scChat: A Large Language Model-Powered Co-Pilot for Contextualized Single-Cell RNA Sequencing Analysis

Yen-Chun Lu[1,2,*], Ashley Varghese[1,*], Rahul Nahar[1], Hao Chen[1], Kunming Shao[2,3], Xiaoping Bao[1,2,#], Can Li[1,#]

1. Davidson School of Chemical Engineering, Purdue University, 480 W. Stadium Ave, West Lafayette, IN 47907, USA
2. Purdue Institute for Cancer Research, Purdue University, 201 S. University St, West Lafayette, IN 47907, USA
3. Department of Biological Sciences, Purdue University, 915 Mitch Daniels Boulevard, West Lafayette, IN 47907, USA

[*]These authors contributed equally

[#]Corresponding authors (X. B., bao61@purdue.edu; C.L., canli@purdue.edu)

Abstract

Data-driven methods, including unsupervised and supervised learning, have become essential tools for analyzing single-cell RNA sequencing (scRNA-seq) data. However, these approaches cannot incorporate research context into the analysis, limiting their ability to provide in-depth, contextual insights to researchers. We introduce scChat, a pioneering AI assistant designed for contextualized scRNA-seq analysis. Unlike existing tools that focus on standard tasks like cell annotation, scChat offers advanced capabilities, including research context-based analysis of experimental outcomes, detailed validation of research hypotheses, identification of potential explanations when hypotheses fail, and suggestions for next-step experimental design. Powered by a large language model (LLM), scChat integrates function calls for quantitative tasks, employs retrieval-augmented generation to reduce hallucination, and uses an LLM-powered search engine to seamlessly incorporate current literature into the analytical conversation.

Introduction

Single-cell RNA sequencing (scRNA-seq) has revolutionized biomedical research by providing a detailed and high-resolution view of the transcriptomic landscape within individual cells[1]. This technology enables researchers to dissect tissue heterogeneity, uncover distinct cell populations, and trace developmental or disease trajectories with unparalleled cellular resolution. Various machine learning algorithms have been developed to analyze scRNA-seq data, including unsupervised clustering methods implemented in software like Seurat[2] and Scanpy[3], as well as supervised[4–9] and self-supervised learning algorithms[10–13]. Deep learning methods, particularly those employing transformer-based architectures like scBERT[10] , scFoundation[13], and scGPT[11], have achieved impressive accuracy in tasks such as cell type annotation. However, a significant limitation of these deep learning approaches is their lack of interpretability, which hinders their usability for biomedical researchers without a computer science background. This barrier restricts interactive tasks such as obtaining biological insights from cell classifications and analyzing cell populations. Furthermore, the purely data-driven nature of ML analysis in scRNA-seq often fails to "contextualize" the specific biomedical problems under investigation, thereby limiting its ability to directly inform experimental design and hypothesis generation.

Large Language Models (LLMs), such as OpenAI's GPT-4[14] (the LLM powering ChatGPT) and Meta's Llama models[15], are trained on extensive datasets covering a broad spectrum of human knowledge. These pretrained LLMs can execute tasks such as code generation, exam-taking, and essay writing at an unprecedented level. To harness the potential of LLMs, various "copilot" tools (AI assistants) have been developed based on these models. For instance, LLaVA-Med[16], BiomedGPT[17], Med-Gemini[18] are chatbots to help doctors analyze medical images and generate diagnosis reports. Despite these commercial successes, AI assistants specifically designed for scientific research and discovery remain relatively scarce. Coscientist[19], a recent AI copilot for chemical research powered by the OpenAI GPT-4 model, can autonomously design, plan and perform complex experiments. Recently, GPT-4 model was empirically tested to annotate cell types and was shown to perform close to human experts[20]. However, to the best of our knowledge, the potential of LLMs for generating more biological insights from the data and answering scientific questions has not been investigated. The ability of LLMs to process vast quantities of information and identify underlying patterns makes them particularly valuable for scientific applications. However, in biomedical sciences, the potential of LLMs is still largely untapped. A significant barrier to their application in scientific research is the phenomenon of "hallucination"[21], where an LLM might present information with unwarranted confidence, even when uncertain. Moreover, pretrained models like GPT-4 excel at qualitative tasks such as explaining technical concepts but struggle with quantitative tasks[22]. As a result, these models cannot be directly applied to analyze scRNA-seq data matrices for tasks retrieving detailed statistics.

To this end, we propose scChat, a platform that combines quantitative statistical learning algorithms, LLMs, and research context to offer contextualized scRNA-seq data analysis capabilities. scChat serves as a co-pilot for scientists, enabling natural language interaction through a graphical user interface (GUI). Beyond standard functionalities like data processing, cell type annotation, subpopulation identification, gene enrichment analysis, and statistical sample comparison, scChat leverages LLMs to provide contextualized insights. These include validating research hypotheses, offering explanations for unexpected experimental outcomes, and suggesting next steps in experimental design, such as treatment strategies for patients. The platform achieves these advanced capabilities by integrating LLMs with function calls, retrieval-augmented generation (RAG), and a powerful LLM-driven search engine, as discussed in the Methods section.

scChat Architecture Overview

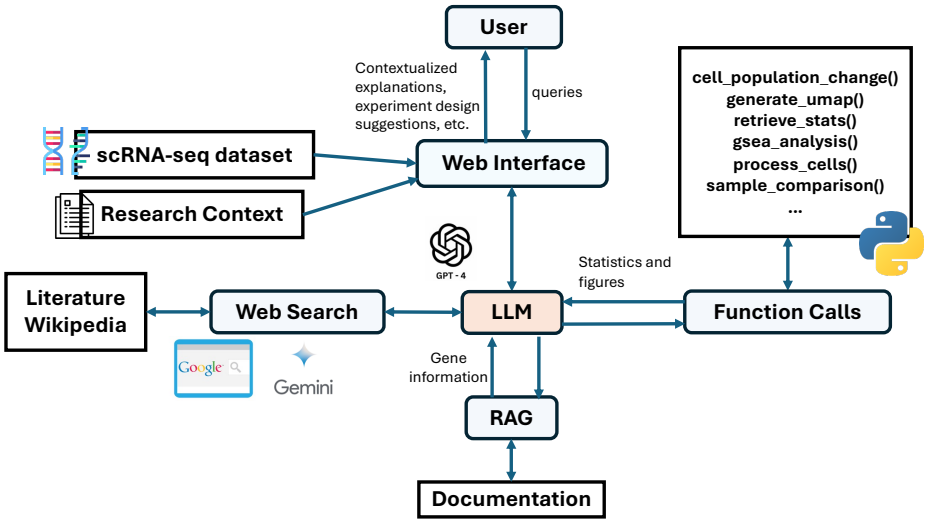

Figure 1 Overview of scChat Architecture

An overview of scChat is shown in **Figure 1**. The inputs to scChat are the raw scRNA-seq data and a research context provided by the user which describes how the data is obtained and the research question/hypothesis to be studied/tested. User interactions are facilitated through a graphical user interface (GUI). The LLM is the "brain" of scChat that orchestrates different tasks. To enable in-depth analysis and minimize hallucinations, the LLM utilizes three main capabilities: automated function calls specifically designed for scRNA-seq analysis, web search powered by Google Gemini, and Retrieval-Augmented Generation (RAG) techniques[23]. The LLM autonomously executes function calls when it identifies that user queries can be more effectively addressed through specialized analyses. These functions encompass a range of tasks, including assessing cell population changes across samples, generating UMAP clusters for annotation, conducting gene set enrichment analysis (GSEA), processing and annotating subpopulations of cells, retrieving relevant statistical data, and comparing different cell samples, such as pre-and post-treatment in cancer studies. The web search functionality is employed to conduct highly contextualized searches, mitigating the risk of hallucinations by preventing the generation of false literature references and supplementing the LLM's inherent knowledge base. In addition, the system incorporates a RAG module to enhance the contextual relevance of the results by augmenting the LLM's generated responses with documentation of all processes, including URLs or references to marker genes used in the analysis.

Results

To demonstrate the capabilities of scChat as a co-pilot for biomedical researchers, we test scChat using two scRNA-seq datasets obtained from two research papers[24,25]. In what follows, we will use the two datasets to showcase how scChat can provide in-depth analyses of the scRNA-seq data and compare the results with the analysis by experienced biomedical researchers in the papers. The detailed chat histories for both papers are shown in the supplementary material where standard quantitative tasks such as cell population clustering and annotation, subpopulation annotation, and differential expression analysis are demonstrated as interactive conversations.

[Evaluation of scChat using the dataset from Bagley *et al.* [24]]

The paper by Bagley *et al.* [24] hypothesized that combining anti-EGFRvIII CAR T cells and anti-PD1 monoclonal antibody pembrolizumab can provide a better treatment for adult, newly diagnosed glioblastoma (GBM) patients. To test the hypothesis, the authors performed scRNA-seq on tumor tissues of three glioblastoma patients' before and after the proposed treatment. When prompted by the user to evaluate the treatment results based on scRNA-seq data and the research context, scChat is able to annotate the cell types and the subpopulations of T-cells using the function calls, which are consistent with the expert annotations by Bagley *et al.* [24] scChat also provides explanations for the annotation results by using the marker genes prepared in a JSON file, which reduces hallucination by providing gene names and URLs for the user's reference through the retrieval augmented generation (RAG) technique. A sample annotation and explanation are listed as follows.

```
Cluster 0: Regulatory T Cells (Tregs).
    Markers: High expression of FOXP3, CTLA4, IL2RA (CD25).
    Rationale: Tregs are characterized by the expression of FOXP3, which is a key transcription factor for their
        development  and function. CTLA4 and IL2RA are also commonly expressed on Tregs and are involved in their
        immunosuppressive functions.
    Links: FOXP3, CTLA4, IL2RA
```

Figure 2 Sample output from scChat for annotating a cell cluster.

Besides doing simple tasks like cell annotation and calculating the cell population changes pre- and post-treatment, scChat can provide more in-depth analysis.

scChat evaluates the T cell exhaustion

Explanations from the paper

- Presence of effector exhausted CD8 T cells after treatment
- Elevated expression of exhaustion markers such as TOX, PDCD1 (PD-1), and EOMES following CAR T-cell infusion
- Increased expression of checkpoint molecules (PD-1, CTLA-4, TIM-3, LAG-3) in post-treatment T cells
- Higher proportions of regulatory T cells (Tregs) observed post-treatment

Explanations from *scChat*

- Upregulation of exhaustion markers, such as PDCD1, TOX, CTLA4, LAG3, and TIGIT post-treatment, indicating exhaustion in most but not all patients
- Increased expression of checkpoint molecules, including PDCD1, CTLA4, and TIGIT, is observed, though some patients show reduced exhaustion, indicating variability
- Increase in regulatory T cells (Tregs) post-treatment, contributing to an immunosuppressive environment, though some patients show a decrease in Tregs
- Despite an increase in effector T cell populations, exhaustion persists in most cases, though some patients exhibit a partial reduction in exhaustion markers

Figure 3 Comparison of explanations from the paper and scChat when prompted with the question "Do the patients show signs of T cell exhaustion?".

Both the paper and scChat analysis emphasize the persistent challenge of T cell exhaustion in the GBM tumor microenvironment (TME) of the GBM patients following CAR T-cell therapy combined with pembrolizumab. The paper highlights the presence of effector-exhausted $CD8^+$ T cells post-treatment, marked by elevated expression of exhaustion markers such as *TOX*, *PDCD1* (*PD-1*), and *EOMES*. Similarly, scChat identifies the upregulation of these markers, particularly prominent in Patient 1, whose T cells display substantial post-treatment exhaustion. In addition, both the paper and scChat report the increased expression of checkpoint molecules, though the paper focuses on *PD-1*, *CTLA-4*, *TIM-3*, and *LAG-3*, while scChat emphasizes *PDCD1*, *CTLA4*, and *TIGIT*, highlighting the continued immune suppression observed after treatment.

For Patient 6, scChat offers a more nuanced view, noting a reduction in regulatory T cells and an increase in effector $CD8^+$ T cells, suggesting a partial alleviation of exhaustion. However, persistent elevation of *NR4A1* indicates that, despite some relief, a stress response remains[26,27]. This aligns with the paper's findings, which reflect significant inter-patient heterogeneity in the levels of exhaustion and immune activation following CAR T-cell therapy. Patient 7 also presents a complex immune response, where scChat reports a downregulation of *HAVCR2* (*TIM-3*), suggesting reduced exhaustion in some T cells, while *TOX* and *EOMES* remain elevated, indicating continued exhaustion in other subsets. The paper similarly underscores this complexity, reporting that T cell exhaustion remains pervasive across various immune compartments, with some variation between patients.

Both the paper and scChat further highlight the immunosuppressive role of regulatory T cells (Tregs). The paper reports an increase in the proportion of immunosuppressive Tregs post-treatment, a finding that is also observed in scChat. This rise in Tregs likely exacerbates immune suppression, counteracting the expansion of effector T cells. Despite some signs of immune activation, both sources underline the persistence of T cell exhaustion and regulatory T cell-mediated immunosuppression, which continue to hinder effective anti-tumor responses.

In summary, while the paper and scChat present some slight differences in patient-specific responses, both converge on a critical point: persistent T cell exhaustion remains a major barrier post-treatment. These findings suggest that, although immunotherapy strategies demonstrate biological activity, they are

insufficient to fully overcome the immunosuppressive environment in GBM. Both analyses emphasize the need for innovative therapeutic combinations to effectively target T cell exhaustion and improve clinical outcomes in patients with newly diagnosed GBM.

scChat provides explanations for the failure of the treatment

The discussion section of the paper focused on providing explanations for the failure of the treatment, which will inform future work and research hypothesis generation. We prompt scChat to provide explanations and compare them with those discussed in the paper. A summary of the comparison is shown in **Figure 4**.

Explanations from the paper

- Concomitant administration of a PD-1 inhibitor
- Tumor heterogeneity
- Additional immunosuppressive elements of the TME
- A lack of lymphodepleting chemotherapy
- Too little residual target tumor following surgery

Explanations from *scChat*

- Checkpoint molecules and exhaustion markers (PD-1, PD-L1)
- Spatial distribution and heterogeneity
- Immunosuppressive myeloid cells (MDSCs and TAMs)
- Inflammatory and pro-tumor cytokines
- Antigen presentation and immune evasion

Sample detailed explanations from *scChat*

Checkpoint Molecules and Exhaustion Markers

- Markers: PD-L1 (CD274), PD-1 (PDCD1), CTLA-4, LAG-3, TIM-3
- Implication: High expression of checkpoint molecules like PD-L1 on myeloid cells can induce T cell exhaustion, reducing the efficacy of CAR T cells. This immune checkpoint inhibition may limit T cell activation, allowing the tumor to persist despite therapy.
    - Literature: Studies show that immune checkpoints such as PD-1 and CTLA-4 are key inhibitors of T cell function in the TME. Overexpression of these molecules often leads to T cell exhaustion and failure of immune-mediated tumor clearance.

Figure 4 Comparison of the explanations from the paper and scChat to explain the treatment failure followed by a sample detailed explanation from scChat.

The first two explanations, tumor heterogeneity and immune checkpoint inhibitors, align closely with those discussed in the original study. Both the original paper and scChat identified immunosuppressive components within the TME, including immunosuppressive myeloid cells. However, scChat did not highlight the immunosuppressive Treg cells that were emphasized in the original study. Furthermore, the original paper identified two clinical factors potentially contributing to the limited efficacy of CAR-T cells: the absence of lymphodepleting chemotherapy and insufficient residual target tumor post-surgery. scChat was not able to refer to the clinical context in its explanation. In contrast, scChat offered two reasonable alternative explanations based on scRNA-seq data analysis: reducing inflammatory and pro-tumor cytokines, and enhancing antigen presentation to counter immune evasion. A sample detailed explanation of scChat is shown in which provides the rationale for the explanation based on marker genes and literature.

[Evaluation of scChat using the dataset from Mathewson et al. [25]]

Mathewson et al. [25] utilized scRNA-seq to examine tumor-infiltrating T cells' gene expression and clonal landscape across 31 patients with gliomas, including isocitrate dehydrogenase (IDH) wild-type GBM and IDH mutant gliomas. The researchers identified potential effectors of anti-tumor immunity, including T cells co-expressing cytotoxic programs and natural killer (NK) cell genes. Notably, they highlighted the NK gene *KLRB1*, which encodes the CD161 receptor, as an inhibitory receptor that could be targeted to enhance T cell-mediated killing of glioma cells. The study provides an atlas of T cells in gliomas and highlights the importance of CD161 as a potential target for immunotherapy. We incorporate this dataset into our analysis using scChat to further validate our platform's ability to identify and characterize key immune cell populations within the TME, while also deriving biological insights within its research context.

scChat evaluates T cell variability across different samples

Explanations from the paper

- Diverse T cell subsets, including $CD8^+$ T cells, $CD4^+$ T cells, and regulatory T cells (Tregs) are identified in glioma samples
- The overall proportions of T cell populations are largely consistent between GBM and IDH-mutant glioma, with some variation in specific T cell states
- $CD8^+$ T cells, expressing cytotoxic genes such as PRF1, GZMB, and GNLY, show variability across samples but remain crucial to anti-tumor immune responses
- Tregs are observed across patient samples, contributing to an immunosuppressive tumor microenvironment

Explanations from *scChat*

- Significant variability in $CD8^+$ T cell counts, ranging from 0.33% to 68.6% $CD45^+$ cells in different patient samples
- High NK/T cell populations in specific samples, such as E61 and MGH206
- Treg expansion is evident in certain patients, with notable increases in E36 (208) and E90 (203), indicating enhanced immunosuppression
- Notable variation in naive T cell counts, with samples such as E42 and MGH225 showing high counts

Figure 5 Comparison of explanations from the paper and scChat when prompted with the question "How do T cell population compositions in the tumor microenvironment differ across glioma samples?".

scChat provides a more detailed view of T cell population variability across patient samples compared to the relatively consistent proportions reported in the paper. While the paper emphasizes a uniformity in T cell distributions between glioblastoma (GBM) and IDH-mutant glioma, scChat identifies substantial fluctuations in $CD8^+$ T cells, NK/T cells, and Tregs, suggesting a more nuanced and individualized immune landscape within gliomas.

In particular, scChat uncovers elevated Treg populations in specific patients, highlighting stronger immunosuppressive environments, which may have significant implications for future immunotherapy strategies. Furthermore, the variability in NK/T cell populations, not as prominently discussed in the paper, points to additional layers of immune complexity within the tumor microenvironment. These findings suggest that scChat offers a personalized perspective, revealing immune heterogeneity that could inform more targeted therapeutic approaches.

scChat evaluates the relationship between cytotoxicity and NK cell-like signature

Explanations from the paper

- Identification of NK cell-like signatures: High cytotoxic $CD8^+$ T cells show increased expression of NK cell-associated genes
- Functional implications: Co-expression of NK cell receptors in cytotoxic $CD8^+$ T cells suggests their potential role in anti-tumor activity
- Lower inhibitory receptor expression: Cytotoxic $CD8^+$ T cells with high NK signatures express lower levels of inhibitory molecules like PDCD1 and CTLA4
- Therapeutic insights: Targeting these NK-like T cell populations could enhance anti-tumor immunity

Explanations from *scChat*

- Analysis of $CD8^+$ T cell cytotoxicity across samples: scChat identified clusters of $CD8^+$ T cells with high cytotoxicity signatures
- Correlations between NK and cytotoxic signatures: scChat demonstrated a significant correlation between NK cell and cytotoxic gene expression scores
- Identification of NK-like $CD8^+$ T cells: scChat identified $CD8^+$ T cells co-expressing NK cell receptors, such as KLRC2, KLRD1, and KLRB1
- Functional differences: scChat suggested that NK-like $CD8^+$ T cells contribute to effective anti-tumor responses due to lower inhibitory markers expression, such as PDCD1 and CTLA4

Figure 6 Comparison of explanations from the paper and scChat when prompted with the question "Based on the analysis of glioma-infiltrating CD8+ T cells from 26 patients, how do cytotoxicity and NK cell-like signatures correlate?".

To further provide therapeutic intuition from this scRNA-seq dataset, both scChat analysis and the published study provide consistent and complementary insights into the relationship between cytotoxicity and NK cell-like signatures in glioma-infiltrating $CD8^+$ T cells. Both analyses highlight a strong association between high cytotoxicity scores and the expression of NK cell markers, suggesting that these T cells may acquire functional traits of NK cells to enhance their anti-tumor activity. The elevated expression of cytotoxic genes such as *PRF1* and *GZMB*, alongside NK receptors including *KLRD1 (CD94)* and *KLRC2 (NKG2C)*, indicates that these cells are well-equipped to target and eliminate tumor cells. This dual functionality could provide them with a competitive advantage in the tumor microenvironment, where traditional $CD8^+$ T cell responses may be dampened.

The scChat analysis supports these findings by confirming the correlation across samples, reinforcing the idea that this NK-like signature is a consistent feature of cytotoxic $CD8^+$ T cells in gliomas. These insights suggest that therapeutically targeting these specific T cell subpopulations, by enhancing their cytotoxic capacity or modulating inhibitory receptors such as *PD-1*, could improve immune responses in glioma patients. The potential to harness this unique subset of T cells offers promising new directions for immunotherapy strategies aimed at overcoming the immune resistance often observed in gliomas.

Conclusions

Compared with traditional tools for scRNA-seq analysis, scChat can leverage the biological knowledge of the LLM to provide in-depth analysis contextualized by the research problem. More importantly, scChat informs subsequent research and experimental design. Several methods including function calls, RAG, and web search were used to enable the LLM to perform quantitative analysis and reduce hallucination.

Methods

[scRNA-seq Dataset]

The scRNA-seq dataset is a primary input to the scChat platform, which accepts data in .h5ad format, a widely used file type for storing single-cell RNA sequencing data. When multiple scRNA-seq samples are collected from different experiments, such as treatment and control groups, scChat allows users to upload an accompanying JSON file. This file labels the data by "sample_id" enabling the platform to accurately distinguish between samples and perform comparative analyses. By leveraging these labels, scChat can analyze changes in cell populations and gene expression levels across different experimental conditions, providing valuable insights into the biological variations between groups.

[Research Context]

In the scChat platform, the research context encompasses the background knowledge, hypotheses, and experimental design that researchers bring to their studies. The hypotheses can be anticipated differences in gene expression between different control groups or the identification of distinct cell types within a population. The experimental design includes sample conditions for treatment and control groups. By integrating this contextual information, scChat tailors the analysis to the unique needs of each study, enhancing its ability to generate insights that are directly applicable and meaningful.

[Web Interface]

The web interface is the central hub for user interaction within the system, providing a seamless and intuitive platform for conducting complex scientific inquiries and analyses. It allows users to input queries and upload various data types for comprehensive analysis. Integrated data visualization tools enable users

to view graphs and charts directly within the interface, enhancing the interpretation of complex datasets. Additionally, export functionalities facilitate smooth transitions to external software, supporting further analysis and study.

[Large Language Model (LLM)]

The Large Language Model (LLM), GPT-4o, is integral to the scChat platform, orchestrating the execution of function calls and Retrieval-Augmented Generation (RAG) processes. We invested substantial effort in fine-tuning critical hyperparameters of GPT-4o, including temperature, Top-p (Nucleus Sampling), and Frequency Penalty, to optimize its performance for scRNA-seq data analysis. These adjustments were meticulously calibrated to strike a balance between generating creative, contextually relevant responses and maintaining the precision necessary for scientific rigor. This optimization ensures that scChat meets the stringent standards required for advanced research and analysis in the scRNA-seq field. To enhance contextualized analysis, all statistics generated by function calls, along with the research context, are saved as part of the LLM's context. This allows GPT-4o to access relevant information when responding to follow-up questions, ensuring that its analyses are grounded in the actual data rather than speculative outputs. The embedded statistics serve to anchor the LLM's responses, minimizing the risk of hallucination and enhancing the reliability of the generated insights.

[Web search]

scChat integrates a web search feature powered by Gemini, tailored to the research context of scRNA-seq analysis. This feature allows users to access the latest scientific literature, tools, and resources directly within the platform. By querying Gemini, researchers can retrieve top scholarly articles, databases, and critical information relevant to their analyses. This integration is not just an add-on but a core component of the research process. It provides immediate access to up-to-date external knowledge, enhancing the analysis by grounding it in current scientific discourse and reducing the risk of LLM hallucinations.

[scRNA-seq data preprocessing and function calls]

LLMs such as GPT-4o, while powerful tools for generating insights and providing contextual understanding are not designed to perform direct quantitative data analysis. This limitation arises because LLMs, although sophisticated, may hallucinate or produce outputs that are not strictly grounded in the underlying data. To address this challenge, scChat uses a series of predefined function calls, powered by the open-source package Scanpy[3] to construct the function calls necessary for analyzing raw scRNA-seq data. When users upload their raw scRNA-seq files onto the platform, scChat automatically initiates a series of preprocessing steps. The platform performs quality control to filter low-quality cells and genes, batch correction to reduce technical variations, and normalization to standardize gene expression across cells. It then selects highly variable genes, enhancing the ability to distinguish between different cell types or states.

Once these preprocessing steps are completed, users can interact with the Large Language Model (LLM) – GPT-4o to initiate downstream analysis. The downstream analysis in scChat is powered by a series of well-defined function calls that will be automatically executed based on the user's queries and the context. The "generate_umap" function, which creates unsupervised clustering 2D UMAP visualizations to understand the structure and relationships between different cell populations. The "retrieve_stats" function allows for detailed annotation by retrieving mean expression values and expression proportions across different genes. Users can then visualize annotated clustering and gene expression patterns with "display_umap" and "display_dotplot", or track changes in cell populations over time or between conditions using the "calculate_cell_population_change" function. For targeted analysis, specific functions, such as "process_cells(cell_tyes)" and "display_annotated_umap(cell_types)", can be tailored for different cell types. The "process_cells" function filters out specific cell types and then re-applies the unsupervised

clustering UMAP, allowing for a focused analysis of selected cell populations. The "display_annotated_umap(cell_types)" function then visualizes the UMAP, displaying the annotated results after the unsupervised clustering has been performed. Additionally, the "differential_expression_genes_comparison" function facilitates differential gene expression comparisons between different patient conditions. Lastly, the "gsea_analysis" function identifies significantly enriched pathways or processes within the dataset, providing a comprehensive understanding of the molecular mechanisms at play. Each of these functions is carefully designed to allow researchers to perform complex analyses and gain meaningful insights from their scRNA-seq data. By providing streamlined and efficient commands, these functions enable researchers to focus on their scientific objectives without getting bogged down in the underlying code.

[Retrieval-Augmented Generation (RAG)]

Ensuring accuracy and relevance in LLMs, especially in specialized domains such as scRNA-seq analysis, is challenging. While LLMs generate text based on trained patterns, they can produce incorrect or shallow responses for complex tasks. Retrieval-augmented generation (RAG) addresses this by integrating external, domain-specific knowledge, allowing the model to retrieve and reference relevant documents or databases.

In scChat, RAG enhances LLMs by integrating a curated set of marker gene documents, designed to assist in accurately annotating cell types and states by providing detailed descriptions of marker genes crucial for interpreting complex data. Each document includes a comprehensive gene description, outlining its biological significance, functional roles, and involvement in cellular processes, essential for understanding the context in which these genes operate. To ensure data validity and credibility, each gene is linked to an external reference URL, directing users to authoritative sources such as the National Center for Biotechnology Information (NCBI) Gene database. This allows users to verify information, ensuring that the LLM's annotations are grounded in reliable scientific knowledge, thereby preventing hallucinations. During analysis, the RAG system dynamically integrates this information into the LLM's responses.

Data availability

scChat is provided as an open-source software package at the GitHub repository https://github.com/li-group/scChat Several scRNA-seq samples and their research contexts are provided in the GitHub repository. A video demo of scChat can be found at https://youtu.be/4LDdncq-sp8

Acknowledgements. We thank Bao and Li laboratory members for their technical assistance. This study was supported by startup funding from the Davidson School of Chemical Engineering and the College of Engineering at Purdue (X.B., C.L.), NIH NCI (grant no. R37CA265926 to X.B.), NSF (grant no. 2143064 and 2424004 to X.B. and C.L.), and Ralph W. and Grace M. Showalter Research Trust Grant (C.L.). The authors also gratefully acknowledge support from the Purdue University Institute for Cancer Research, P30CA023168, Purdue Institute for Integrative Neuroscience (PIIN), Bindley Biosciences Center, and the Walther Cancer Foundation.

References

(1) Ziegenhain, C.; Vieth, B.; Parekh, S.; Reinius, B.; Guillaumet-Adkins, A.; Smets, M.; Leonhardt, H.; Heyn, H.; Hellmann, I.; Enard, W. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol Cell* **2017**, *65* (4), 631-643.e4. https://doi.org/10.1016/j.molcel.2017.01.023.

(2) Hao, Y.; Stuart, T.; Kowalski, M. H.; Choudhary, S.; Hoffman, P.; Hartman, A.; Srivastava, A.; Molla, G.; Madad, S.; Fernandez-Granda, C.; Satija, R. Dictionary Learning for Integrative,

Multimodal and Scalable Single-Cell Analysis. *Nat Biotechnol* **2024**, *42* (2), 293–304. https://doi.org/10.1038/s41587-023-01767-y.
(3) Wolf, F. A.; Angerer, P.; Theis, F. J. SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis. *Genome Biol* **2018**, *19* (1), 15. https://doi.org/10.1186/s13059-017-1382-0.
(4) Lin, Y.; Cao, Y.; Kim, H. J.; Salim, A.; Speed, T. P.; Lin, D.; Yee Hwa Yang, J. ScClassify: Hierarchical Classification of Cells. https://doi.org/10.1101/776948.
(5) Pliner, H. A.; Shendure, J.; Trapnell, C. Supervised Classification Enables Rapid Annotation of Cell Atlases. *Nat Methods* **2019**, *16* (10), 983–986. https://doi.org/10.1038/s41592-019-0535-3.
(6) Kong, W.; Fu, Y. C.; Holloway, E. M.; Garipler, G.; Yang, X.; Mazzoni, E. O.; Morris, S. A. Capybara: A Computational Tool to Measure Cell Identity and Fate Transitions. *Cell Stem Cell* **2022**, *29* (4), 635-649.e11. https://doi.org/10.1016/j.stem.2022.03.001.
(7) Alquicira-Hernandez, J.; Sathe, A.; Ji, H. P.; Nguyen, Q.; Powell, J. E. ScPred: Accurate Supervised Method for Cell-Type Classification from Single-Cell RNA-Seq Data. *Genome Biol* **2019**, *20* (1). https://doi.org/10.1186/s13059-019-1862-5.
(8) Tan, Y.; Cahan, P. SingleCellNet: A Computational Tool to Classify Single Cell RNA-Seq Data Across Platforms and Across Species. *Cell Syst* **2019**, *9* (2), 207-213.e2. https://doi.org/10.1016/j.cels.2019.06.004.
(9) Wang, J.; Ma, A.; Chang, Y.; Gong, J.; Jiang, Y.; Qi, R.; Wang, C.; Fu, H.; Ma, Q.; Xu, D. ScGNN Is a Novel Graph Neural Network Framework for Single-Cell RNA-Seq Analyses. *Nat Commun* **2021**, *12* (1), 1882. https://doi.org/10.1038/s41467-021-22197-x.
(10) Yang, F.; Wang, W.; Wang, F.; Fang, Y.; Tang, D.; Huang, J.; Lu, H.; Yao, J. ScBERT as a Large-Scale Pretrained Deep Language Model for Cell Type Annotation of Single-Cell RNA-Seq Data. *Nat Mach Intell* **2022**, *4* (10), 852–866. https://doi.org/10.1038/s42256-022-00534-z.
(11) Cui, H.; Wang, C.; Maan, H.; Pang, K.; Luo, F.; Duan, N.; Wang, B. ScGPT: Toward Building a Foundation Model for Single-Cell Multi-Omics Using Generative AI. *Nat Methods* **2024**, *21* (8), 1470–1480. https://doi.org/10.1038/s41592-024-02201-0.
(12) Szałata, A.; Hrovatin, K.; Becker, S.; Tejada-Lapuerta, A.; Cui, H.; Wang, B.; Theis, F. J. Transformers in Single-Cell Omics: A Review and New Perspectives. *Nat Methods* **2024**, *21* (8), 1430–1443. https://doi.org/10.1038/s41592-024-02353-z.
(13) Hao, M.; Gong, J.; Zeng, X.; Liu, C.; Guo, Y.; Cheng, X.; Wang, T.; Ma, J.; Zhang, X.; Song, L. Large-Scale Foundation Model on Single-Cell Transcriptomics. *Nat Methods* **2024**, *21* (8), 1481–1491. https://doi.org/10.1038/s41592-024-02305-7.
(14) OpenAI. GPT-4 Technical Report. **2023**.
(15) Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; Scialom, T. Llama 2: Open Foundation and Fine-Tuned Chat Models. **2023**.
(16) Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; Gao, J. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. **2023**.
(17) Zhang, K.; Zhou, R.; Adhikarla, E.; Yan, Z.; Liu, Y.; Yu, J.; Liu, Z.; Chen, X.; Davison, B. D.; Ren, H.; Huang, J.; Chen, C.; Zhou, Y.; Fu, S.; Liu, W.; Liu, T.; Li, X.; Chen, Y.; He, L.; Zou, J.; Li, Q.; Liu, H.; Sun, L. BiomedGPT: A Generalist Vision-Language Foundation Model for Diverse Biomedical Tasks. **2023**. https://doi.org/10.1038/s41591-024-03185-2.
(18) Yang, L.; Xu, S.; Sellergren, A.; Kohlberger, T.; Zhou, Y.; Ktena, I.; Kiraly, A.; Ahmed, F.; Hormozdiari, F.; Jaroensri, T.; Wang, E.; Wulczyn, E.; Jamil, F.; Guidroz, T.; Lau, C.; Qiao, S.;

Liu, Y.; Goel, A.; Park, K.; Agharwal, A.; George, N.; Wang, Y.; Tanno, R.; Barrett, D. G. T.; Weng, W.-H.; Mahdavi, S. S.; Saab, K.; Tu, T.; Kalidindi, S. R.; Etemadi, M.; Cuadros, J.; Sorensen, G.; Matias, Y.; Chou, K.; Corrado, G.; Barral, J.; Shetty, S.; Fleet, D.; Eslami, S. M. A.; Tse, D.; Prabhakara, S.; McLean, C.; Steiner, D.; Pilgrim, R.; Kelly, C.; Azizi, S.; Golden, D. Advancing Multimodal Medical Capabilities of Gemini. 2024.

(19) Boiko, D. A.; MacKnight, R.; Kline, B.; Gomes, G. Autonomous Chemical Research with Large Language Models. *Nature* **2023**, *624* (7992), 570–578. https://doi.org/10.1038/s41586-023-06792-0.

(20) Hou, W.; Ji, Z. Assessing GPT-4 for Cell Type Annotation in Single-Cell RNA-Seq Analysis. *Nat Methods* **2024**, *21* (8), 1462–1465. https://doi.org/10.1038/s41592-024-02235-4.

(21) Dziri, N.; Milton, S.; Yu, M.; Zaiane, O.; Reddy, S. On the Origin of Hallucinations in Conversational Models: Is It the Datasets or the Models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp 5271–5285. https://doi.org/10.18653/v1/2022.naacl-main.387.

(22) Lewkowycz, A.; Andreassen, A.; Dohan, D.; Dyer, E.; Michalewski, H.; Ramasesh, V.; Slone, A.; Anil, C.; Schlag, I.; Gutman-Solo, T.; Wu, Y.; Neyshabur, B.; Gur-Ari, G.; Misra, V. Solving Quantitative Reasoning Problems with Language Models. **2022**.

(23) Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.; Rocktäschel, T.; Riedel, S.; Kiela, D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., Lin, H., Eds.; Curran Associates, Inc., 2020; Vol. 33, pp 9459–9474.

(24) Bagley, S. J.; Binder, Z. A.; Lamrani, L.; Marinari, E.; Desai, A. S.; Nasrallah, M. P.; Maloney, E.; Brem, S.; Lustig, R. A.; Kurtz, G.; Alonso-Basanta, M.; Bonté, P. E.; Goudot, C.; Richer, W.; Piaggio, E.; Kothari, S.; Guyonnet, L.; Guerin, C. L.; Waterfall, J. J.; Mohan, S.; Hwang, W. T.; Tang, O. Y.; Logun, M.; Bhattacharyya, M.; Markowitz, K.; Delman, D.; Marshall, A.; Wherry, E. J.; Amigorena, S.; Beatty, G. L.; Brogdon, J. L.; Hexner, E.; Migliorini, D.; Alanio, C.; O'Rourke, D. M. Repeated Peripheral Infusions of Anti-EGFRvIII CAR T Cells in Combination with Pembrolizumab Show No Efficacy in Glioblastoma: A Phase 1 Trial. *Nat Cancer* **2024**, *5* (3), 517–531. https://doi.org/10.1038/s43018-023-00709-6.

(25) Mathewson, N. D.; Ashenberg, O.; Tirosh, I.; Gritsch, S.; Perez, E. M.; Marx, S.; Jerby-Arnon, L.; Chanoch-Myers, R.; Hara, T.; Richman, A. R.; Ito, Y.; Pyrdol, J.; Friedrich, M.; Schumann, K.; Poitras, M. J.; Gokhale, P. C.; Gonzalez Castro, L. N.; Shore, M. E.; Hebert, C. M.; Shaw, B.; Cahill, H. L.; Drummond, M.; Zhang, W.; Olawoyin, O.; Wakimoto, H.; Rozenblatt-Rosen, O.; Brastianos, P. K.; Liu, X. S.; Jones, P. S.; Cahill, D. P.; Frosch, M. P.; Louis, D. N.; Freeman, G. J.; Ligon, K. L.; Marson, A.; Chiocca, E. A.; Reardon, D. A.; Regev, A.; Suvà, M. L.; Wucherpfennig, K. W. Inhibitory CD161 Receptor Identified in Glioma-Infiltrating T Cells by Single-Cell Analysis. *Cell* **2021**, *184* (5), 1281-1298.e26. https://doi.org/10.1016/j.cell.2021.01.022.

(26) Srirat, T.; Hayakawa, T.; Mise-Omata, S.; Nakagawara, K.; Ando, M.; Shichino, S.; Ito, M.; Yoshimura, A. NR4a1/2 Deletion Promotes Accumulation of TCF1+ Stem-like Precursors of Exhausted CD8+ T Cells in the Tumor Microenvironment. *Cell Rep* **2024**, *43* (3), 113898. https://doi.org/10.1016/j.celrep.2024.113898.

(27) Liu, X.; Wang, Y.; Lu, H.; Li, J.; Yan, X.; Xiao, M.; Hao, J.; Alekseev, A.; Khong, H.; Chen, T.; Huang, R.; Wu, J.; Zhao, Q.; Wu, Q.; Xu, S.; Wang, X.; Jin, W.; Yu, S.; Wang, Y.; Wei, L.; Wang, A.; Zhong, B.; Ni, L.; Liu, X.; Nurieva, R.; Ye, L.; Tian, Q.; Bian, X.-W.; Dong, C. Genome-Wide Analysis Identifies NR4A1 as a Key Mediator of T Cell Dysfunction. *Nature* **2019**, *567* (7749), 525–529. https://doi.org/10.1038/s41586-019-0979-8.