CZI Data Insights Page (with instructions)

Title: Atlas-scale hierarchical identification of cell types and functions

Project Purpose (one sentence - 147/200 characters): To develop automated computational tools to predict cellular types and properties for single-cell data leveraging known hierarchical relationships.

Abstract (243/250 words):
A key deliverable of cell atlas projects is an operational definition of cell type that allows the automated labeling of cells in new datasets. So far, this goal is largely aspirational: single-cell annotation remains a painfully manual task, and insufficient training data and inconsistent labels have limited the utility of previous cell type classification models. However, many millions of cells with standardized labels have recently become available, most notably in the cell x gene Census dataset.

Here, we propose a classification framework that leverages the hierarchical nature of cell ontology labels to probabilistically classify both the cell types and functional properties of unseen cell profiles. Using a simple but powerful technique called marginalization classification for cells (McCells), we will train a model that outputs a probability distribution over the cell ontology for each cell or cluster. Crucially, our approach leverages the hierarchical nature of cell ontology labels during training, ensuring that (1) cells labeled with either general or specific labels contribute; and (2) the label probabilities for each data point obey the conditional independence relationships encoded in the ontology structure. McCells also elegantly models multiple inheritance in the cell ontology and is compatible with a variety of model architectures. Additionally, we will leverage cell ontology annotations of cell functions to predict functional descriptions of cells from gene expression. Ultimately, McCells will allow users to traverse the cell ontology and identify the probability of predicted cell types and functions at varying levels of specificity.

Keywords: Cell type identification, neural network, hierarchical classification, cell ontology

Project/Work Plan (974/1000 words):
Recent efforts have made huge strides toward experimentally characterizing the normal complement of cell types in the human, mouse, and other animals. Thus, the time is ripe for creating a model that encapsulates an operational definition of cell type and enables automated labeling of cells in new datasets.

The Cell Ontology[1] classifies cell types across different organisms and tissues in a directed acyclic graph, and contains over 2,700 cell types. The CZ CELLxGENE Census labels cells with cell ontology terms (**Fig. 1a**). The Census is the only large-scale dataset with these labels consistently applied. We plan to leverage this metadata to build a model that predicts cell type and cell function.

Here we propose to develop computational tools that will allow users to automatically annotate single-cell data in an accurate and consistent manner by leveraging the Census to develop a cross-tissue model. Our tool will streamline the cell annotation process while providing consistent results across different datasets.

Our approach is distinct from previous efforts to build cell type classifiers in three key ways: (1) we fully leverage the hierarchical nature of the cell ontology labels during training; (2) we output a probability distribution over the hierarchical terms in the cell ontology; and (3) we leverage cell functional annotations to predict both cell functions and cell types. CellO[2] outputs hierarchical cell type probabilities but was trained on bulk RNA-seq data. OnClass[3] predicts cell ontology annotations and can predict labels for unseen cell types but does not fully utilize the hierarchical structure during training. Similarly, MARS[4] can predict annotations for

unseen cell types but uses a flat classifier and thus does not fully utilize the hierarchical information either. None of these approaches use the functional annotations of the cell ontology.

Hierarchical Prediction of Cell Types

We will implement the Marginalization Classification (MC) approach[5]. In a previous study, when compared to a variety of other models that incorporate hierarchically organized labels, this model performed the best[5]. The MC approach returns a normalized probability distribution across the final level of the hierarchy (leaf nodes), then calculates internal node probabilities as the sums of the probabilities of children nodes (**Fig. 1c**). Importantly, both internal and leaf node probabilities are compared with the true labels during training, ensuring that the learned probability distribution respects conditional independence relationships encoded in the hierarchical structure. Given that the cell ontology is a directed acyclic graph (DAG) rather than a tree, with multiple inheritance (**Fig. 1b**), we extend the MC approach to DAGs. This involves identifying the set of unique leaf nodes descending from each internal node and summing the probabilities of the leaves in this set (**Fig. 1d**). This modification ensures that internal node probabilities reflect the likelihood of their descendant leaf nodes without exceeding 1.

The MC approach is compatible with many different types of classifiers. We will explore deep learning approaches using both fully-connected layers and self-attention layers. The loss function is a sum of leaf loss and internal node loss. We use cross-entropy for leaf nodes (a multi-class, single-label classification problem) and binary cross-entropy loss for internal nodes (a multi-class, multi-label classification problem). We evaluate the performance of the model with leaf Accuracy, internal Accuracy, and Macro-F1 Score.

We will train McCells on all data in the census that meets our filtering criteria. For example, we plan to remove cell types whose only label is more than 5 levels above the leaf level of the Cell Ontology. We will also focus on cell types with sufficient scRNA-seq observations initially and remove epigenomic and spatial modalities.

When applied to new datasets, the output will be the cell type furthest down the Cell Ontology that is predicted with a probability above a threshold user-defined threshold (say, 0.9). Users can also obtain per-cell or per-cluster probabilities for the whole ontology and visualize them on a hierarchical diagram.

In preliminary analysis, we tested our approach using a subset of the Census cells that have Hematopoietic Cell as a parent node (**Fig. 2**). Using a simple MLP neural net architecture, we achieved high accuracy and macro-F1 score on the leaf nodes (**Fig. 2a-b**). We evaluated the model on an in-house dataset from human bone marrow cells and found that it gave good predictions with well-calibrated probabilities (**Fig. 2c-d**). For example, the cluster we had annotated as “MEP” had highest probability for the cell ontology term “megakaryocyte-erythroid progenitor” and “myeloid lineage restricted progenitor” (**Fig. 2d**).

After the tool is built and trained on this dataset, we will train it with the whole Census dataset that meets our quality control conditions. We envision that McCells could be integrated into CELLxGENE to aid in cell type labeling.

Hierarchical Prediction of Cell Functions

A relatively under-explored aspect of the cell ontology is that it describes cells in terms of their functional properties. For example, the cell ontology term “macrophage” (CL:0000235) is equivalent to (myeloid leukocyte and CAPABLE OF antigen processing and presentation of peptide or polysaccharide antigen via MHC class II and CAPABLE OF phagocytosis and CAPABLE OF pseudopodium organization). The functions that cells are CAPABLE OF belong in turn to the gene ontology and are hierarchically organized, and cell types inherit functional terms from their parents. Thus, cells labeled with cell ontology labels are also labeled with rich

hierarchical functional annotations, which makes it possible to train a classifier to predict the functions that cells can perform. To our knowledge, our approach would be the first to make use of these functional annotations in the context of scRNA-seq data. Upon training, users can explore the probabilities of specific or general cell functions for individual cells or clusters in a hierarchical fashion. This facet of our proposal holds significant potential, as predicting the functions of cells based on their ontological annotations can offer valuable insights, particularly when exploring properties of cells from new species, disease contexts, or perturbation studies.

Utility (312/500 words):
Defining the cellular “parts list” of the human body is a foundational task for biomedical science. A human cell atlas promises to advance our understanding of cell biology, development, the physiology of normal tissue function, and the molecular basis of disease. A key deliverable of cell atlas projects is an operational definition of cell type that allows the automated labeling of cells in new datasets. So far, this goal is largely aspirational and single-cell annotation remains a painfully manual task.

While many classification models have been trained to enable labeling single-cell data[6–9], insufficient training data and inconsistent labels have limited the utility of these approaches. These two problems are related, because the highly manual and decentralized nature of cell annotation has led to widely differing nomenclature and level of detail in cell labels. Converting raw outputs of single-cell transcriptomic data into a usable and actionable format can be a time-consuming task. Traditionally, unsupervised learning is used to group cell types, which must then be manually classified into a cell type. But the nuances and personal choices inherent in this process mean there is little standardization between different research groups [4].

Our tool will streamline and automate the process of annotating new single cell datasets. We will make it robust to accepting and processing new datasets from various sources. We will integrate our tool with LIGER, Seurat, and scanpy packages so that users can classify cells directly on their dataset objects. We anticipate that our tool will turn a process that can take days or weeks into a process that will take minutes or hours. There will always be a need for confirming the annotations, and our tool will be designed in such a way to make this as seamless a process as possible. Finally, our tool will provide a confidence level for identified cell types that will enable future research on high-confidence datasets.

Additionally, we anticipate applying this tool to the entire Census dataset that did not meet our quality threshold for training to re-label all cell types. This will improve quality control across the Census by having consistently applied labels, at an appropriate level in the ontology. We plan to regularly re-train and update the tool every six months following every new Long-term supported (LTS) Census release.

We will make the tool available as a python package through PyPI and scverse.

Data: CELL x GENE Census

Milestones and Deliverables (250 words):

DEI Plan (250 words):

We strongly believe in the importance of distributing software tools that are user-friendly and reproducible. We will develop a well-documented R or Python package with a manual (R vignette or Python ReadTheDocs page) and post it on the appropriate language-specific repository (CRAN or Bioconductor for R, PyPI for Python). To

facilitate reproducibility and reuse, we will provide RMarkdown or Jupyter Notebook tutorials and additional notebooks showing how we performed the analyses of spatial transcriptomic datasets. We will also apply to offer a tutorial, presentation, or poster on single-cell analysis of cell-cell signaling at a leading bioinformatics conference such as RECOMB, ISMB, PSB, or ACM-BCB.

Community Statement (250 words):
Our work develops fundamental analytical tools with very broad and important implications. We anticipate that a general model for predicting cell types and functions will have very broad implications for research into cell differentiation and development, the physiology of normal tissue function, and disease. Thus, our work both builds on and has the potential to amplify and enrich the utility of many single-cell datasets funded by CZI. We are committed to building, applying, and disseminating tools for identifying cell types and cellular properties, and are excited about the potential applications of these tools within the CZI network and the broader scientific community.

References:

1) Bard, J., Rhee, S. Y., & Ashburner, M. (2005). An ontology for cell types. Genome biology, 6(2), R21. https://doi.org/10.1186/gb-2005-6-2-r21
2) Dhall, A., Makarova, A., Ganea, O., Pavllo, D., Greeff, M., and Krause, A., "Hierarchical Image Classification using Entailment Cone Embeddings", arXiv e-prints, 2020. doi:10.48550/arXiv.2004.03459.
3) CZ CELLxGENE Discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. CZI Single-Cell Biology, Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M. Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, J. Michael Cherry, Tiffany Chi, Jennifer Chien, Leah Dorman, Nayib Gloria, Pablo Garcia-Nieto, Mim Hastie, Daniel Hegeman, Jason Hilton, Timmy Huang, Amanda Infeld, Ana-Maria Istrate, Ivana Jelic, Kuni Katsuya, Yang Joon Kim, Karen Liang, Mike Lin, Maximilian Lombardo, Bailey Marshall, Bruce Martin, Fran McDade, Colin Megill, Nikhil Patel, Alexander Predeus, Brian Raymor, Behnam Robatmili, Dave Rogers, Erica Rutherford, Dana Sadgat, Andrew Shin, Corinn Small, Trent Smith, Prathap Sridharan, Alexander Tarashansky, Norbert Tavares, Harley Thomas, Andrew Tolopko, Meghan Urisko, Joyce Yan, Garabet Yeretssian, Jennifer Zamanian, Arathi Mani, Jonah Cool, Ambrose Carr. bioRxiv 2023.10.30.563174; doi: https://doi.org/10.1101/2023.10.30.563174
4) A comprehensive mouse kidney atlas enables rare cell population characterization and robust marker discovery. Claudio Novella-Rausell, Magda Grudniewska, Dorien J. M. Peters, Ahmed Mahfouz. bioRxiv 2022.07.02.498501; doi: https://doi.org/10.1101/2022.07.02.498501
    a) Now published in iScience doi: 10.1016/j.isci.2023.106877. https://www.biorxiv.org/content/10.1101/2022.07.02.498501v1.full

Previous cell type annotation methods:

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1862-5
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6791524/
https://www.celltypist.org/ (https://doi.org/10.1101/2023.05.01.538994, https://doi.org/10.1126/science.abl5197)
https://www.sciencedirect.com/science/article/pii/S2001037021000192
https://www.nature.com/articles/s41467-021-25725-x
https://pubmed.ncbi.nlm.nih.gov/33077966/
https://www.nature.com/articles/s41592-022-01651-8

Figures (2 pages)

Diagrams in this Slide Deck for Easy Editing

Other Figures/Images linked to lower in this document

- **Figure 1:**
  - **Diagram of MC approach - DRAFT ON SLIDE 1**
  - **Summed probabilities - label with cell types - DRAFT ON SLIDE 2**
  - **Example parts of Cell Ontologies - show hierarchical relationships - from OWL website - get screen grab**
- **Figure 2:**
  - **Accuracy and F1 scores - make into line plots**
  - **UMAP embedding of output layer for whole dataset**
    - **Color by true cell type label**
  - **Umap of Chen's dataset colored by his labels**
  - **Umap of Chen's dataset with our predicted labels**
  - **Pick a cluster from Chen's dataset - show diagram of cell ontology with probabilities - average all cells from cluster - DRAFTS ON SLIDE 3 and 4**

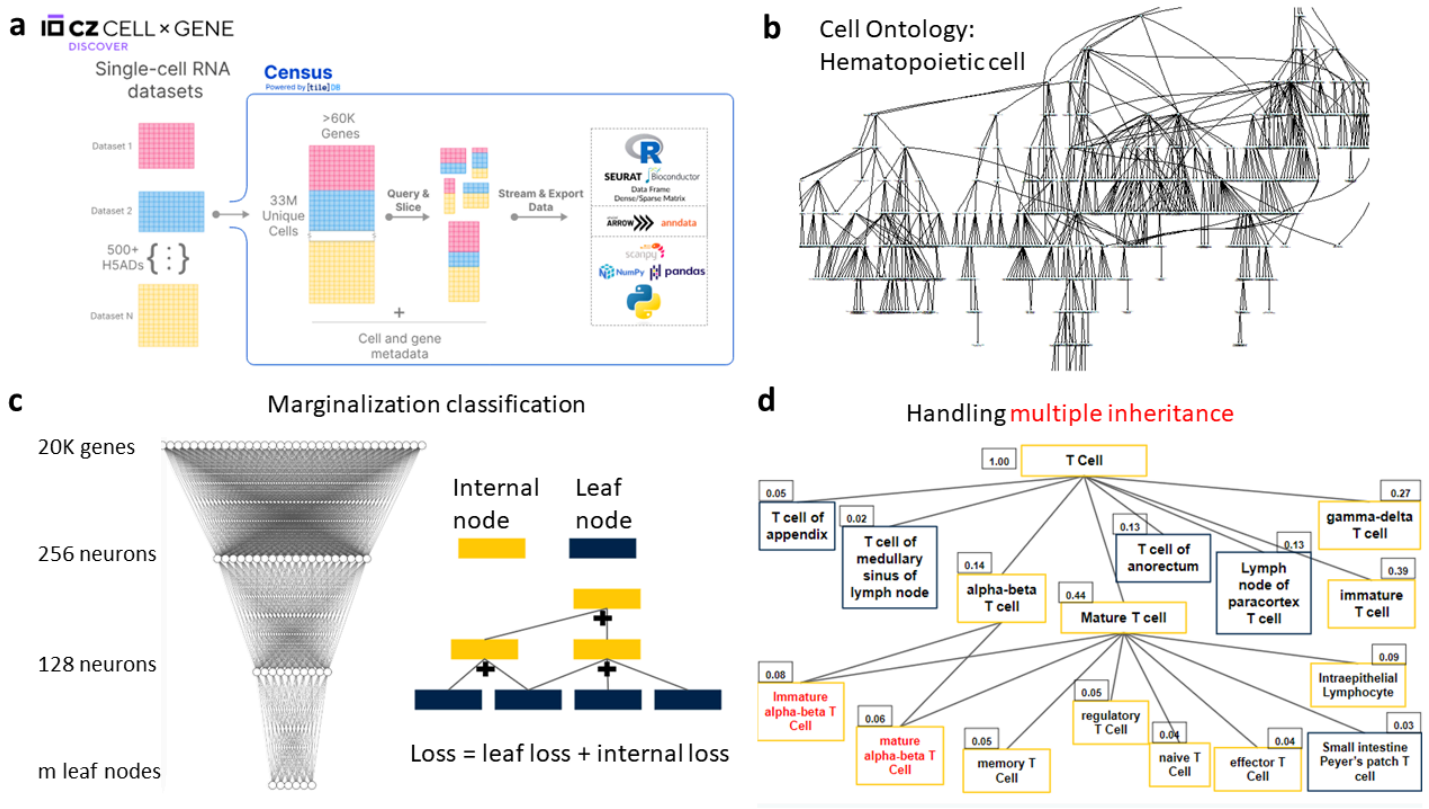


Figure 1: Diagram of Marginalization Classification for Cells. a. We will leverage the CZI CELL x GENE Census dataset to train a hierarchical classification model called Marginalization Classification for Cells (McCells) that predicts cell type and function. The dataset contains many millions of cells labeled with cell ontology terms, which include functional annotations and are arranged in a hierarchy. **b**. Structure of cell ontology terms that are descendants of “Hematopoietic Cell” (CL:0000988). Note the complexity of the structure, including multiple inheritance, making it a directed acyclic graph rather than a simple tree. **c**. Diagram of marginalization classification approach. We train a classification model (such as a multilayer perceptron, shown on the left, or neural network with self-attention layers) to directly output leaf node probabilities and calculate internal node probabilities by recursively summing the probabilities of descendants. Importantly, the network is trained to minimize a loss that compares the predicted probabilities to true labels for both leaf nodes and internal nodes. **d**. To extend the marginalization classification approach to labels with DAG structures, we must account for multiple inheritance to avoid internal nodes with probability exceeding 1. We do this by summing the probabilities for unique leaf nodes descended from each internal node, so that no leaf contributes multiple times to any given internal node. An example of this strategy is shown for a sub-graph of the cell ontology that contains multiple inheritance.

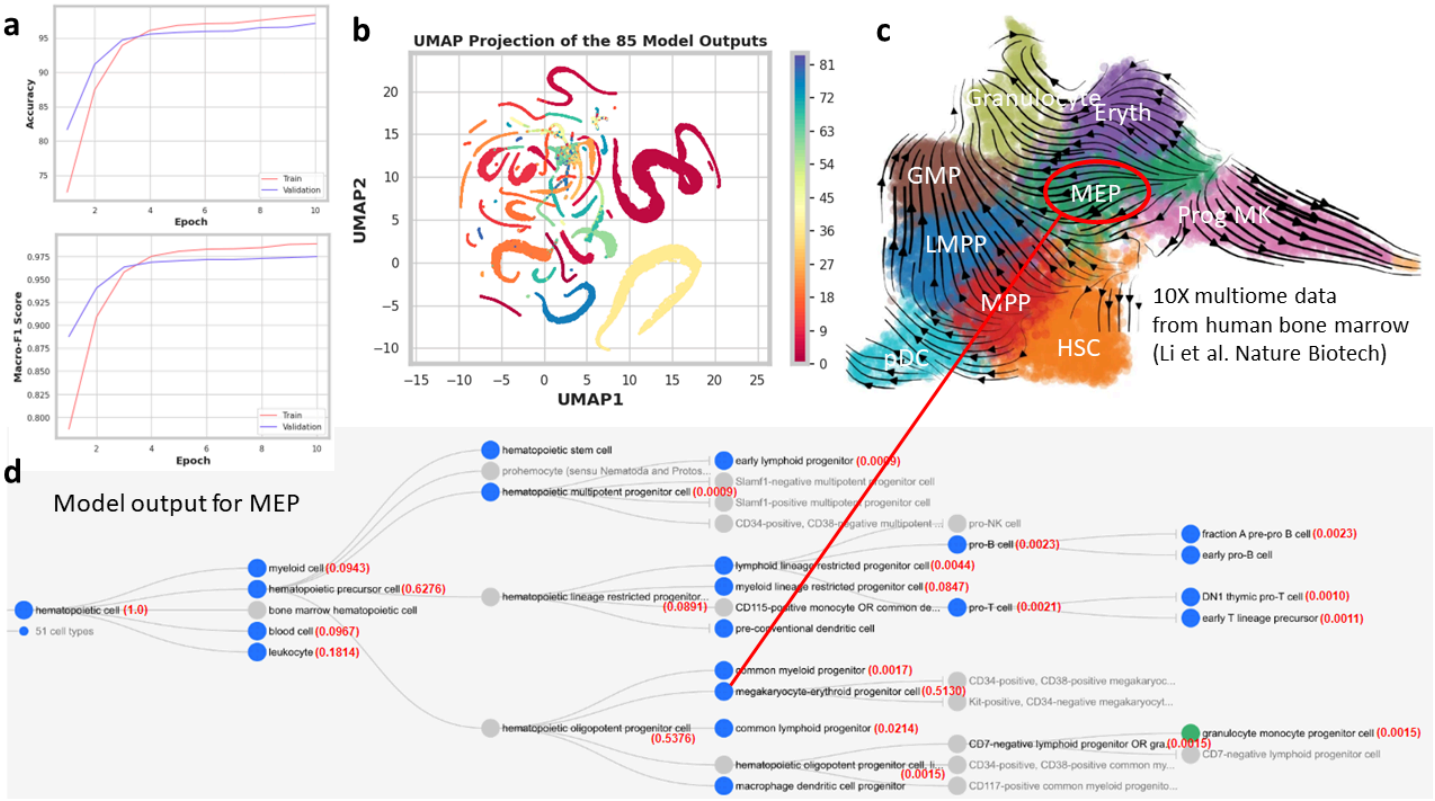


Figure 2: Preliminary Results on Hematopoietic Cells Sampled from Census. a. Line plots of accuracy and macro-F1 for a marginalization classification model trained for 10 epochs on a stratified subset of 392K hematopoietic cells sampled from the Census dataset. We used a multilayer perceptron architecture with 2 hidden layers. The subset contained 85 leaf node terms and 54 internal node terms and was balanced to reflect the same cell type proportions as the full 2.5 million cell dataset (restricted to human, 10X v3 3' scRNA-seq). **b**. UMAP projection of 85 leaf node probabilities, colored by true label. Note that the plot looks "stringy" because the neuron activations have been passed through a softmax layer, making values similar for cells within each type. **c**. UMAP and velocity stream plot of held-out dataset (not in the Census) that we used to test the trained model. Data is human bone marrow cells sequenced with 10X multiome (single-nucleus RNA-seq and ATAC-seq) from our recent publication, Li et al. Nature Biotechnology 2023. We predicted cell type labels for each cell in this dataset. **d**. Example output for the cells we manually annotated as "MEP". Predicted probabilities for each cell ontology term are shown in red. Each node is a cell ontology term; edges represent inheritance; blue nodes are terms with cell observations; and grey nodes are unobserved terms. Note that descendants are not shown for some of the terms to save space.

References

1. Diehl, A. D. *et al.* The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J. Biomed. Semantics* **7**, 44 (2016).

2. Bernstein, M. N., Ma, Z., Gleicher, M. & Dewey, C. N. CellO: comprehensive and hierarchical cell type classification of human cells with the Cell Ontology. *iScience* **24**, 101913 (2021).

3. Wang, S. *et al.* Leveraging the Cell Ontology to classify unseen cell types. *Nat. Commun.* **12**, 5556 (2021).

4. Brbić, M. *et al.* MARS: discovering novel cell types across heterogeneous single-cell experiments. *Nat. Methods* **17**, 1200–1206 (2020).

5. Dhall, A. *et al.* Hierarchical Image Classification using Entailment Cone Embeddings. *arXiv [cs.CV]* (2020) doi:10.48550/ARXIV.2004.03459.

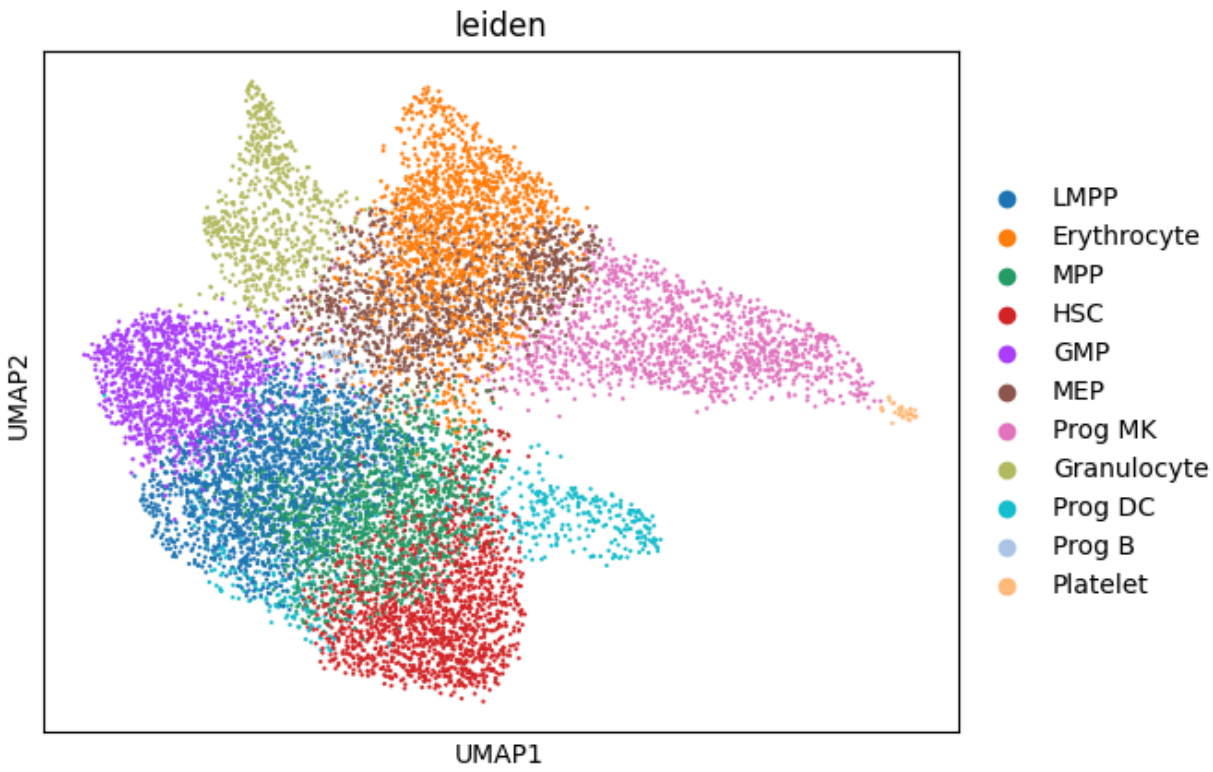
6. Brbić, M. *et al.* Annotation of spatially resolved single-cell data with STELLAR. *Nat. Methods* **19**, 1411–1418 (2022).

7. Pasquini, G., Rojo Arias, J. E., Schäfer, P. & Busskamp, V. Automated methods for cell type annotation on scRNA-seq data. *Comput. Struct. Biotechnol. J.* **19**, 961–969 (2021).

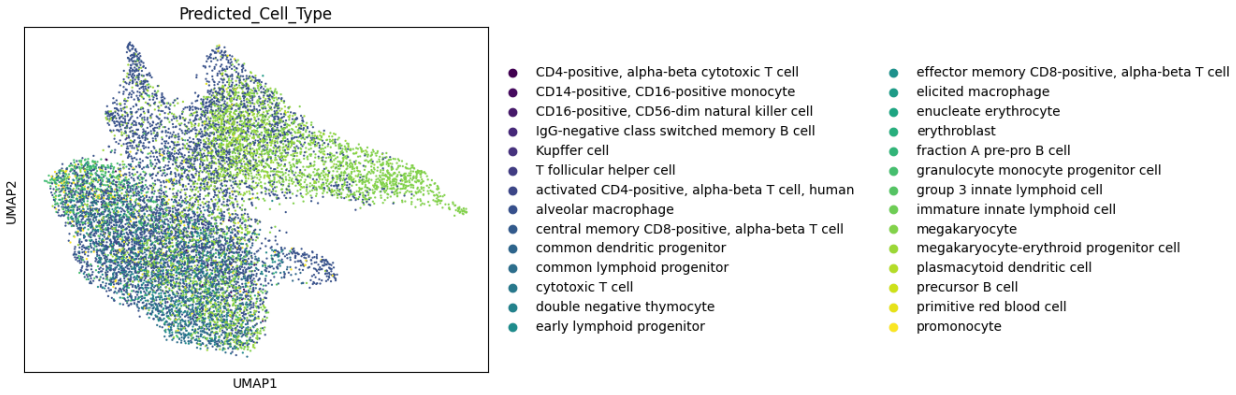
8. Xu, C. *et al.* Automatic cell type harmonization and integration across Human Cell Atlas datasets. *bioRxiv* 2023.05.01.538994 (2023) doi:10.1101/2023.05.01.538994.

9. Pliner, H. A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *Nat. Methods* **16**, 983–986 (2019).

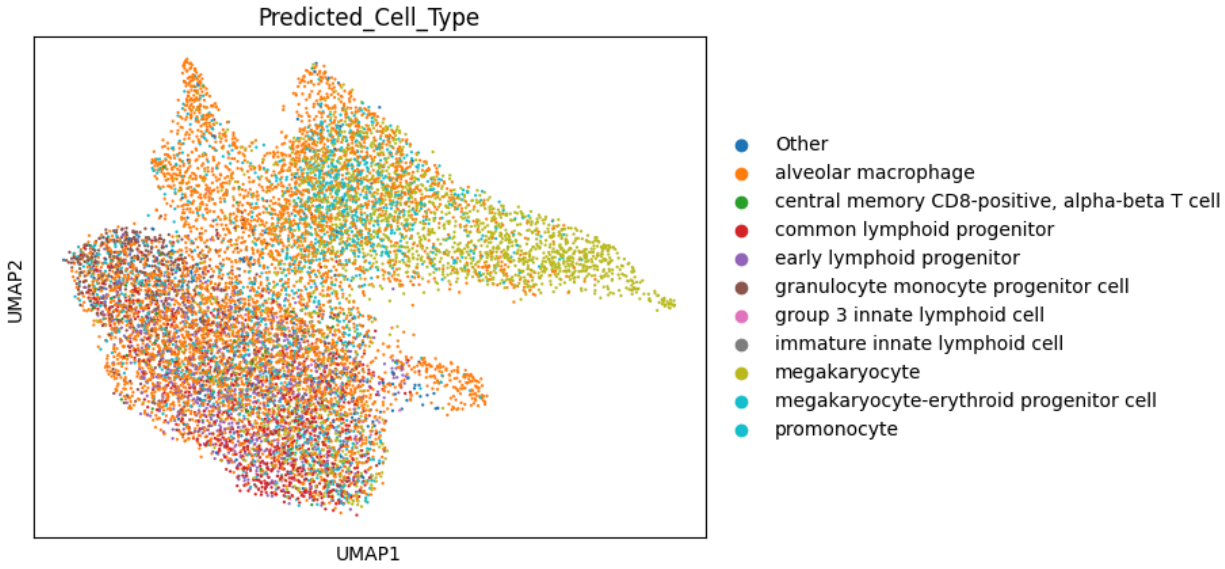
UMAP from Chen’s Dataset with his labels:

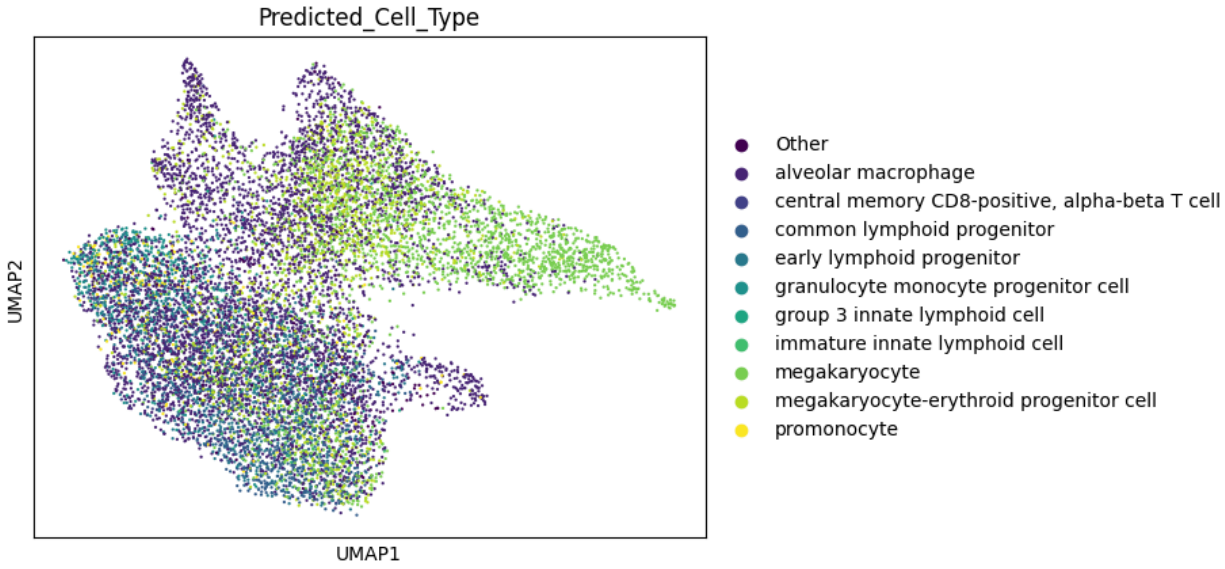


UMAP of Chen’s dataset with our predicted outputs

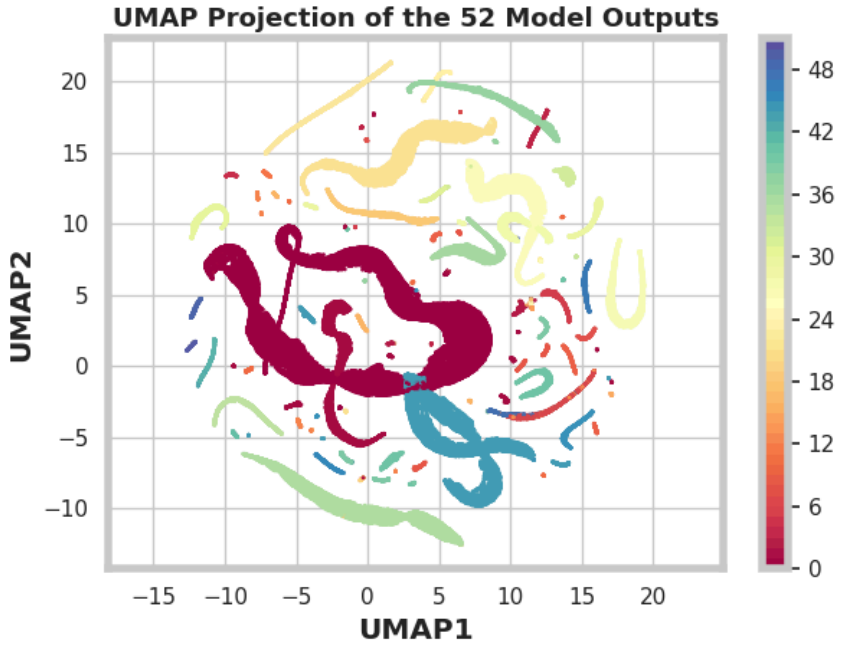


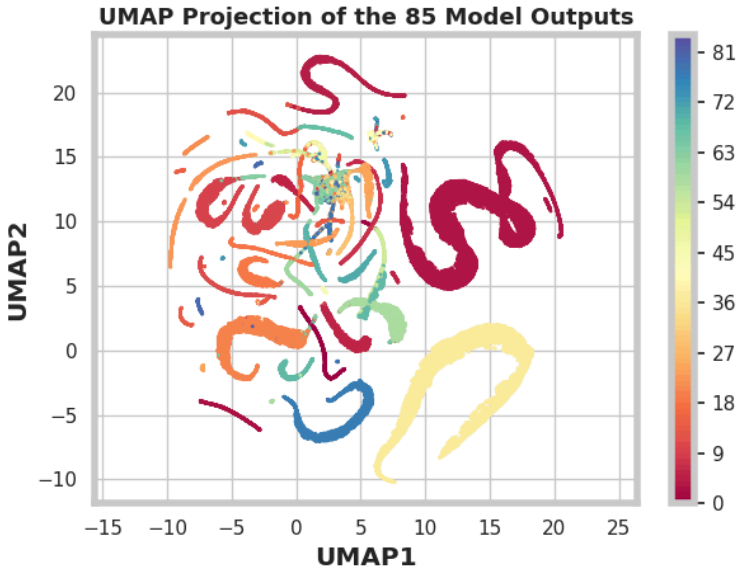
Top 10 categories + other

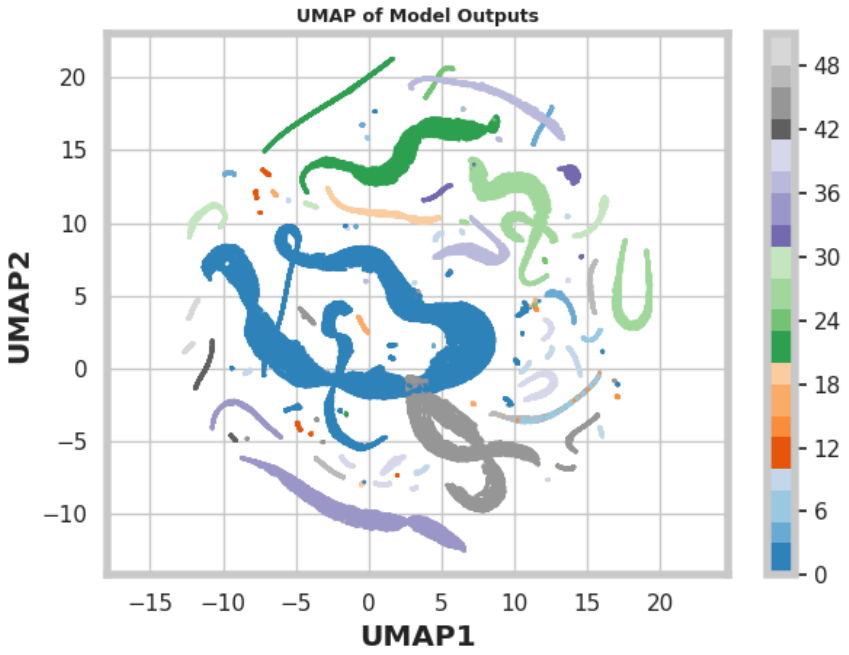



Predicted_Cell_Type
UMAP2
UMAP1
Other
alveolar macrophage
central memory CD8-positive, alpha-beta T cell
common lymphoid progenitor
early lymphoid progenitor
granulocyte monocyte progenitor cell
group 3 innate lymphoid cell
immature innate lymphoid cell
megakaryocyte
megakaryocyte-erythroid progenitor cell
promonocyte

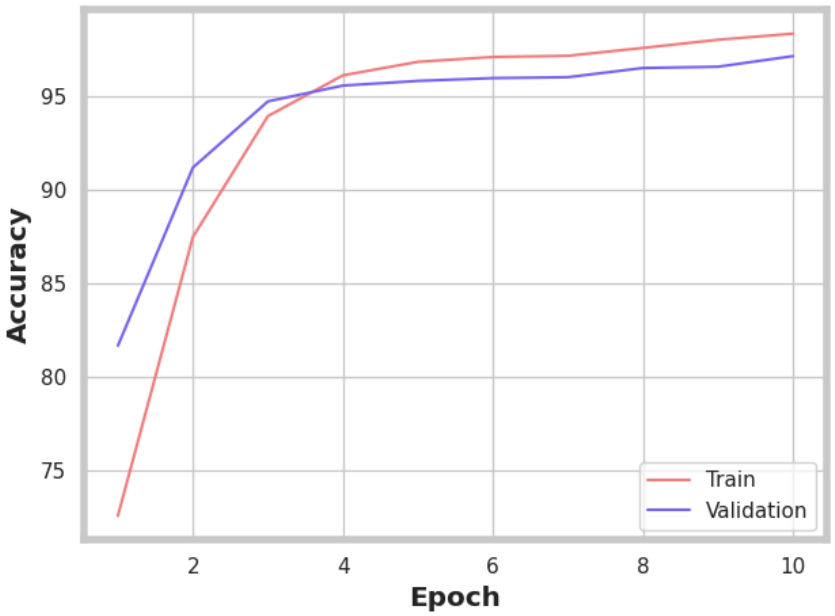
UMAP of Model Outputs



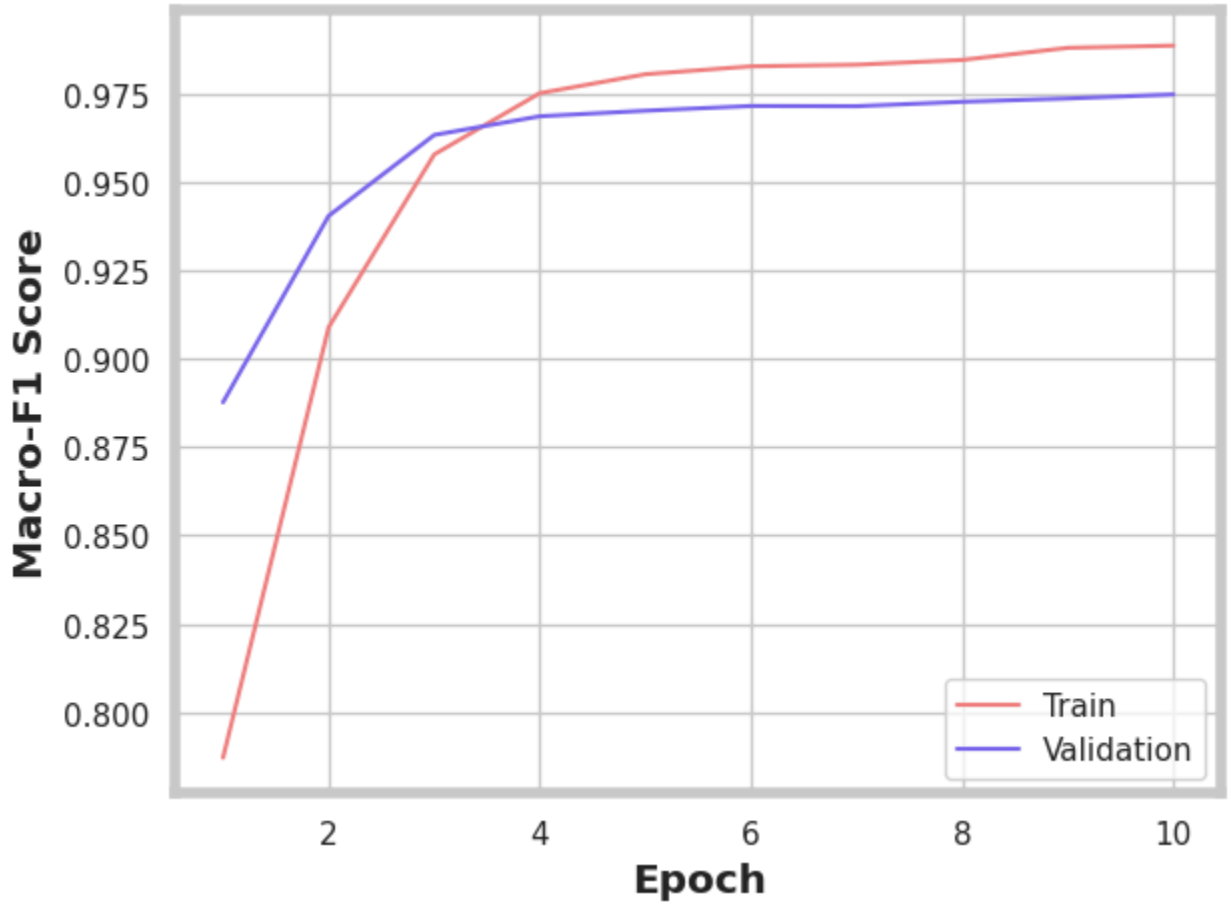



UMAP of Model Outputs
UMAP2
UMAP1
20
15
10
5
0
−5
−10
−15
−10
−5
0
5
10
15
20
48
42
36
30
24
18
12
6
0

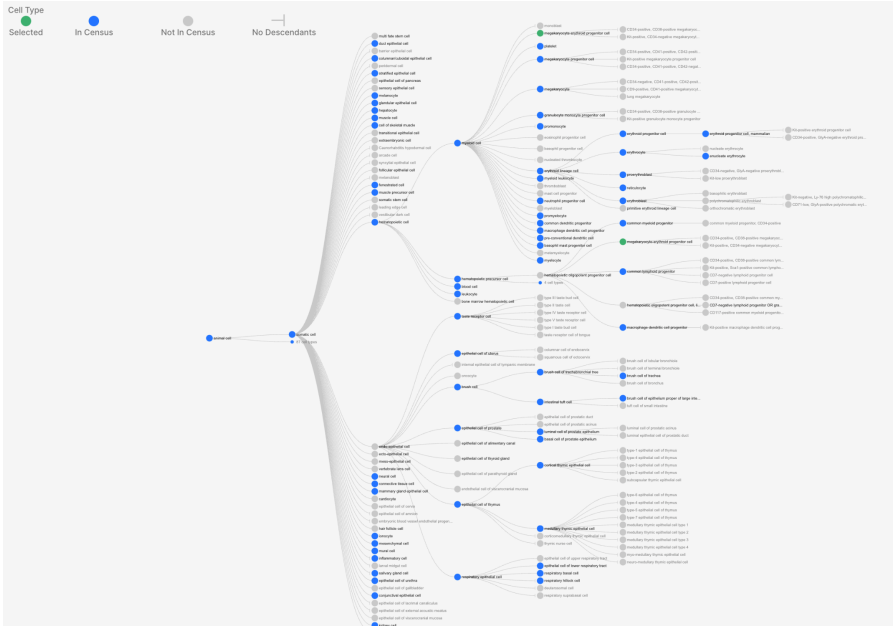
Accuracy of Current Model:



Macro-F1 Score:



Example Part of Cell Ontology: Screenshot from [CellXGene](). Depending on the shape of the image we want, we could change which cells are displayed.



Here's a slightly less busy version, but it also doesn't show as many layers.