

Volume 41, Issue 8
August 2025

Article Contents

Abstract

1 Introduction

2 Methods

JOURNAL ARTICLE

SCassist: an AI based workflow assistant for single-cell analysis

FREE

Vijayaraj Nagarajan ✉, Guangpu Shi, Samyuktha Arunkumar, Chunhong Liu, Jaanam Gopalakrishnan, Pulak R Nath, Junseok Jang, Rachel R Caspi

Bioinformatics, Volume 41, Issue 8, August 2025, btaf402, <https://doi.org/10.1093/bioinformatics/btaf402>

Published: 12 July 2025 **Article history** ▼



PDF

Split View



Cite



Permissions



Share ▼

Vijay Nagarajan PhD
Laboratory of Immunology,
National Eye Institute,
Bethesda 20892, USA

Abstract

Summary

Single-cell RNA sequencing (scRNA-seq) data analysis often involves complex iterative workflow

The Agony of "Big Data" in Biology



- **Single-Cell Analysis:** Studying a forest by looking at every single leaf, individually. We do that with cells.
 - 5,000 cells x 30000 genes
 - Lot of leaves with lot of properties
- **Problem**
 - Which cells are good? (QC)
 - How to make them comparable? (Normalization)
 - How to group them? (Clustering)
 - How are those groups communicating ? (Interactions)
 - How are those groups fate decided ? (Trajectories)
 - How are those groups regulated ? (Multi-modal)
- **Human cost**
 - Too many parameters to choose objectively. Takes more than a PhD, a therapist, and a lot of coffee...

LLMs: Our New Overlords

Or just really good interns



- **They...**

- Write poems, generate code, order coffee...
- Anything they can't?

- **Biomedical applications**

- They can now infer
- Do we need a PhD anymore?

- **Hallucinations**

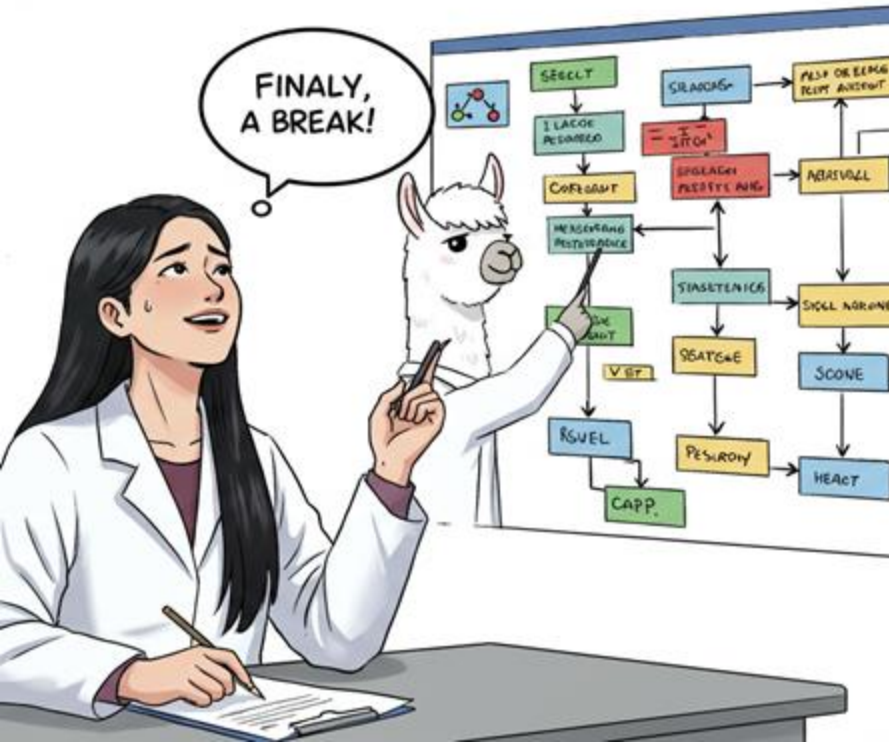
- Bug or a feature
- Care why they make up stuff?

- **RAG**

- Give them a cheat sheet
- So they can't make up too much information

SCassist: AI Workflow Assistant For Single-Cell

You didn't ask, but probably need



- **Novel**

- AI that holds your hand through the *entire* process, not just one tiny step

- **Empowers Users**

- At least gives you someone else to blame for the parameters

- **Familiar Environment**

- Simple R package, no fancy pytorch or langchain ?

- **Flexible, Cost-Effective & Confidential**

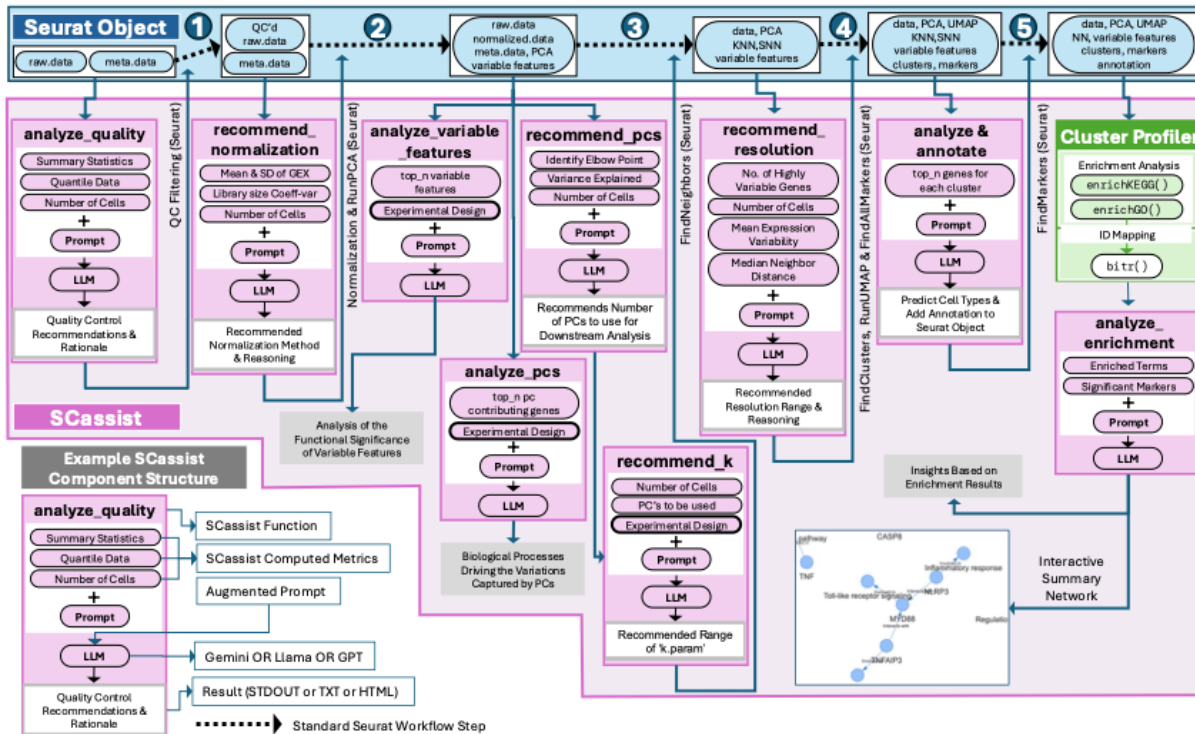
- Choose your AI poison: Google, OpenAI, or the free one that runs on your laptop (and probably makes it sound like a jet engine)

- **Open-Source**

- Because we believe in sharing the pain... I mean the progress..

SCassist Under the Hood

It's not magic, just a lot of "If" statements



• Key Logic

- Takes data, computes/extracts metrics/tools
- Creates augmented prompt (polite instructions)
- Prompt sent to the LLM (for thinking or pretending to be thinking)
- LLM responds, we parse it and show it to user

• LLM Choice

- Google (you need a really big brain)
- OpenAI (you don't like Google)
- Ollama (you want it free or you are paranoid)

SCassist Functions

Because you can't be expected to know everything

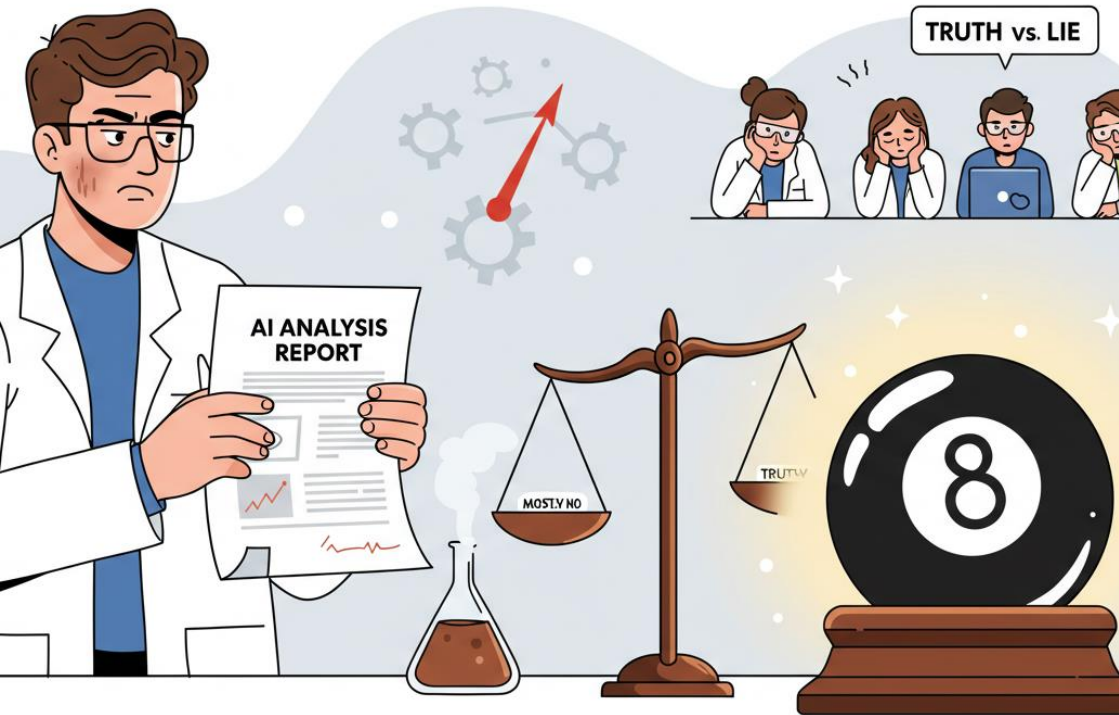


Key Functions

- `SCassist_analyze_quality()` : “is your data good, bad, or just really, really sad.”
- `SCassist_recommend_normalization()` : “suggests the best way to make your data comparable, so you don't have to pretend you understand the math.”
- `SCassist_analyze_variable_features()` : “explains who is the driver and what is being driven.”
- `SCassist_analyze_and_annotate()` : “predicts what your cells are, so you can finally thank the robot.”
- `SCassist_analyze_enrichment()` : “turns your messy biological results into something that sounds profound.”

Evaluation

Proof we didn't just build a fancy magic 8-Ball



- **Datasets**

- We re-analyzed our own published data (because if it can't handle *our* data, what hope is there?)

- **Groundedness Score**

- Did the AI make stuff up? We checked. Mostly not. (unlike some news sources.)

- **Contextual Relevance**

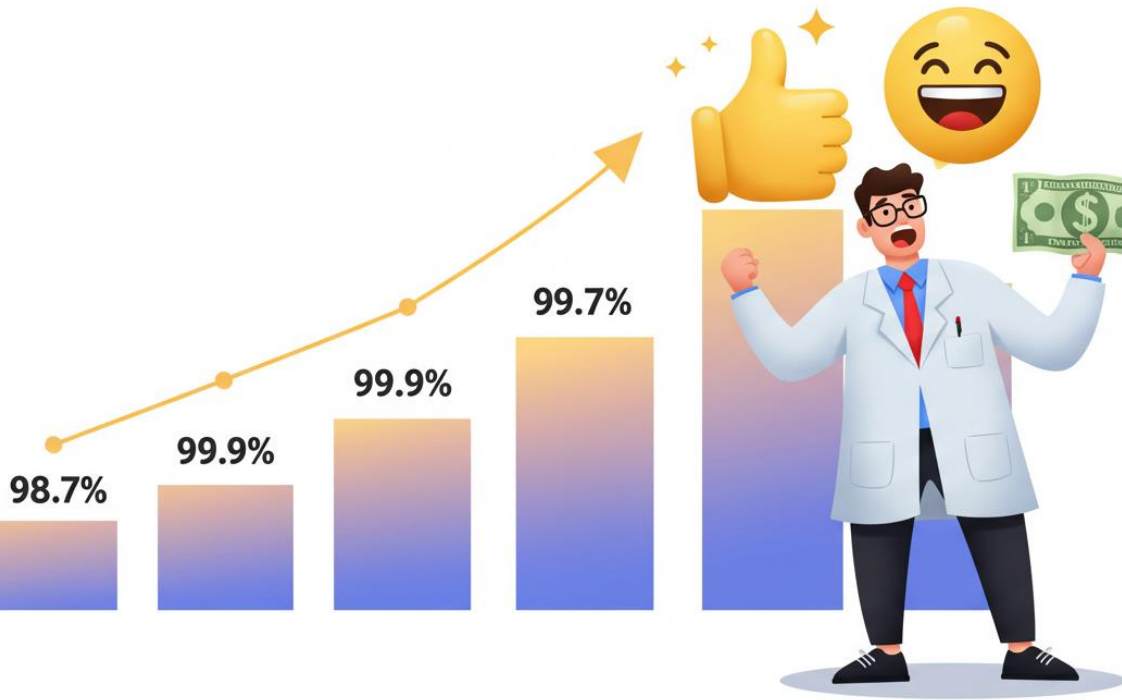
- Did the AI actually “understand” what it was talking about? We used a fancy AI called BERT to check (because apparently, BERT knows everything.)"

- **Human Evaluation**

- We subjected 8 poor souls (scientists!) to our reports. They rated us on 'Accuracy,' 'Relevance,' 'Clarity,' 'Trustworthiness,' and 'Overall Satisfaction.' (we bribed them with coffee, which didn't seem to have worked)

SCassist Performance

Surprisingly not terrible!



- **Groundedness**

- 98.7% and 99.9% groundedness. So almost no made-up facts!

- **Semantic Similarity**

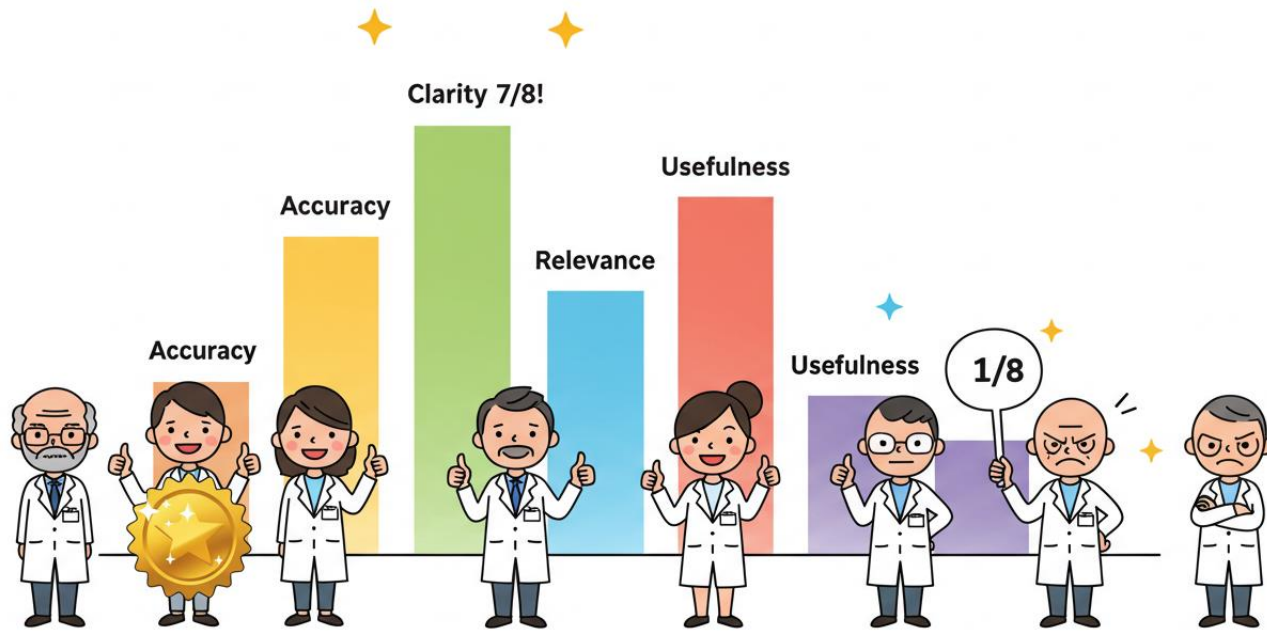
- 76% and 74% semantic similarity. It mostly understood the assignment.

- **Cost**

- A whopping \$2.07 for two months of testing (cheaper than a fancy coffee, more useful than a motivational poster.)

SCassist Performance

Humans were impressed (mostly)



- **Overall**

- Statistically significant 'yay!' from the evaluators. ($p=0.0001122$)

- **Clarity**

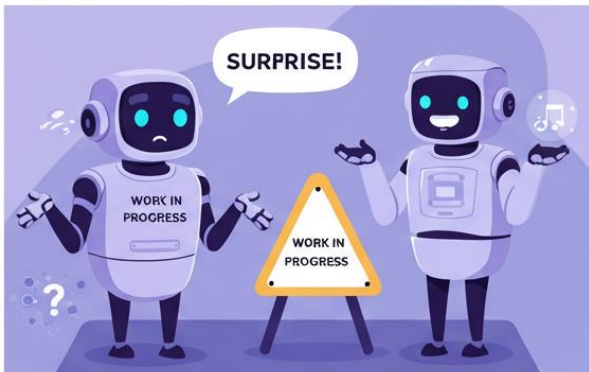
- SCassist is exceptionally clear. 7 out of 8 evaluators found it 'useful' (the 8th was probably just having a bad day).

- **Robustness**

- Works equally well for senior scientists who've seen it all, and junior scientists who are still terrified of complex software.
- Also works on different datasets. It's not picky.

Limitations & Future Directions

We're not perfect (Yet)



• Limitations

- AI models change. So, the future results might be *even better* (or slightly different).
- It's a guide, not a god (yet). You still have to think. Sorry.

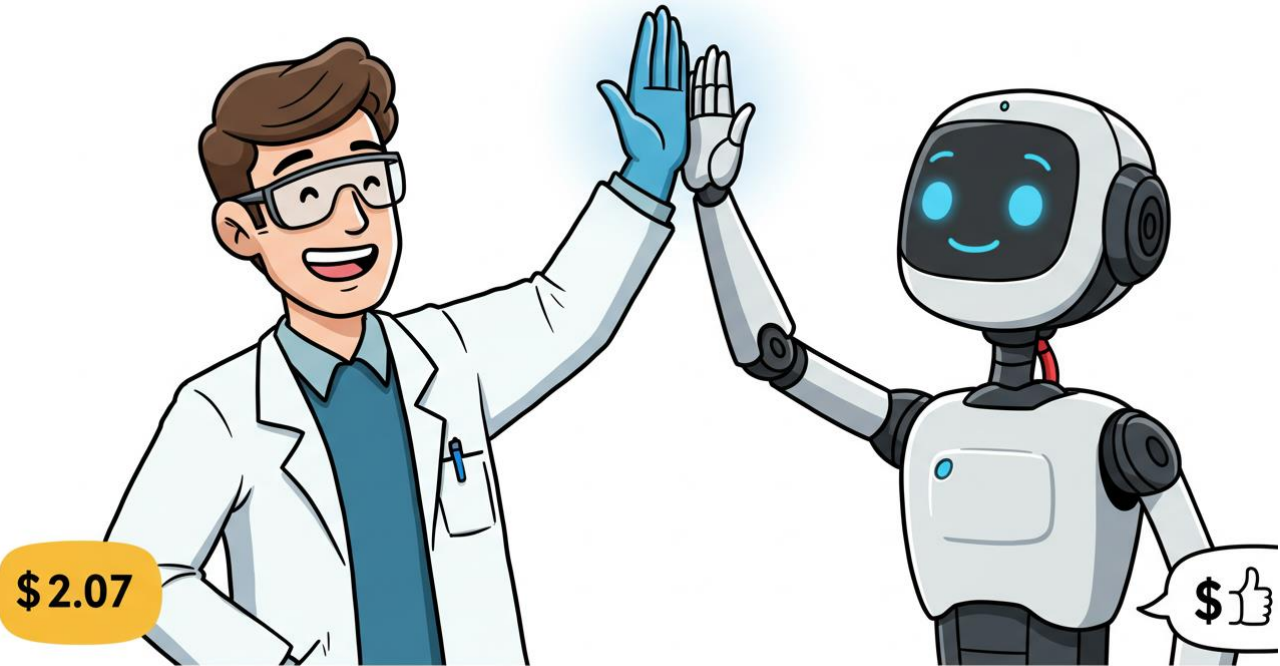


• Future Directions

- Interaction and trajectory: Because cells don't just sit there (they have feelings and travel plans)
- Multi-modal and spatial data (because one type of data isn't confusing enough)
- Automation: We're teaching the AI to click buttons for you (soon, it'll do your laundry)

Conclusion & Call to Action

Go forth and analyze!



• Key Takeaways

- An AI assistant to help you analyze single-cell data
 - It's accurate, it's relevant, it's clear, and humans mostly liked it.
 - Works for everyone, from the seasoned pro to the 'what's a computer?' beginner.
 - It's cheap! (unless you have a really expensive GPU for Ollama, then it's FREE)
 - Accelerates discovery, or at least, reduces your screen time staring blankly at plots.

• Call to Action

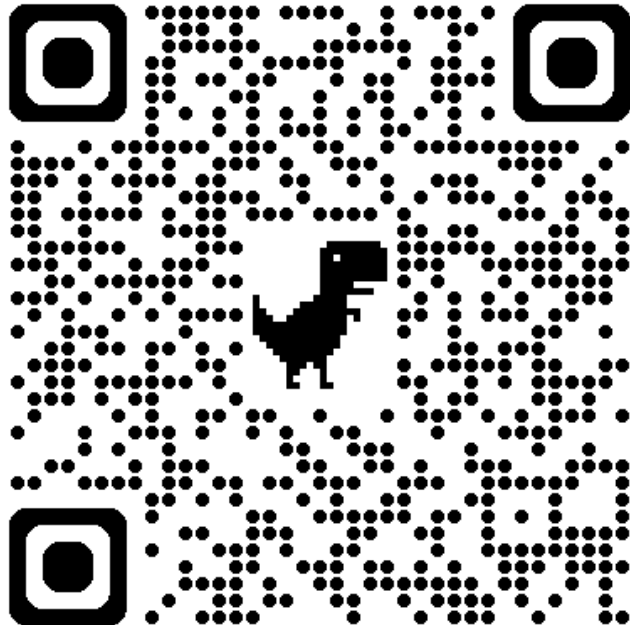
- **GitHub:** <https://github.com/NIH-NEI/SCassist>
 - (Go on, you know you want to give us a STAR.)
- **Installation:** It's an R package, check our GitHub for easy instructions.
- **Tutorials:** Yes, we even wrote instructions. Shocking, I know.

Thanks to:
Dr. Rachel Caspi
Dr. Charles Egwuagu
Dr. Han-Yu Shih

nagarajanv@nih.gov

V2 Update Preview

- CellChat Integration for Interaction Inference
 - Fork the current version, Star the repo and Email nagarajanv@nih.gov to receive the beta version of this update



IAN: Intelligent Analysis for Omics-Based Understanding of Biological Systems

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

[HOME](#) | [SUBMIT](#) | [FAQ](#) | [BLOG](#) |

New Results

[Follow this preprint](#) [Previous](#)

IAN: An Intelligent System for Omics Data Analysis and Discovery

Posted March 11, 2025.

[Vijayaraj Nagarajan, Guangpu Shi, Reiko Horai, Cheng-Rong Yu, Jaanam Gopalakrishnan, Manoj Yadav, Michael H Liew, Calla Gentilucci, Rachel R Caspi](#)

doi: <https://doi.org/10.1101/2025.03.06.640921>

[Download PDF](#)

[Print/Save Options](#)

[Data/Code](#)

