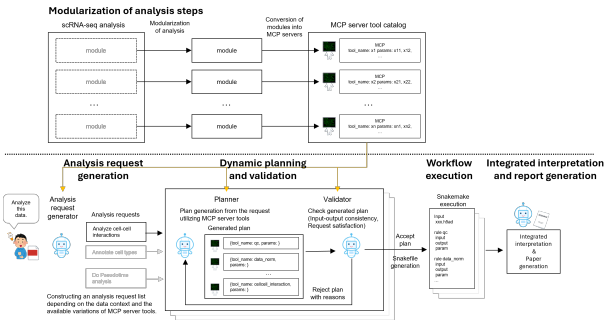

Fig. 1: Overview of the proposed scRNA-seq analysis framework. An LLM generates and validates analysis plans over modularized analysis steps exposed as MCP server components, which are compiled into reproducible workflows and executed in parallel for integrated interpretation and report generation.

3. Experimental Evaluation

We evaluated the framework using three scRNA-seq datasets (PBMC [8], Mouse Neuron [9], and Pancreas [10]) and three LLMs (gpt-4o, gpt-5, and gpt-5-

mini), with validation enabled or disabled. For each condition, ten analysis queries were issued, resulting in a total of 180 workflows. We evaluated Snakefile generation success, Snakemake run success, and query fulfillment, where query fulfillment measures whether the executed analysis plan actually satisfies the requested analysis objective.

Snakefile generation success rates and Snakemake run success rates were similar regardless of whether validation was enabled (Fig. 2). In contrast, a clear difference was observed in query fulfillment: the average fulfillment rate increased from 0.34 without validation to 0.78 with validation (Fig. 2). This result arose because, without validation, analysis plans that were technically executable but insufficient to satisfy the requested analysis were allowed to run. Consequently, high execution success does not necessarily guarantee analytical validity or request satisfaction. These findings demonstrate that explicit validation of analysis plans is essential.

Across all datasets and models, validation consistently improved query fulfillment (Fig. 2), demonstrating that explicit validation is essential for automated and scalable exploration of scRNA-seq analyses, as high execution success alone does not guarantee analytical validity or request satisfaction.

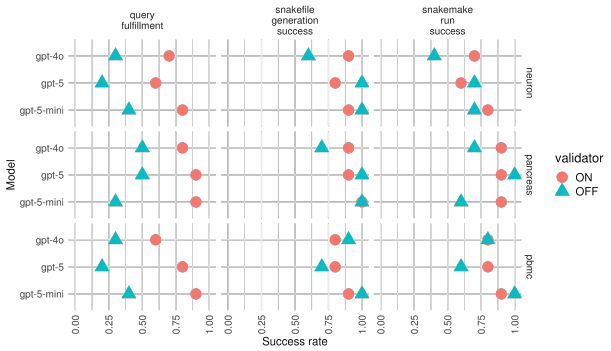

Fig. 2: Experimental results across three scRNA-seq datasets (PBMC, Neuron, and Pancreas) and three LLMs (gpt-4o, gpt-5, and gpt-5-mini). Each point represents the success rate over ten analysis queries for query fulfillment, Snakefile generation success, and Snakemake run success. Circles indicate validation enabled and triangles indicate validation disabled.

4. Conclusion

We presented a scRNA-seq analysis framework that bridges LLM-based planning and workflow languages, enabling automated, validated, and scalable exploration of diverse analysis strategies. Evaluation across 180 workflows showed that plan validation substantially improves query fulfillment, highlighting the importance of validation for executing analyses that satisfy analytical objectives rather than merely running successfully.

References

[1] Gunsagar S Gulati, Jeremy Philip D'Silva, Yunhe Liu, Linghua Wang, and Aaron M Newman. Profiling cell identity and tissue architecture with single-cell and spatial transcriptomics. *Nat. Rev. Mol. Cell Biol.*, 26(1):11–31, January 2025.

[2] Greg Gibson. Perspectives on rigor and reproducibility in single cell genomics. *PLoS Genet.*, 18(5):e1010210, May 2022.

[3] Luke Zappia and Fabian J Theis. Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape. *Genome Biol.*, 22(1):301, October 2021.

[4] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*, 35(4):316–319, April 2017.

[5] Johannes Köster and Sven Rahmann. Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, October 2012.

[6] Samuel Alber, Bowen Chen, Eric Sun, Alina Isakova, Aaron J Wilk, and James Zou. Cell voyager: AI comp bio agent generates new insights by autonomously analyzing biological data.

[7] Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, Di Yin, Shruti Marwaha, Jennefer N Carter, Xin Zhou, Matthew Wheeler, Jonathan A Bernstein, Mengdi Wang, Peng He, Jingtian Zhou, Michael Snyder, Le Cong, Aviv Regev, and Jure Leskovec. Biomni: A general-purpose biomedical AI agent. *bioRxivorg*, June 2025.

[8] 10x Genomics. 10k peripheral blood mononuclear cells (pbmcs) from a healthy donor, single indexed, universal 3' dataset. `https://www.10xgenomics.com/jp/datasets/10-k-peripheral-blood-mononuclear-cells-pbm-cs-from-a-healthy-donor-single-indexed-3-1-standard-4-0-0`, 2020. [Dataset].

[9] 10x Genomics. 10k mouse e18 combined cortex, hippocampus and subventricular zone cells, chromium gem-x single cell 3' dataset. `https://www.10xgenomics.com/datasets/10k-Mouse-Neurons-3p-gemx`, 2024. [Dataset].

[10] Maayan Baron, Adrian Veres, Samuel L Wolock, Aubrey L Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K Wagner, Shai S Shen-Orr, Allon M Klein, Douglas A Melton, and Itai Yanai. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst*, 3(4):346–360.e4, October 2016.