# AI AGENT FOR DATA-DRIVEN HYPOTHESIS GENERATION IN SINGLE-CELL TRANSCRIPTOMICS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Large Language Models (LLMs) have the ability to utlilize expert knowledge and simulate human thinking, which potentially makes them instrumental for a variety of scientific tasks. However, since scientific data is heterogeneous, often presented in the form of unordered tables, bridging the gap between unstructured non-textual data and the language processing capabilities of LLMs remains an open challenge. Agentic AI offers a promising approach by enabling LLMs to interactively query datasets for relevant information. Here, we explore the application of this agentic paradigm to single-cell transcriptomic analysis, with a specific focus on cell type annotation. Our results show that when LLMs are equipped with data-querying capabilities, their performance in annotating cell types improves significantly compared to single-shot prompting. Furthermore, we provide a proof of concept illustration of how our method can be used to integrate diverse single-cell datasets (e.g., cell census), ensuring consistent annotation across multiple sources, facilitating meta-analysis across big sample cohorts.

## 1 INTRODUCTION

Large Language Models (LLMs) have proven to be highly useful across a variety of scientific domains. For instance, there are specialized models like ChemCrow (Bran et al. (2023)), which predict properties of chemical reactions, or BrainGPT (Luo et al. (2024)) that analyzes neuroscience data. Their success stems from their ability to effectively utilize vast amounts of human knowledge, encoded in scientific literature, and simulate basic human reasoning patterns. Moreover, the recent development of agentic paradigms has further enhanced these capabilities by allowing LLMs not only to process information but also to actively plan, query, and execute tasks. These developments can solve a fundamental problem for a wide-scale adoption of AI system for the analysis of diverse scientific datasets.

LLMs have attracted significant attention in the field of single-cell transcriptomics as well, spurring a surge of attempts to integrate these models with transcriptomic data. Many of these approaches, such as CellWhisperer (Schaefer et al. (2024)), rely on fine-tuning the model so that it processes cells as additional tokens—a method that, while powerful, demands extensive training and does not fully leverage the model's inherent reasoning capabilities. An alternative strategy converts the data into a text format, as seen in implementations like GPTCelltype (Hou & Ji (2024)). However, this text-based approach is constrained by the limited context window of LLMs, preventing the inclusion of all relevant data. These challenges highlight the need for a more dynamic system – one that can interact with and query extensive datasets in a flexible manner.

To address these challenges, we developed LAMBDA (Language Agent for Molecular Biological Data Analysis), an LLM-based agent that bridges large language models (e.g., Gemini model family (Team et al. (2023)), Gpt-4o (Hurst et al. (2024)), or Claude (Anthropic (2024)) with single-cell data. Unlike previous approaches such as GPTCelltype and CellWhisperer, LAMBDA supports bidirectional interaction: it not only retrieves data for the model but also allows the model to query the data. For that we devised a protocol that facilitates the interaction between LLMs and the data that overcomes context window size limitations, mitigates most of the LLM hallucinations issues and helps the agent to converge to an optimal solution. The principles that we implement in LAMBDA comprise a general strategy that can be used to perform data-driven hypothesis generation and testing in single cell omic data using LLMs.

As a proof of concept, we explore the role of LAMBDA as a cell typing assistant. Annotation of cell types is integral to the analysis of single cell data and is most frequently performed either manually or using label transfer methods (Domínguez Conde et al., 2022) which rely on existing annotations. Unlike these approaches, LAMBDA offers an automatic way to perform cell typing independent of human input based exclusively on LLM knowledge. This feature is helpful for the annotation of novel datasets and mitigating the effects of human biases. To showcase the advantages of agentic mode of using LLMs to single shot prompting, we include comparison of these two strategies on atlas-level datasets. Finally, we discuss the perspective of using LAMBDA for meta analysis over large collections of datasets, such as CELLxGENE Discover Census (Program et al. (2025)), to identify gene expression and cell type abundance patterns associated with various covariates.

## 2 Related work

### 2.1 Single cell omic data

Single-cell omics technologies measure molecular characteristics in individual cells, providing high-resolution profiles of cellular states and functions. Unlike bulk assays, which obscure heterogeneity by averaging measurements across cell populations, single-cell data illuminates cell-to-cell variability and enables the identification of rare cell types and subpopulations. Single-cell transcriptomics is one of the most widely used types of exeriment. It quantifies the abundance of individual RNA transcripts within each cell, revealing cell-specific gene expression patterns. Such data is critical for dissecting complex biological processes, including developmental lineages, immune responses, disease mechanisms, and cellular responses to stimuli.

A powerful application of single-cell omics is the construction of single-cell atlases. These atlases aim to comprehensively map all cell types within an organism, tissue, or organ, providing a foundational resource for understanding cellular organization and function. By integrating data from numerous single-cell experiments, these atlases capture the full spectrum of cellular states and their relationships to each other. Initiatives like the Human Cell Atlas (Atlas, 2018) are generating comprehensive maps of the human body, promising to accelerate discoveries in basic biology and medicine.

### 2.2 LLMs in single cell omics

Several strategies have been developed to use LLMs in single-cell analysis. Broadly, these approaches can be categorized into three groups: (i) tokenization-based , (ii) methods utilizing single-shot prompting, and (iii) agent-based techniques.

Examples of **tokenization-based methods** include scBERT (Yang et al. (2022)), and the more recent CellWhisperer (Schaefer et al. (2024)). These models aim to learn a mapping from gene expression to the token space of LLM. While these approaches are potentially powerful, they present notable challenges. First, the fine-tuning process is computationally expensive, and applying it to state-of-the-art models is often impractical given that many high-performing models are closed source. Second, performance of these methods heavily rely on the quality of data annotations used for fine-tuning, which may be inconsistent or inaccurate, resulting in the model copying human biases.

**Single-shot prompting methods**, such as GPTCelltype (Hou & Ji (2024)), provide gene expression information to an LLM directly in the form of text without requiring fine-tuning. The primary advantage of this approach is its ease of integration with off-the-shelf models. However, these methods may be limited by the context window of an LLM, since full gene expression information can not be passed as a single input. Moreover, LLM are not good at handling numerical data.

**Agent-based methods** potentially may solve the drawbacks of single-shot methods by providing an LLM with the ability to query the dataset in an interactive manner as well as equipping it with the ability to perform numerical computations using external tools. An example of an agent for single cell data is CellAgent (Xiao et al. (2024)). However, this method primarily focuses on the utilization of tools rather than enhancing the model's inherent reasoning capabilities by allowing multi step reasoning process.

### 2.3 Prompting LLMs

Apart from equipping LLM with agentic capabilities, the key challenges of building LAMBDA included mitigating LLM hallucinations and forcing it to reason over the input data integrating various aspects of it. These things can be addressed by refining the prompts to an LLM and so we present two notable prompting strategies: Chain-of-Thought (CoT) (Wei et al. (2022)) and Tree-of-Thought (ToT) prompting (Yao et al. (2024)).

Chain-of-Thought prompting encourages models to articulate a sequential series of reasoning steps. Rather than leaping directly to an answer, the model is guided to break down a complex problem into logical, incremental stages. This approach mirrors human problem-solving, decomposing intricate tasks into smaller, manageable parts, thereby not only improving the accuracy of the final result but also providing transparency into the model's reasoning process.

Tree-of-Thought prompting takes this idea further by allowing the model to explore multiple reasoning pathways simultaneously. Instead of adhering to a single, linear sequence, the model branches out to evaluate diverse solution strategies in parallel. This branching mechanism enables the model to consider various perspectives and converge on a more robust solution, effectively mimicking a decision-making process where multiple scenarios are weighed before arriving at a final answer.

## 3 Methods

### 3.1 Overview of LAMBDA

LAMBDA is an LLM-based agent that is aimed to bring expert knowledge of LLMs into single cell analysis by allowing the model to interactively query the data, use statistical tests and remember results of its intermediate thinking steps. On a conceptual level LAMBDA is best described as a sequence of 4 steps: 1) retrieval of the relevant data, 2) data-driven hypothesis generation by LLM, 3) data-based hypothesis testing using LLM criteria, 4) aggregation of the hypotheses.

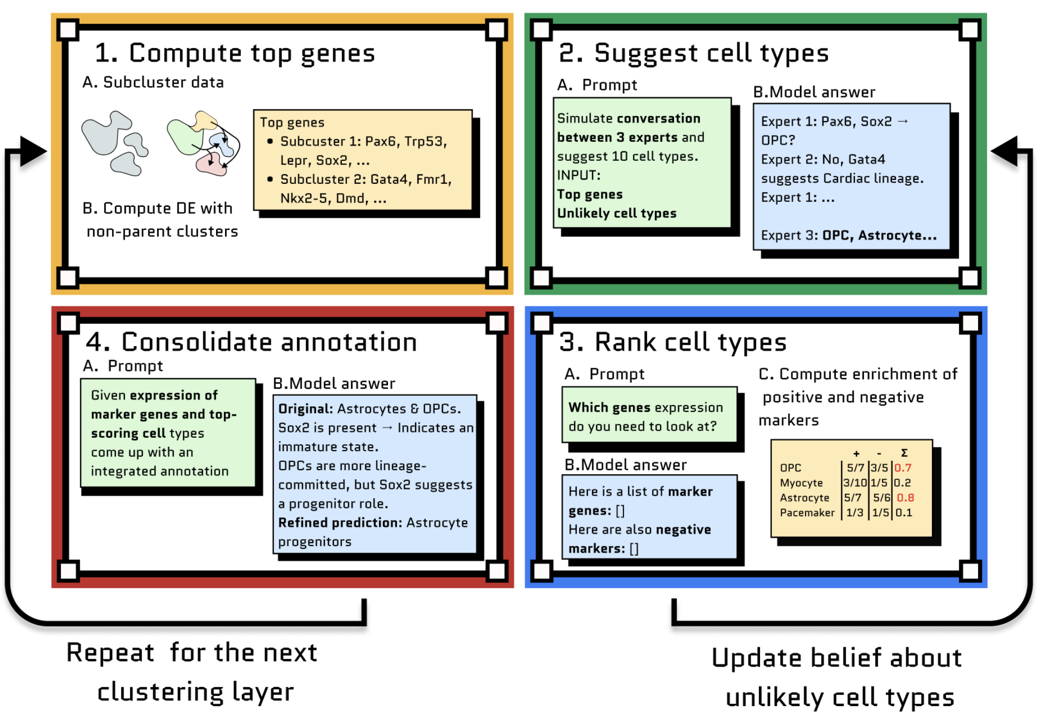


Figure 1: Schematic illustrating key steps of the LAMBDA pipeline for cell type annotation. Note that the prompts used in this schematic are illustrative, the real prompts can be found in the supplementary.

For cell typing, LAMBDA applies these 4 steps operating on the level of individual cell clusters (**Figure 1**). The annotation pipeline starts with 2-level clustering of the data and the identification of top enriched genes within each of the subclusters of the analyzed clusters accounting for potential cell type heterogeneity within cluster. Next, using these enriched genes as input, LAMBDA prompts the model to suggest ten potential cell types (Step 2). Subsequently, in Step 3, the model is queried for marker genes associated with each of these proposed cell types, enabling their ranking. The marker genes identified in Step 3 are then fed back into the model (Step 2), prompting the generation of ten new potential cell types, considering those previously deemed less likely. This iterative cycle of Steps 2 and 3 is repeated multiple times. This iterative application enforces a more comprehensive exploration of potential cell type assignments, recovering cell types that might have been overlooked in initial iterations. Finally, in Step 4, the most probable annotations are aggregated and consolidated into a final cell type assignment.

#### 3.1.1 Eliciting Reasoning in LLMs

LAMBDA uses top enriched genes within each cluster and optionally tissue of origin and a list of unlikely cell types to predict cell type label. To ensure the LLM provides a comprehensive analysis and explores a wide range of possibilities, we use a "tree-of-thought"-like prompting strategy. The prompt guides the LLM to simulate a discussion between two experts, followed by a summary from a third expert (Supplementary text S2 and example response Supplementary text S6). Here is a simplified example:

```
Your task is to simulate a report by three expert biologists about the
identification of cell types based on the observed data.

The main part consists of 3 rounds, within each round 2 experts describe
various traits of the cells debating the position of each other.
Following the debate, a third expert provides a concluding review and
suggests 10 potential cell types.
```

This strategy resembles tree-of-thought prompting in that it also involves a simulated conversation between experts, each exploring different solution paths. However, it diverges from tree-of-thought in that the experts debate not the final solution (cell type annotation) itself, but rather the underlying traits of the cells. This creates a rich, multifaceted description, focusing on diverse features and preventing the model from fixating on only a limited set of features. Furthermore, the role of the third expert differs; instead of contributing to the trait discussion, this expert synthesizes the preceding discussion to generate ten probable cell type annotations, ensuring broad coverage of plausible hypotheses.

#### 3.1.2 Mitigating LLM hallucinations

A major hurdle in using LLMs for scientific research is their tendency to "hallucinate," meaning they report facts that are not real. Addressing this issue was crucial during the development of LAMBDA. The two steps of our framework that are the most vulnerable to hallucinations are cell type generation and marker gene quering.

Although LLMs may sometimes invent cell types, we try to minimize this by requiring it to use terms from the established cell ontology. While this approach does not make the model to use the exact ontology terms, it generally helps to keep the output consistent with accepted biological classifications:

```
• Exclude any entries that highlight specialized or non-standard
functions (e.g., B cell-interacting DC) rather than recognized
classifications.
• Exclude entries that specify unique expression patterns not used in
conventional nomenclature, such as cell ontology (e.g., TCF4+ DC).
```

The next crucial step in the pipeline is scoring marker gene signatures reported by the LLM. The results of this step determine which cell types are considered for the final cell type assignment. Therefore, the consistency and reliability of the signature are important. To compile this signature the model is instructed to produce a set of marker genes expected to be enriched (positive markers). Given that some cell types or stages are distinguished by the absence of particular genes, we also ask it to report genes anticipated to be depleted (negative markers). Differential gene expression of these genes is then computed within subclusters compared to a reference set of cells. Based on the number of significant hits, the normalized score is computed (Formula 1).

$$\text{score} = \left(\frac{\text{num\_of\_significant\_positive}}{\text{num\_of\_positive}}\right) \times \alpha + \left(\frac{\text{num\_of\_significant\_negative}}{\text{num\_of\_negative}}\right) \times (1-\alpha) \quad (1)$$

To ensure LLM reports a marker signature covering diverse aspects of cell identity, we first prompt it to construct a hierarchical tree of cell types based on ”is subtype” relationships (e.g., CD4 T cells as a subtype of T cells, which are in turn subtypes of lymphocytes). The model then reports gene markers for each level of this hierarchy. An excerpt from the prompt is provided below:

```
Construct a hierarchy for each cell type that represents how specific the
cell type definition is: child shares attributes with parent (T helper is
a T cell) but parent doesn't share attributes with child (T cell is not
a T helper).  Include in this tree siblings, cousins and cell types with
similar transcriptional signature...
```

To further reduce spurious associations, we take the initially generated marker signature and ask the model to validate it (**Supplementary text S4**). The model is required to provide justifications for why certain genes are associated with a specific cell type and why others are not:

```
Discuss the functions of each gene in the report and provide its HGNC
symbol, for each gene explicitly state whether its expression is HIGH,
LOW, ABSENT or UNKNOWN in {cts}. {organism}...
```

We found that this strategy increased the consistency of marker genes between runs and reduced the number of spuriously associated genes. An alternative strategy to mitigate the stochasticity of an LLM would be to run this prompt multiple times and use an aggregated signature; however, this incurs additional computational overhead. Therefore, we opted for the described approach.

### 3.1.3 Implementing decision making for hierarchical analysis

Since LLMs can not analyze the expression patterns of every single cell in a large dataset, we give the model summarized data at the cluster level. A potential downside of this approach is the loss of information regarding intra-cluster heterogeneity. We address this using a hierarchical analysis of clusters, which works in two main ways.

First, when we analyze a cluster, we further divide it into subclusters and compute both enriched genes and differential gene expression in each subcluster. As well as detecting within cluster heterogenity, the added benefit of this hierarchical approach is that it helps us use the right cells as a reference for differential gene expression. In our setup, which includes parent supercluster, the specific cluster being analyzed, and the smaller subclusters within it, we compute differential gene expression of each subcluster relative to all the cells in the supercluster exclusing the analyzed cluster itself. This ensures that when looking at specific subtypes (like subtypes of CD4 cells), the comparison is done to a relevant group (like all T cells, but and not CD4 or any random group of cells in the dataset).

Second, if at the consolidation step LAMBDA detects cluster heterogeneity, it can decide to analyze each subcluster separately. If this happens, the model keeps the original cluster-level annotation but allows the LLM some flexibility to adjust the subcluster annotations. The decision to cluster further is done if different subclusters have different highest scoring cell types.

## 4 Results

### 4.1 Agent-based method surpasses single-shot prompting strategy

To demonstrate the advantages of an agent-based approach in leveraging LLMs for cell type annotation, we applied our method to two single-cell atlases: one of the human intestinal tract (Elmentaite et al. (2021)) and another of the human lung (Sikkema et al. (2023)). In each case, the LLM was tasked with predicting cell types, and the resulting predictions were compared against the atlas annotations.

For evaluation, we defined five categories to characterize the relationship between the predicted cell types and the atlas annotations:

| Category | Description |
|---|---|
| **Identical** | The predicted cell type exactly matches the atlas annotation. |
| **Subtype** | The predicted cell type is a subtype of the atlas annotation. |
| **Supertype** | The predicted cell type is a supertype of the atlas annotation. |
| **Sibling** | The predicted cell type is a sibling of the atlas annotation; both share an immediate common parent in the cell hierarchy (e.g., Th1 and Th2 are siblings because they are both subtypes of T helper cells). |
| **Unrelated** | The predicted cell type is too distinct from the atlas annotation. |

Table 1: Categories of Predicted Cell Type Relationships with Atlas Annotations

Our results indicate that the agent-based prompting approach significantly outperforms the single-shot strategy across both datasets (**Figure 2 A-B**). For example, in the intestinal tract atlas, the number of identical matches with the agent-based method was more than twice that achieved by the single-shot approach. Furthermore, LAMBDA was able to identify a greater number of subtypes, revealing cell types that had been annotated too coarsely in the original atlas. Conversely, the single-shot strategy tended to produce overly coarse annotations, as evidenced by the predominance of the *supertype* category.

Our results indicate that the agent-based prompting approach significantly outperforms the single-shot strategy across both datasets (**Figure 2 A-B**). For example, in the intestinal tract atlas, the agent-based method produced more than twice as many identical matches as the single-shot approach. In addition, LAMBDA identified more subtypes, showing that some cell types had been grouped too broadly in the original atlas. On the other hand, the single-shot strategy tended to create overly broad annotations, as shown by the large number of results in the *supertype* category.

#### 4.1.1 Discrepancies with atlas annotation

To investigate the reasons for discrepancies between LAMBDA's annotations and the atlas, we analyzed marker gene signature enrichment for both sets of annotations, as shown in **Figure 2 C**. This analysis revealed several patterns. For example, cells annotated as GIP cells in the atlas were assigned a more specific label by LAMBDA: SST+, PYY+, GCG+, CCK+ enteroendocrine cells. Examination of the expression of these marker genes confirmed their presence within the cluster. This suggests that the original atlas annotation, while not incorrect, was overly broad, and LAMBDA's approach provided a more refined and biologically detailed classification. In another instance, LAMBDA predicted gamma-delta T cells in a population annotated as ILC3 in the atlas. Intriguingly, the marker gene signature associated with the atlas's ILC3 annotation showed only weak enrichment in these cells, suggesting a potential misannotation or limitation in the original atlas.

Beyond discrepancies with existing annotations, our analysis also highlighted challenges related to the inherent complexity of single-cell data. In one case, LAMBDA predicted gamma-delta T cells, while the atlas labeled the same population as ILC2. However, closer examination revealed that these cells displayed marker signatures characteristic of both cell types. This ambiguity suggests that these cells might be doublets or representing a mixture of distinct cell populations. Similarly, in another scenario, cells predicted to be L cells by LAMBDA were annotated as N cells in the atlas. Gene expression profiling suggested these cells were likely a heterogeneous mixture or, again, indicative of a transitional state, blurring the lines between distinct cell identities.

These observations collectively indicate that discrepancies in cell type annotations can arise from multiple factors. One factor is the inherent difficulty for any model, including LLMs, in fully capturing the subtle and nuanced gene expression patterns that define every cell type. Furthermore, the presence of doublets within cell clusters introduces ambiguity, leading to mixed marker signatures that complicate accurate annotation. The intrinsic heterogeneity of certain cell clusters, encom-

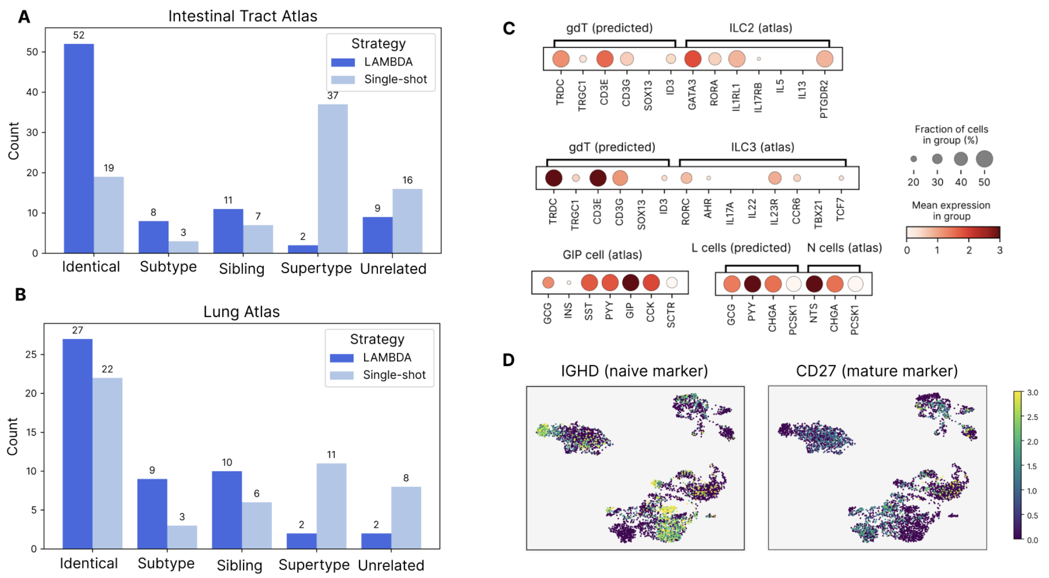


Figure 2: LAMBDA outperforms single-shot prompting strategies and identifies in consistencies within cell atlases. A-B. accuracy of predictions of an agent based and single shot strategies. C. Analysis of the cell clusters that were wrongly annotated by LAMBDA, gene signatures for the atlas annotation and predicted annotation are shown. D. LAMBDA identifies heterogeneity within clusters from atlas annotation.

passing multiple distinct cell populations, also presents a challenge to achieving a single, precise annotation. Finally, it is important to acknowledge that limitations and potential inaccuracies within the original atlas annotations themselves can contribute to apparent discrepancies. Understanding these multifaceted reasons is crucial for refining cell type annotation methodologies and interpreting single-cell data.

LAMBDA's ability to detect mixed populations is particularly intriguing, as overly broad annotations are common in atlases. For instance, in the lung atlas, the original annotation grouped all B cells under a single label. LAMBDA, however, identified distinct subpopulations corresponding to naive and mature B cells. Analysis of the marker distributions for these two states (**Figure 2 D**) confirmed the presence of several naive and several mature populations, underscoring the enhanced resolution provided by our agent-based method.

### 4.2 Analysis of CELLxGENE Discover Census

Finally, we show a major benefit of automatic cell typing that goes beyond analyzing just one dataset: the ability to do meta-analysis. We believe LAMBDA's full potential is unlocked when it is used on diverse collections of single-cell datasets, like those in the CELLxGENE Discover Census. To show this, we sampled a fraction of the B cells from the CELLxGENE with a focus on tissue diversity, creating a collection of over 200,000 cells from 119 datasets. These cells came from 88 tissues from healthy samples and 36 disease conditions. We clustered the cells using scVI embeddings (Lopez et al. (2018)), and then used LAMBDA to get replace the original cell type labels which belonged to different levels of cell type hierarchy with the uniform ones.

The final annotation included 16 different labels (**Figure 3 A**). It's important to note that not all the annotated cells were classified as B cell subtypes. For example, one cluster was identified as oligodendrocytes, and enrichment analysis showed high expression of the genes PLP1, MBP, and CRYAB, which are known markers for oligodendrocytes (Kim et al. (2021); Solly et al. (1996); Kuipers et al. (2017)) (**Figure S1**). This shows how LAMBDA can improve existing atlas annotations by consistently assigning cell type labels across different clusters and finding cell populations that might have been incorrectly labeled or missed in the original atlas.

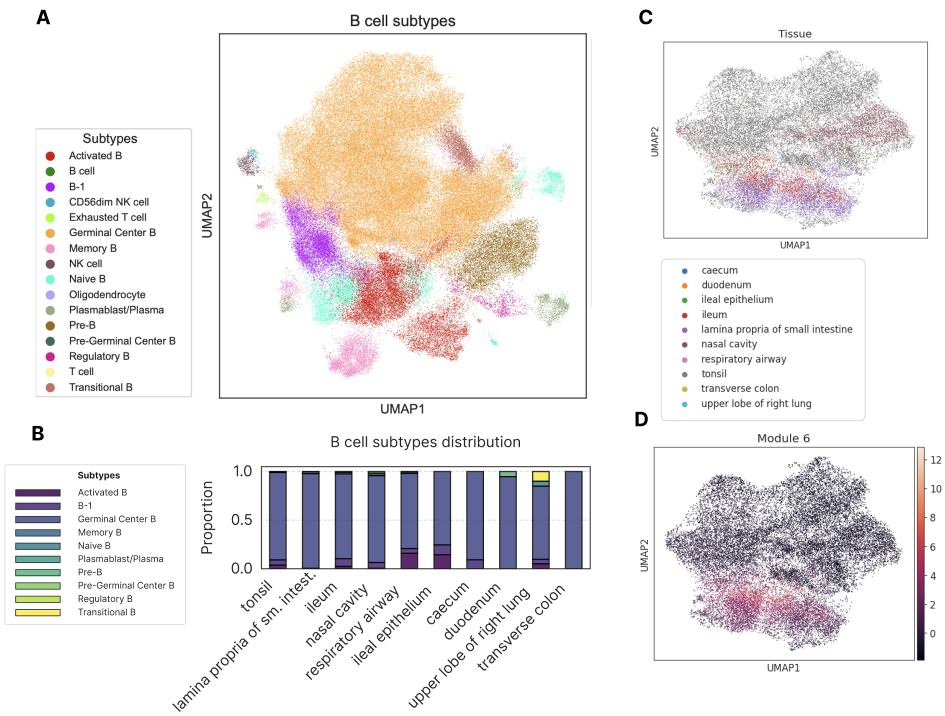


Figure 3: Uniform annotation allows identification of tissues with similar subtype composition.

Furthermore, we demonstrate how the resulting annotations can facilitate meta-analysis across various metadata categories, including age, sex, disease, and tissue. To achieve this, we performed clustering of tissues based on the proportions of different cell types (**Figure S2 A**). This approach is particularly advantageous, as it circumvents the challenge posed by the original data, where cells are annotated at varying levels of granularity across different datasets. Our method enabled the identification of tissue groups with similar B cell subtype compositions.

To understand why certain tissues clustered together, we zoomed in on one particular cluster and investigated which cell types were most abundant. This revealed a high proportion of germinal center B cells (**Figure 3 B**). Next, we wanted to see if this cell type showed differences across the various tissues. Using Hotspot, we identified several gene modules (**Figure S3 A**), and, strikingly, some of these modules were found only in specific tissue subgroups: the ileum, duodenum, and lamina propria—all parts of the digestive system (**Figure 3 C-D**). This finding demonstrates the power of our approach to uncover new biological insights.

## 5 Conclusion

The ever-growing volume of single-cell omics data holds immense potential for biological discovery, but realizing this potential hinges on developing methods for automated, data-driven hypothesis generation and testing. A parallel revolution is occurring with the rise of LLMs. Taken together, LLMs and single-cell omics offer a powerful synergy. To harness this potential, we introduce LAMBDA, an AI agent that enables LLM to actively explore single cell data. LAMBDA does not just passively process data; it actively interrogates it, generating hypotheses and testing them against the evidence, much like a scientist would. When applied to cell typing, this dynamic strategy leads to significantly more accurate annotations compared to single-shot prompting of LLMs and, unlike traditional methods, it does not require any reference data.

By enabling uniform cell typing across vast, heterogeneous collections of single cell datasets, it unlocks the door to the meta-analyses, allowing researchers to explore biological questions across a multitude of conditions and physiological contexts.

## REFERENCES

Anthropic. Claude: Large language model, 2024. URL `https://www.anthropic.com`. Accessed: 2025-02-13.

Human Cell Atlas. Human cell atlas. *Human cell atlas sequences first 250K developmental cells*, 2018.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

C Domínguez Conde, C Xu, LB Jarvis, DB Rainbow, SB Wells, T Gomes, SK Howlett, O Suchanek, K Polanski, HW King, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594):eabl5197, 2022.

Rasa Elmentaite, Natsuhiko Kumasaka, Kenny Roberts, Aaron Fleming, Emma Dann, Hamish W King, Vitalii Kleshchevnikov, Monika Dabrowska, Sophie Pritchard, Liam Bolt, et al. Cells of the human intestinal tract mapped across space and time. *Nature*, 597(7875):250–255, 2021.

Wenpin Hou and Zhicheng Ji. Assessing gpt-4 for cell type annotation in single-cell rna-seq analysis. *Nature Methods*, pp. 1–4, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Dongkyeong Kim, Hongjoo An, Chuandong Fan, and Yungki Park. Identifying oligodendrocyte enhancers governing plp1 expression. *Human molecular genetics*, 30(23):2225–2239, 2021.

Hedwich F Kuipers, Jane Yoon, Jack Van Horssen, May H Han, Paul L Bollyky, Theo D Palmer, and Lawrence Steinman. Phosphorylation of $\alpha$b-crystallin supports reactive astrogliosis in demyelination. *Proceedings of the National Academy of Sciences*, 114(9):E1745–E1754, 2017.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

Xiaoliang Luo, Akilles Rechardt, Guangzhi Sun, Kevin K Nejad, Felipe Yáñez, Bati Yilmaz, Kangjoo Lee, Alexandra O Cohen, Valentina Borghesani, Anton Pashkov, et al. Large language models surpass human experts in predicting neuroscience results. *Nature human behaviour*, pp. 1–11, 2024.

CZI Cell Science Program, Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, et al. Cz cellxgene discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Research*, 53(D1):D886–D900, 2025.

Moritz Schaefer, Peter Peneder, Daniel Malzl, Mihaela Peycheva, Jake Burton, Anna Hakobyan, Varun Sharma, Thomas Krausgruber, Joerg Menche, Eleni M Tomazou, et al. Multimodal learning of transcriptomes and text enables interactive single-cell rna-seq data exploration with natural-language chats. *bioRxiv*, pp. 2024–10, 2024.

Lisa Sikkema, Ciro Ramírez-Suástegui, Daniel C Strobl, Tessa E Gillett, Luke Zappia, Elo Madissoon, Nikolay S Markov, Laure-Emmanuelle Zaragosi, Yuge Ji, Meshal Ansari, et al. An integrated cell atlas of the lung in health and disease. *Nature medicine*, 29(6):1563–1577, 2023.

SK Solly, J-L Thomas, M Monge, C Demerens, C Lubetzki, MV Gardinier, J-M Matthieu, and B Zalc. Myelin/oligodendrocyte glycoprotein (mog) expression is associated with myelin deposition. *Glia*, 18(1):39–48, 1996.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Yihang Xiao, Jinyi Liu, Yan Zheng, Xiaohan Xie, Jianye Hao, Mingzhi Li, Ruitao Wang, Fei Ni, Yuxiao Li, Jintian Luo, et al. Cellagent: An llm-driven multi-agent framework for automated single-cell data analysis. *bioRxiv*, pp. 2024–05, 2024.

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866, 2022.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

## 6 SUPPLEMENTARY

**S1. Gene functions prompt**

```
You are compiling a nuanced report of gene functions. Within
each subcluster identify SEVERAL groups of genes with cell
type or biological function association. Genes which were
not assigned to any module should be put into other module.
{context}

Output format is a JSON dictionary. The top-level keys should
be the subcluster names ("Subcluster 1", "Subcluster 2", etc.).
Each subcluster key should have a value that is a dictionary.
Within these inner dictionaries, the keys will be module names
(e.g., "MHC Class II Presentation", "Antigen processing", avoid
referencing specific cell types) and the values will be lists
of genes (the items within the parentheses).

Don't add any disclaimers.

{subcluster_expression}
```

**S2. Cell type hypothesis generation prompt**

```
# Characterization of Cellular Identity

Your task is to simulate a collective report of three expert
biologists about the identification of cell types based on
the observed data. They analyze expression patterns, cellular
location, and other characteristics of a cell cluster split
into several subclusters.

## Report structure
1. The main part of the report is a structured report between
two experts. It consists of 3 rounds, within each round 2
experts describe various traits of the cells debating the
position of each other. Most likely the cells is question are
subtypes of {cell_context} but alternative possibilities should
be explored.
```

```
2. Following the debate, a third expert provides a concluding
review.  This review suggests {num_hypotheses} potential cell
types, each hypothesis referencing a specific cell subtype, not
a broad category (e.g., Th1 cells, not Th cells).

3. Expert 3 output should be structured as follows, the
name of the cell type enclosed in double square brackets:
"[[...]]". Example of the format: "1. [[Cell subtype name]]:
The presence of A-E genes and absence of K-Q genes hints at
this cell-subtypes..."
    • Exclude any entries that highlight specialized or
non-standard functions (e.g., \B cell-interacting DC") rather
than recognized classifications.
    • Exclude entries that specify unique expression patterns
not used in conventional nomenclature, such as cell ontology
(e.g., \TCF4+ DC").
    • Retain commonly accepted classifications, including
well-known markers (e.g., \CD4 T cells") or standard
tissue-specific designations (e.g., \tissue resident T cells").
    • Make sure that reported cell types are present in
{location} of {organism}.

## DATA
{DATA}
```

**S3. Marker gene prompt**

```
**Gene Markers for {cts}**

As a molecular biology and histology expert, identify and
annotate cell types based on marker gene expression.

1) Construct a hierarchy for each cell type that represents how
specific the cell type definition is: child shares attributes
with parent (T helper is a T cell) but parent doesn't share
attributes with child (T cell is not a T helper).  Include
in this tree siblings, cousins and cell types with similar
transcriptional signature.

2) Identify Positive Markers: List 10-15 reliable marker
genes highly expressed in "{cts}", covering various aspects
of its identity across different hierarchical levels, skip
non-specific levels. {context} {organism}

    - For example, for effector Th1 cells, include markers
for T cells (CD3D, CD3E, CD3G, CD247, CD2, CD5, CD28, PTPRC),
effector T helper cells (CD40LG, CD69, IL2RA, HLA-DRB1), and T
helpers of type 1 (TBX21, IFNG, IL12RB2, CXCR3, STAT4).

3) Identify Related Cell Types: Which cell types can also
express some of the marker genes listed above?

4) Identify Negative Markers: Report 5-10 genes expressed in
the cell types from point 2 but *not* in {cts}. These negative
```

```
markers are crucial for distinguishing {cts} from other similar
cell types, activation states, or differentiation stages.

5) In the end provide a list of all the positive markers
(only those that are not absent) in double square brackets
("[[...]]") and negative markers in double angular brackets:
"<<...>>".
```

**S4. Validation of marker genes prompt**

```
Analyze the marker gene report with a goal of compiling a lists
of positive and a list of negative marker genes for {cts}.

1) Discuss the functions of each gene in the report and provide
its HGNC symbol, for each gene explicitly state whether its
expression is HIGH, LOW, ABSENT or UNKNOWN in {cts}. {organism}

2) Provide summary lists: positive markers (HIGH) for {cts} as
a list in double square brackets ("Positive: [[HGNC1, HGNC2]]")
and negative markers (only ABSENT and LOW) for {cts} as a list
in double angular brackets ("Negative: <<HGNC1, HGNC2>>"). Use
HGNC gene symbols.

{positive}
{negative}
```

**S5. Consolidate cell type annotations**

```
You are provided with the results of single cell RNA-sequencing
analysis of a cell cluster. Based on the preliminary annotation
and reported expression patterns come up with a consolidated
annotation that summarizes the cell types.

Preliminary annotation: {annotation}.

1) Construct a hierarchy for each cell type that represents how
specific the cell type definition is: child shares attributes
with parent (T helper is a T cell) but parent doesn’t share
attributes with child (T cell is not a T helper). Include
in this tree siblings, cousins and cell types with similar
transcriptional signature.

2) Revise reported expression patterns. And decide if there is
enough evidence for each proposed cell type. For that simulate
a discussion between two experts trying to settle down on one
cell type. Each marker gene is important.

3) Provide academically acceptable consolidated name that
reflects subtype content, stage and gene expression. For
example: "T helper cells (Th1 and Th17)" or "B cells (pro-B
and pre-B stages)".
    •   Make sure to come up with a name that represents
diversity of input cell types in the preliminary annotation.
```

```
    •   Provide one final annotation, enclosed in double square
brackets (e.g., [[cell type (...) and ...]]).
    •   Avoid vague terms.


4) Narrow down the cell type annotation to the likeliest cell
ontology terms (without ids) and output this annotation in
double angular brackets: <<term1, term2>>.

{DATA}
```

**S6. Example cell type response**

```
Example LLM Answer:
## Expert Debates

### Round 1

**Expert 1:**
Looking at the expression patterns across all subclusters, the
presence of TPSAB1, TPSB2, CPA3, and HDC across all subclusters
is quite significant. These markers are classic indicators of mast
cells. Additionally, markers such as MS4A2, KIT, and FCER1A/FCER1G
support this identification due to their role in mast cell function
and signaling.

**Expert 2:**
I agree, the presence of TPSAB1, TPSB2, and CPA3 is compelling for
mast cells. However, I want to point out the consistent expression of
genes involved in immune signaling such as GATA2 and IL1RL1. These are
also found in basophils, which are known to have overlapping markers
with mast cells. The expression of GATA2, in particular, is noteworthy
as it’s essential for the development of basophils and mast cells
alike.

### Round 2

**Expert 1:**
While basophils might share some markers, the presence of CPA3
and HPGDS is more specific to mast cells rather than basophils.
Furthermore, the cellular structure and movement genes such as VIM
and RGS proteins suggest active tissue-resident cells, which are
characteristic of mast cells and less so of circulating basophils.

**Expert 2:**
You make a good point about CPA3 and HPGDS. However, I’d like to
emphasize the expression of ALOX5AP and SIGLEC6 found in some
subclusters, which also appear in basophils. The SIGLEC6 expression,
although less common, could indicate a regulatory function that
basophils perform in specific contexts, such as allergic responses
in the lung tissue.

### Round 3

**Expert 1:**
Considering the overall expression profile, the evidence heavily
favors mast cells, specifically those playing a role in histamine and
inflammatory responses in the lung. The diverse expression of immune
signaling and cellular movement genes supports their functionality as
tissue-resident effector cells.
```

```
**Expert 2:**
I concede that the mast cell evidence is strong given the CPA3
expression. Nonetheless, the presence of some basophil markers
suggests there might be a mixed population or a transitional state
captured in these subclusters, which is not uncommon in complex
tissues like the lung.

## Expert 3: Concluding Review

1. [[Connective Tissue Mast Cells]]: High levels of TPSAB1, TPSB2, and
CPA3 suggest these cells play a critical role in histamine response
within the lung.

2. [[Mucosal Mast Cells]]: The presence of KIT, IL1RL1, and HPGDS
supports identification as mast cells involved in mucosal immunity,
closely linked with allergic reactions.

3. [[Basophils]]: Shared expression of GATA2 and MS4A2 hints at
potential basophil presence, which can be involved in allergic
responses.

4. [[Interleukin-4 Producing Mast Cells]]: The immune signaling
profile suggests specialization in cytokine production, a known
function of certain mast cell subtypes.

5. [[Proinflammatory Mast Cells]]: Expression of FOS and JUNB
indicates active roles in inflammation, typical of mast cells
responding to tissue damage.

6. [[Tissue-Resident Mast Cells]]: High VIM and RGS proteins suggest a
stable, resident population within lung tissue.

7. [[IgE-Activated Mast Cells]]: FCER1A and FCER1G expression
indicates these cells may be involved in IgE-mediated allergic
reactions.

8. [[Cysteinyl Leukotriene Producing Mast Cells]]: Presence of LTC4S
and HPGD aligns with mast cells known for leukotriene synthesis,
important in asthma.

9. [[Allergen-Responsive Mast Cells]]: Gene expression patterns
suggest a role in immediate hypersensitivity reactions, crucial in
allergenic environments.

10. [[Neuropeptide-Activated Mast Cells]]: Expression of SLC18A2
may hint at interactions with neuropeptides, indicative of a sensory
function within the lung.
```

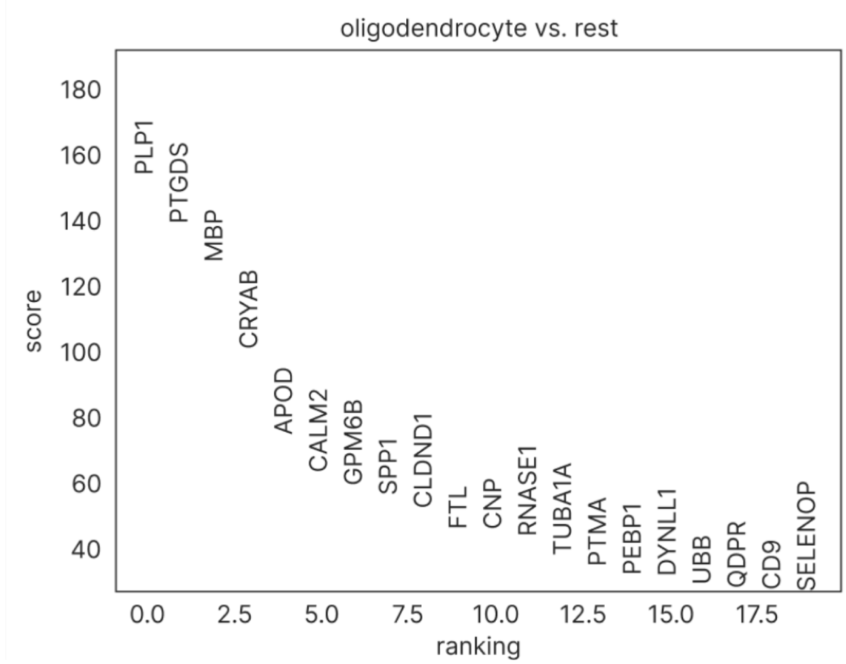


Figure S1: Enrichment of oligodendrocyte markers in a cell type which had been labeled as B cells in the CELLxGENE census

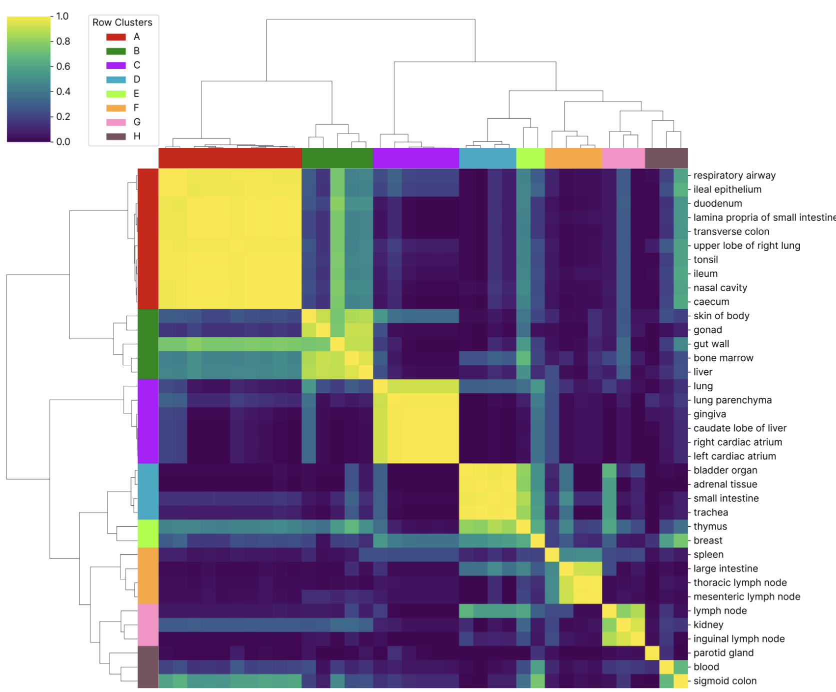


Figure S2: Clustering of tissues based on proportions of B cells

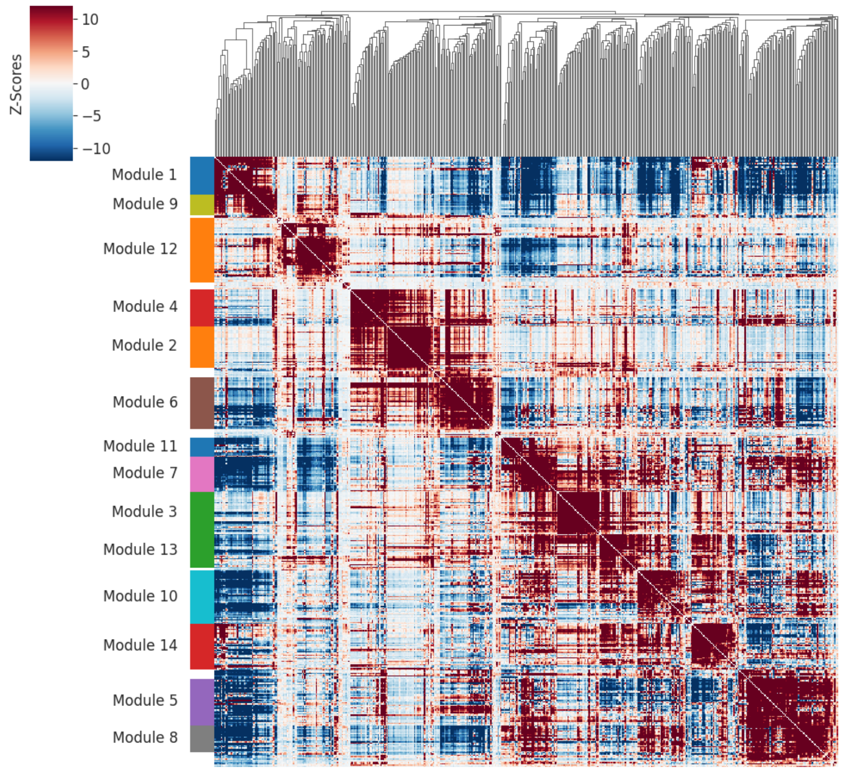


Figure S3: Modules identified with Hotspot in germinal center B cells.