JOINT EMBEDDING OF TRANSCRIPTOMES AND TEXT ENABLES INTERACTIVE SINGLE-CELL RNA-SEQ DATA EXPLORATION VIA NATURAL LANGUAGE

Moritz Schaefer [*, †,‡] mschaefer@cemm.at
Peter Peneder [*, §,¶] peter.peneder@ccri.at
Daniel Malzl [‡,‖,**] daniel@menchelab.com
Anna Hakobyan [‡,§] anna.hakobyan@ccri.at
Varun Sharma [†,‡,‖, **] vsharma@cemm.at
Thomas Krausgruber [†,‡] tkrausgruber@cemm.at
Jörg Menche [††,‖,**,‡‡,‡] joerg.menche@univie.ac.at
Eleni M. Tomazou [§] eleni.tomazou@ccri.at
Christoph Bock [†,‡] cbock@cemm.at

ABSTRACT

Single-cell RNA sequencing (scRNA-seq) has revolutionized our understanding of cellular states, but interpreting the vast data it generates remains challenging. Here, we introduce CellWhisperer, a multimodal machine learning model that bridges the gap between transcriptomics data and natural language, enabling intuitive interaction with scRNA-seq datasets. Trained on the bulk RNA-seq data for over 650,000 samples and their textual annotations from the Gene Expression Omnibus (GEO), CellWhisperer employs contrastive learning to create a joint embedding space, enabling tasks such as cell retrieval based on free-text queries and zero-shot classification of cell types. We show that these abilities extend to scRNA-seq datasets with a broad range of cell types. Integrated into the CELLxGENE browser, this allows biologists to explore and label single-cell transcriptomes using natural language queries. Our experiments show that CellWhisperer can accurately annotate cellular states, beyond standard cell types, without relying on reference datasets. This work paves the way for accessible and nuanced interpretations of scRNA-seq data, including those that are poorly covered by reference data, leveraging the power of natural language in transcriptomics research.

1 INTRODUCTION

Single-cell transcriptomes, i.e. the expression levels of most or all genes in a given cell, provide a rich representation of cellular states and are obtainable in high throughput at low cost. In recent years, single-cell RNA sequencing (scRNA-seq) has emerged as a powerful and widely used assay to dissect transcriptional cell states at an immense scale, with efforts to capture and annotate comprehensive sets of organismal cell states such as the Human Cell Atlas (Regev et al., 2018). Yet, there is a significant annotation gap, and the vast amount of biological information captured by scRNA-seq experiments is challenging to interpret. Transferring annotations from high-quality, expert-annotated

[*]These authors contributed equally.
[†]Institute of Artificial Intelligence, Center for Medical Data Science, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria
[‡]CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria
[§]St. Anna Children's Cancer Research Institute (CCRI), Vienna, Austria
[¶]Doctoral School in Microbiology and Environmental Science, University of Vienna, Vienna, Austria
[‖]Max Perutz Labs, Vienna, Austria
[**]Department of Structural and Computational Biology, Center for Molecular Biology, University of Vienna, Vienna, Austria
[††]Ludwig Boltzmann Institute for Network Medicine at the University of Vienna, Vienna, Austria
[‡‡]Faculty of Mathematics, University of Vienna, Vienna, Austria

reference datasets provides an efficient starting point, but is limited by their availability (which is biased toward early-adopting areas of biology) and the kind of annotations that are provided (often only coarse-grained cell types). Bulk RNA-seq, in contrast to scRNA-seq, measures transcriptomes for a sample as a whole, averaging across sample-intrinsic heterogeneity. Bulk RNA-seq datasets have been collected for almost two decades in repositories such as the Gene Expression Omnibus (GEO), which enforce extensive and quality-controlled annotations for each sample, leading to an ever-growing and extremely diverse census of biological samples.

Here, we leverage the entire human part of GEO to train *CellWhisperer*, a multimodal machine learning model that enables dynamic cell annotation by using unconstrained natural language. Specifically, we perform contrastive learning through the CLIP approach (Radford et al., 2021) to learn a joint embedding space from bulk RNA-seq data for over 650,000 samples and their textual annotations. We demonstrate that this enables free-text cell retrieval and cell type classification for scRNA-seq data. To make these capabilities easily accessible, we integrate CellWhisperer into the widely used CELLxGENE browser, enabling interactive analysis of single-cell transcriptomes with natural language in the web browser (early demo available[1]). Our work demonstrates the feasibility of adding natural language queries as a key channel for interacting with scRNA-seq datasets.

2 BACKGROUND

Our work builds on established contrastive learning methods and recent transcriptomics models.

CLIP (Contrastive Language-Image Pretraining) was originally developed to learn visual concepts from natural language supervision, making use of the vast amount of labeled images on the web (Radford et al., 2021). The core idea of CLIP is to use contrastive learning to align the two embeddings generated from the two input modalities. The model architecture therefore consists of two *towers*, which separately embed the two modalities for a given pair, upon which an additional layer brings the two embeddings into the same dimensionality (e.g. 2048). During training, an InfoNCE-based loss maximizes the cosine-similarity between matching transcriptome-text pairs, whereas the cosine-similarity between all other (unmatching) pairs in a given batch is minimized.

Subsequent developments of CLIP have focused on improving training efficiency and data requirements, as the original CLIP model was trained on 400 million data points for 12 days on 592 V100 GPUs (Radford et al., 2021). Specifically, Zhai et al. (2021) initialized both the vision and text models of CLIP with pre-trained weights and fine-tuned only their BERT-based text tower, which improved model performance at drastically reduced training cost.

Whereas CLIP and LiT have focused on the image and text domain, recent works have established foundation models also for the transcriptomics domain. The Geneformer model uses six transformer encoder layers to process individual transcriptomes as "sentences of genes" ranked by their expression. The model was trained on 30 million single-cell transcriptomes and enabled fine-tuning prediction tasks with low data requirements (Theodoris et al., 2023). Equally noteworthy is the scGPT model, which further incorporates expression values, rather than just gene ranks, and is demonstrated to enable numerous cell-level prediction tasks (Cui et al., 2024).

3 METHODS

We design and train a first-of-its-kind multimodal transcriptome-text model by curating a novel dataset with pairs of transcriptomes and natural language annotations from GEO and adopting the CLIP/LiT approach for contrastive learning on the two modalities (see Fig. 1).

3.1 UTILIZING GEO AS A LARGE-SCALE TRANSCRIPTOMICS DATASET WITH NATURAL LANGUAGE ANNOTATIONS

To create our multimodal dataset, we relied on GEO and the NCBI Sequence Read Archive (SRA). These two connected databases provide convenient access to transcriptome profiles together with high-quality textual annotations. Starting from the full set of human SRA samples, we derived

[1] http://cellwhisperer.bocklab.org/

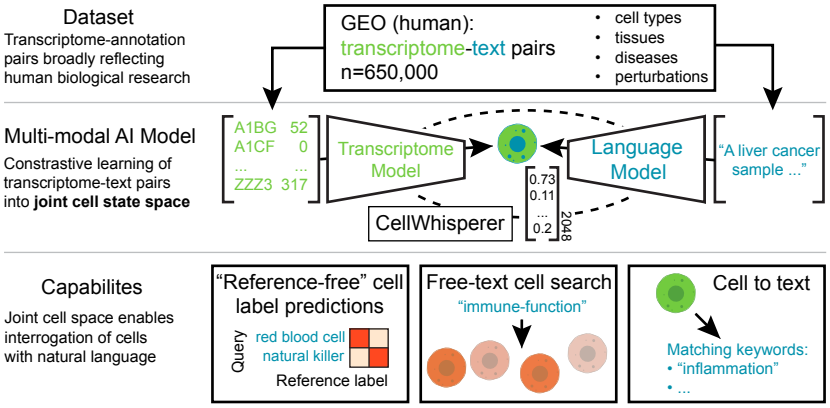


Figure 1: Overview of CellWhisperer, a multimodal model trained on transcriptome-text pairs from GEO that allows interrogations of cells with natural language

a training dataset of 650,000 data points after quality filtering. We also derived a thematically coherent validation set with 15,000 primary tissue samples associated with a disease state, which we used to monitor training and avoid overfitting.

Our transcriptome tower operates on gene-level read counts, which are computed and provided in a standardized manner by Lachmann et al. (2018) for the human part of GEO/SRA. The sample-level annotations were first normalized by mapping their structured metadata to controlled vocabularies using the MetaSRA pipeline (Bernstein et al., 2017). To train a model capable of interacting with natural language, we then employed an LLM, which converted the structured annotations into a natural language form (e.g., *The sample is a Jurkat T cell line with ectopic expression of MALT1-R149A [...]*). See Appendix A for details).

3.2 Design of CellWhisperer, a multimodal transcriptome-text model

To enable interactions with transcriptomic data via natural language, we built a multimodal model named *CellWhisperer*, which consists of two *towers* — a transcriptome model and a text model. The outputs of the two towers are each brought into a 2048-dimensional space through processing via two learnable linear layers, separated by a ReLU nonlinearity, following batch layer normalization as done by Shrivastava et al. (2023). We follow the LiT approach and use pre-trained components for both the transcriptome tower and the text tower. For encoding the transcriptome, we employ the Geneformer model, which has been trained on 30M single-cell transcriptomes (Theodoris et al., 2023). For encoding the textual annotations, we rely on BioBERT (Lee et al., 2020), which has been trained on large biomedical text corpora. We also explored the usage of alternative models, such as scGPT (Cui et al., 2024) for the transcriptome tower and BioGPT (Luo et al., 2022) for the text tower, which led to similar, albeit slightly less stable, results.

This integration of the two modalities enables several ways to interact with transcriptomics data using natural language. First, free-text queries can be embedded alongside the transcriptome readouts from a scRNA-seq experiment. Then, the cosine similarity between two modalities can be computed in a cell-by-cell manner to provide a continuously-valued estimate and ranking of how cells match the query. As originally demonstrated by Radford et al. (2021) in the image domain, this approach can also be leveraged for reference-free zero-shot cell classification, where an arbitrary number of labels can be embedded and compared with a single cell, to rank which of the labels match best.

We integrate these capabilities into the widely used *CELLxGENE* cell browser package (CZI Single-Cell Biology Program et al., 2023), enabling interactive language-based exploration and comprehension of cellular states in the web browser.

3.3 TRAINING OF CELLWHISPERER

As demonstrated first by Zhai et al. (2021), we observed optimal training performance when we initialized the two towers with pre-trained weights, and fine-tuned only the text model with the final shared embedding layers for both towers.

We trained our model with a combination of the InfoNCE-based loss employed by Radford et al. (2021) and the JSD-based loss from Shrivastava et al. (2023), which has been demonstrated to perform favorably in low data regimes. We performed linear learning rate warm-up over the first 10,000 steps to a maximum learning rate of 0.0001. Training was performed on a single A100-80GB GPU over 5 hours with a batch size of 64 with additional gradient accumulation over 32 batches. We used our validation set to select the best model checkpoint during training (see Appendix B).

4 EXPERIMENTS

After training, we first confirmed that CellWhisperer was able to confidently retrieve the matching text label for a given transcriptome from our validation set (ROC-AUC: 0.924 on deduplicated data, see Appendix B for details).

Exploring CellWhisperer's novel capabilities, we generated an atlas of the human transcriptome with cluster-level natural language annotations (see Appendix C and Appendix Fig. 2). This atlas, in combination with CellWhisperer, enables investigation and mapping of arbitrary concepts to human biology, which we demonstrated via a comprehensive number of disease names from the OMIM database (see Appendix C), exhibiting cluster-level variances for potential disease involvements (see Appendix Fig. 3).

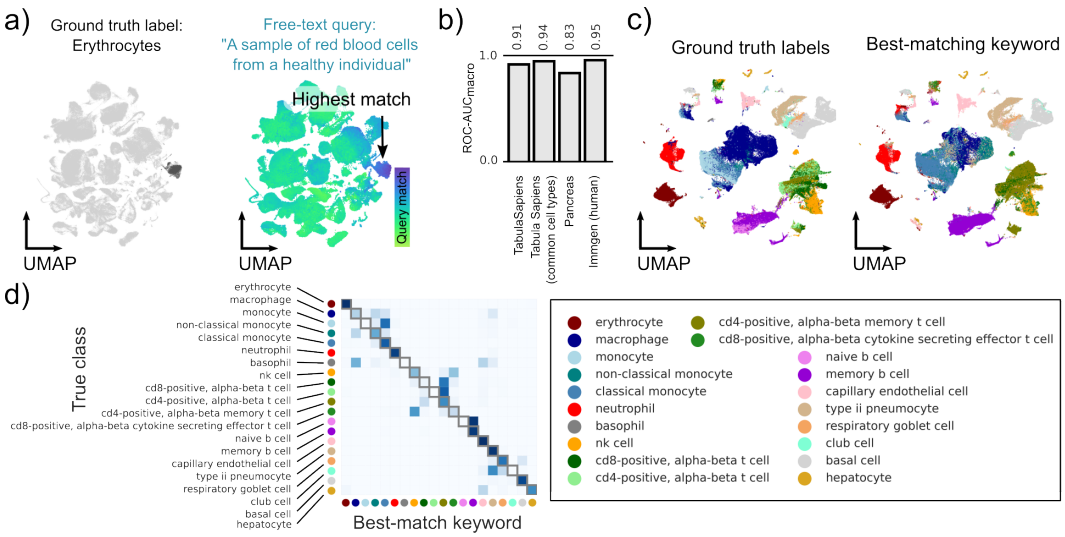


Figure 2: Reference-free zero-shot annotation of single-cell data with CellWhisperer
a) UMAP representation of the Tabula Sapiens dataset, embedded with CellWhisperer (transcriptomes). Cell coloring is indicated by the panel heading, showing the correspondence between a ground truth label and free-text query scores. b) Zero-shot cell type classification performance in evaluation datasets. c,d) UMAPs (c) and confusion matrix (d) comparing the ground-truth cell type labels and best-matching keywords for the subset of 20 common cell types from blood, lung, and liver from Tabula Sapiens.

4.1 CELLWHISPERER ZERO-SHOT ANNOTATES SCRNA-SEQ DATA WITHOUT REFERENCE DATA

It is established that traditional cell type annotations fall short in capturing the full spectrum of states that cells acquire, but generating comprehensive cell-level annotations in an automated fashion is a

major challenge in the field. Language-based multimodal models like CellWhisperer are a promising solution, as they have the flexibility to generate labels at arbitrary granularity.

To evaluate CellWhisperer's ability to generate meaningful annotations for single cells, we tested whether CellWhisperer could match transcriptomes to their corresponding expert-annotated cell type labels, which we provided to CellWhisperer as free-text. This reference-free zero-shot task showcases the model's ability to annotate cells without the need for fine-tuning or user-provided pre-labeled data (in contrast to the present paradigm of fine-tuning transcriptome embedding models or using their embeddings for reference-mapping). We focused our validations on Tabula Sapiens, a comprehensive human cell atlas spanning nearly 500,000 cells from 24 organs. (Tabula Sapiens Consortium et al., 2022) We first tested CellWhisperer's ability to handle free-text queries by scoring single cells by their similarity to the text. For instance, we found that querying cells in the Tabula Sapiens dataset for "A sample of red blood cells from a healthy individual" correctly retrieved cells labeled as erythrocytes (Fig. 2a). We found that this observation generalizes to many other cell types, and quantified the model's performance on all 177 cell types in the Tabula Sapiens dataset, achieving a ROC-AUC score of 0.91. This is especially remarkable given that the model has been trained only on data from GEO, which primarily consists of bulk RNA-seq data.

The representation of cell types in GEO varies, with rare cell types being less well represented than common ones. To assess how this would impact predictive performance, we derived a subset of Tabula Sapiens consisting of cell types that were most common in blood, liver, and lung (see Appendix D). As expected, we observed that those cells could be classified especially well, with many of the mistakes resulting from mix-ups of closely related cell types such as "monocytes" vs "classical monocytes", or subgroups of T-cells (Fig. 2c-d; ROC-AUC = 0.94). Furthermore, the transcriptome latent space spanned by CellWhisperer clustered the cell types in this subset in a well separable manner (Fig. 2c). We also assessed the performance of our model on two further datasets: i) a challenging multi-technique single-cell dataset of cells from the human pancreas (Luecken et al., 2022), in which CellWhisperer achieves a ROC-AUC score of 0.83 (Fig. 2b), and ii) a bulk dataset of human immune cells (ROC-AUC=0.95; Fig. 2b). We further explored whether fine-tuning on scRNA-seq data could enhance our model, and indeed observed improved cell type classification performance (see Appendix E for details).

In summary, CellWhisperer allows cell type annotation with arbitrary labels, providing a proof of principle for leveraging free-text annotations of bulk data for single-cell data analysis.

4.2 CellWhisperer facilitates identification of structural cells exhibiting immune functions through scRNA-seq exploration in the web browser

Having validated our model on annotation tasks for scRNA-seq data, we integrated it into the widely used *CELLxGENE Explorer* scRNA-seq analysis tool, enabling natural language-based analysis of large-scale scRNA-seq datasets without any need for writing analysis code (Fig. 3a, Appendix F)

To demonstrate CellWhisperer's functionality as part of the CELLxGENE web browser, we set out to explore the Tabula Sapiens dataset in an interactive manner. As a relevant example of exploratory research, we aimed at identifying human non-hematopoietic structural cells that display immune functions, which have recently been systematically characterized in mice (Krausgruber et al., 2020). When performing a free-text search via CellWhisperer for "structural cells with immune functions", we observed strong query matches for cells annotated as fibroblast and endothelial cells as well as myeloid cells (Fig. 3b). Fibroblasts are established to mediate immune functions (Davidson et al., 2021), so we focused our analysis on the less explored endothelial cells (Fig. 3c). Filtering for cells labeled as "endothelial cells" in Tabula Sapiens revealed that a subset of them localized in the myeloid cluster. Using CellWhisperer to match them with biological keywords led to high matches for terms containing "macrophage" and "monocyte". Indeed, these cells also strongly expressed macrophage markers LYZ and CD14, so we excluded them from the downstream analysis (Fig. 3d). The remaining endothelial cells exhibited considerable variability in terms of how strongly cells matched the search query (Fig. 3c,e) and we suspected that endothelial cells varied in their propensity for immune function. To test this hypothesis, we selected the top and bottom 1000 cells (with respect to query-matching score) and performed a differential gene expression analysis from which we analyzed the top 50 upregulated genes with Enrichr (Chen et al., 2013). Indeed, the top matches in the KEGG pathway library (Kanehisa et al., 2023) overall pointed toward immune functions,

including terms such as *Antigen-processing and presentation* and the *TNF signaling pathway*, indicating an active role in immunity for a subset of endothelial cells (Fig. 3e).

Taken together, CellWhisperer, in combination with CELLxGENE, enabled us to perform exploratory research leading to novel evidence for the presence of non-hematopoietic structural cells with immune functions in the human endothelium.

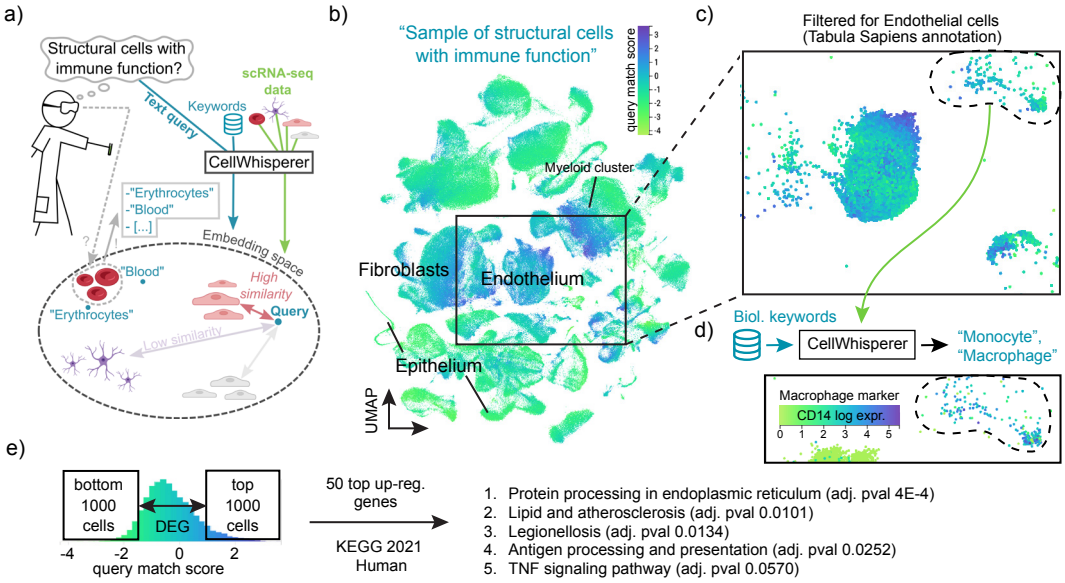


Figure 3: Interactive scRNA-seq analysis with CellWhisperer in the web browser identifies immune-related endothelial cells in human.
a) Overview of CellWhisperer's capabilities for scRNA-seq exploration: Scoring cells by similarity to a user query and describing user-selected cells. b) UMAP of Tabula Sapiens on the CellWhisperer transcriptome embedding. Cells are highlighted by their similarity to the embedded search query. Cluster labels are based on Tabula Sapiens annotations. c) Filtering for "endothelial cells" (Tabula Sapiens annotation). d) CellWhisperer and macrophage marker expression identify a subset of endothelial-labeled cells as myeloid cells. e) Differential Gene Expression (DEG) analysis of top 1000 vs. bottom 1000 free-text query-matching cells (from (b,c)), followed by enrichment analysis. Myeloid-identified cells were excluded. The top 5 resulting terms are shown on the right.

5 Discussion

Building on the success of multimodal contrastive learning in visual domains (Radford et al., 2021), we applied this approach to transcriptomics with our model, CellWhisperer. Utilizing public transcriptomics data, CellWhisperer connects gene expression profiles with natural language, thus facilitating intuitive data analysis. The performance demonstrated in our validations suggests that natural language could soon become a primary interface for exploring cellular states. However, our model also exposes limitations in its predictions, for example caused by biases in the training data. Future enhancements therefore include the incorporation of more varied data and the refinement of data processing techniques. We also anticipate the development of conversational interfaces for discussing transcriptomic data, potentially through CLIP-based LLM-models models like LLaVA (Liu et al., 2023a). Through such an LLM integration, CellWhisperer could eventually even function as an agent, conversing with more sophisticated language models such as GPT-4, and thereby enabling complex, semi-automated transcriptomics analyses and integration with broader scientific resources (Liu et al., 2023b).

REFERENCES

Matthew N Bernstein, Anhai Doan, and Colin N Dewey. MetaSRA: normalized human sample-specific metadata for the sequence read archive. *Bioinformatics*, 33(18):2914–2923, September 2017. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btx334. URL http://dx.doi.org/10.1093/bioinformatics/btx334.

Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R Clark, and Avi Ma'ayan. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics*, 14:128, April 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-128. URL http://dx.doi.org/10.1186/1471-2105-14-128.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, pp. 1–11, 2024.

CZI Single-Cell Biology Program, 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, Pablo Garcia-Nieto, Nayib Gloria, 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, and Ambrose Carr. CZ CELL×GENE discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. November 2023. URL https://www.biorxiv.org/content/10.1101/2023.10.30.563174.abstract.

Sarah Davidson, Mark Coles, Tom Thomas, George Kollias, Burkhard Ludewig, Shannon Turley, Michael Brenner, and Christopher D Buckley. Fibroblasts as immune regulators in infection, inflammation and cancer. *Nature reviews. Immunology*, 21(11):704–717, November 2021. ISSN 1474-1733, 1474-1741. doi: 10.1038/s41577-021-00540-z. URL http://dx.doi.org/10.1038/s41577-021-00540-z.

Philip A Ewels, Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso, and Sven Nahnsen. The nf-core framework for community-curated bioinformatics pipelines. *Nature biotechnology*, 38(3):276–278, March 2020. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-020-0439-x. URL http://dx.doi.org/10.1038/s41587-020-0439-x.

Tracy S P Heng, Michio W Painter, and Immunological Genome Project Consortium. The immunological genome project: networks of gene expression in immune cells. *Nature immunology*, 9 (10):1091–1094, October 2008. ISSN 1529-2908, 1529-2916. doi: 10.1038/ni1008-1091. URL http://dx.doi.org/10.1038/ni1008-1091.

Minoru Kanehisa, Miho Furumichi, Yoko Sato, Masayuki Kawashima, and Mari Ishiguro-Watanabe. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic acids research*, 51(D1): D587–D592, January 2023. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkac963. URL http://dx.doi.org/10.1093/nar/gkac963.

Thomas Krausgruber, Nikolaus Fortelny, Victoria Fife-Gernedl, Martin Senekowitsch, Linda C Schuster, Alexander Lercher, Amelie Nemc, Christian Schmidl, André F Rendeiro, Andreas Bergthaler, and Christoph Bock. Structural cells are key regulators of organ-specific immune responses. *Nature*, 583(7815):296–302, July 2020. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-020-2424-4. URL http://dx.doi.org/10.1038/s41586-020-2424-4.

Alexander Lachmann, Denis Torre, Alexandra B Keenan, Kathleen M Jagodnik, Hoyjin J Lee, Lily Wang, Moshe C Silverstein, and Avi Ma'ayan. Massive mining of publicly available RNA-seq data from human and mouse. *Nature communications*, 9(1):1366, April 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-03751-6. URL http://dx.doi.org/10.1038/s41467-018-03751-6.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, February 2020. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btz682. URL http://dx.doi.org/10.1093/bioinformatics/btz682.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. April 2023a. URL http://arxiv.org/abs/2304.08485.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang, Jianfeng Gao, and Chunyuan Li. LLaVA-Plus: Learning to use tools for creating multimodal agents. November 2023b. URL http://arxiv.org/abs/2311.05437.

Malte D Luecken, M Büttner, K Chaichoompu, A Danese, M Interlandi, M F Mueller, D C Strobl, L Zappia, M Dugas, M Colomé-Tatché, and Fabian J Theis. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, 19(1):41–50, January 2022. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-021-01336-8. URL http://dx.doi.org/10.1038/s41592-021-01336-8.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6), November 2022. ISSN 1467-5463, 1477-4054. doi: 10.1093/bib/bbac409. URL http://dx.doi.org/10.1093/bib/bbac409.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. Orca 2: Teaching small language models how to reason. November 2023. URL http://arxiv.org/abs/2311.11045.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Aviv Regev, Sarah Teichmann, Orit Rozenblatt-Rosen, Michael Stubbington, Kristin Ardlie, Ido Amit, Paola Arlotta, Gary Bader, Christophe Benoist, Moshe Biton, Bernd Bodenmiller, Benoit Bruneau, Peter Campbell, Mary Carmichael, Piero Carninci, Leslie Castelo-Soccio, Menna Clatworthy, Hans Clevers, Christian Conrad, Roland Eils, Jeremy Freeman, Lars Fugger, Berthold Goettgens, Daniel Graham, Anna Greka, Nir Hacohen, Muzlifah Haniffa, Ingo Helbig, Robert Heuckeroth, Sekar Kathiresan, Seung Kim, Allon Klein, Bartha Knoppers, Arnold Kriegstein, Eric Lander, Jane Lee, Ed Lein, Sten Linnarsson, Evan Macosko, Sonya MacParland, Robert Majovski, Partha Majumder, John Marioni, Ian McGilvray, Miriam Merad, Musa Mhlanga, Shalin Naik, Martijn Nawijn, Garry Nolan, Benedict Paten, Dana Pe'er, Anthony Philippakis, Chris Ponting, Steve Quake, Jayaraj Rajagopal, Nikolaus Rajewsky, Wolf Reik, Jennifer Rood, Kourosh Saeb-Parsy, Herbert Schiller, Steve Scott, Alex Shalek, Ehud Shapiro, Jay Shin, Kenneth Skeldon, Michael Stratton, Jenna Streicher, Henk Stunnenberg, Kai Tan, Deanne Taylor, Adrian Thorogood, Ludovic Vallier, Alexander van Oudenaarden, Fiona Watt, Wilko Weicher, Jonathan Weissman, Andrew Wells, Barbara Wold, Ramnik Xavier, Xiaowei Zhuang, and Human Cell Atlas Organizing Committee. The human cell atlas white paper. October 2018. URL http://arxiv.org/abs/1810.05192.

Aman Shrivastava, Ramprasaath R Selvaraju, Nikhil Naik, and Vicente Ordonez. CLIP-Lite: Information efficient visual representation learning with language supervision. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 8433–8447. PMLR, 2023. URL https://proceedings.mlr.press/v206/shrivastava23a.html.

Tabula Sapiens Consortium, Robert C Jones, Jim Karkanias, Mark A Krasnow, Angela Oliveira Pisco, Stephen R Quake, Julia Salzman, Nir Yosef, Bryan Bulthaup, Phillip Brown, William Harper, Marisa Hemenez, Ravikumar Ponnusamy, Ahmad Salehi, Bhavani A Sanagavarapu, Eileen Spallino, Ksenia A Aaron, Waldo Concepcion, James M Gardner, Burnett Kelly, Nikole Neidlinger, Zifa Wang, Sheela Crasta, Saroja Kolluru, Maurizio Morri, Serena Y Tan, Kyle J Travaglini, Chenling Xu, Marcela Alcántara-Hernández, Nicole Almanzar, Jane Antony, Benjamin Beyersdorf, Deviana Burhan, Kruti Calcuttawala, Matthew M Carter, Charles K F Chan, Charles A Chang, Stephen Chang, Alex Colville, Rebecca N Culver, Ivana Cvijović, Gaetano D'Amato, Camille Ezran, Francisco X Galdos, Astrid Gillich, William R Goodyer, Yan Hang, Alyssa Hayashi, Sahar Houshdaran, Xianxi Huang, Juan C Irwin, Sori Jang, Julia Vallve Juanico, Aaron M Kershner, Soochi Kim, Bernhard Kiss, William Kong, Maya E Kumar, Angera H Kuo, Baoxiang Li, Gabriel B Loeb, Wan-Jin Lu, Sruthi Mantri, Maxim Markovic, Patrick L McAlpine, Antoine de Morree, Karim Mrouj, Shravani Mukherjee, Tyler Muser, Patrick Neuhöfer, Thi D Nguyen, Kimberly Perez, Nazan Puluca, Zhen Qi, Poorvi Rao, Hayley Raquer-McKay, Nicholas Schaum, Bronwyn Scott, Bobak Seddighzadeh, Joe Segal, Sushmita Sen, Shaheen Sikandar, Sean P Spencer, Lea C Steffes, Varun R Subramaniam, Aditi Swarup, Michael Swift, Will Van Treuren, Emily Trimm, Stefan Veizades, Sivakamasundari Vijayakumar, Kim Chi Vo, Sevahn K Vorperian, Wanxin Wang, Hannah N W Weinstein, Juliane Winkler, Timothy T H Wu, Jamie Xie, Andrea R Yung, Yue Zhang, Angela M Detweiler, Honey Mekonen, Norma F Neff, Rene V Sit, Michelle Tan, Jia Yan, Gregory R Bean, Vivek Charu, Erna Forgó, Brock A Martin, Michael G Ozawa, Oscar Silva, Angus Toland, Venkata N P Vemuri, Shaked Afik, Kyle Awayan, Olga Borisovna Botvinnik, Ashley Byrne, Michelle Chen, Roozbeh Dehghannasiri, Adam Gayoso, Alejandro A Granados, Qiqing Li, Gita Mahmoudabadi, Aaron McGeever, Julia Eve Olivieri, Madeline Park, Neha Ravikumar, Geoff Stanley, Weilun Tan, Alexander J Tarashansky, Rohan Vanheusden, Peter Wang, Sheng Wang, Galen Xing, Les Dethlefsen, Camille Ezran, Astrid Gillich, Yan Hang, Po-Yi Ho, Juan C Irwin, Sori Jang, Rebecca Leylek, Shixuan Liu, Jonathan S Maltzman, Ross J Metzger, Ragini Phansalkar, Koki Sasagawa, Rahul Sinha, Hanbing Song, Aditi Swarup, Emily Trimm, Stefan Veizades, Bruce Wang, Philip A Beachy, Michael F Clarke, Linda C Giudice, Franklin W Huang, Kerwyn Casey Huang, Juliana Idoyaga, Seung K Kim, Christin S Kuo, Patricia Nguyen, Thomas A Rando, Kristy Red-Horse, Jeremy Reiter, David A Relman, Justin L Sonnenburg, Albert Wu, Sean M Wu, and Tony Wyss-Coray. The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376 (6594):eabl4896, May 2022. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.abl4896. URL http://dx.doi.org/10.1126/science.abl4896.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, and Patrick T Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, June 2023. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-023-06139-9. URL http://dx.doi.org/10.1038/s41586-023-06139-9.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and Fine-Tuned chat models. July 2023. URL http://arxiv.org/abs/2307.09288.

Unknown author. RNAseq profiling of defined immunocyte subsets from human blood, healthy volunteers. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE227743, 2023. Accessed: 2024-01-20.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, A Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. LiT: Zero-Shot transfer with locked-image text tuning. *Proceedings / CVPR, IEEE*

Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 18102–18112, November 2021. ISSN 1063-6919, 2575-7075. doi: 10.1109/CVPR52688.2022.01759. URL http://openaccess.thecvf.com/content/CVPR2022/html/Zhai_LiT_Zero-Shot_Transfer_With_Locked-Image_Text_Tuning_CVPR_2022_paper.html.

APPENDIX

A DATASET CURATION

Our training and validation datasets are generated by combining curated natural language annotations with raw gene-level RNA-seq read counts.

TRAINING SET GENERATION

The training set was generated based on the human ARCHS[4] dataset (v2.2, Date: 5-30-2023, 722,425 samples) (Lachmann et al., 2018), which comprehensively computes mouse and human GEO RNA-seq samples providing a standardized read count matrix. To generate the metadata annotations, we start from the ARCHS[4]-provided NCBI accessions (precedence of experiment accession >BioSample accession >GEO accession) and use the Entrez API, implemented in Biopython, to map them to their respective SRA UIDs. We then used these UIDs to retrieve raw metadata from the SRA database. We processed the raw metadata to contain ASCII characters only using the unidecode package and subsequently curated it using the MetaSRA pipeline (Bernstein et al., 2017), leading to 720,710 annotations. We leveraged the estimation of the *sample type confidence* by the MetaSRA pipeline, which indicated the pipeline's confidence for correct sample type classification, to filter the least "recognizable" samples from the dataset (`sample_type_confidence` <0.5). We also excluded samples that appeared in the validation set, resulting in the final set of 653,688 samples for our training set. To compress the structured metadata from GEO/MetaSRA into concise natural language summaries with a focus on the biologically relevant information, akin to CellWhisperer's intended use cases, we employed an open-weight LLM (Orca 2 13B (fine-tuned from Llama 2), (Mitra et al., 2023; Touvron et al., 2023). The following zero-shot prompt was used for each sample in combination with the YAML-formatted metadata preprocessed by MetaSRA.

> *Using the provided YAML annotation of an RNA sequencing study for a single sample, please describe the sample in natural language. Your response should be brief and focus solely on the biological aspects of the sample that contribute to understanding its cell state, such as cell type and any pertinent conditions or treatments.*
>
> *Specifically, exclude all technical and methodological details of the sequencing process, including but not limited to library type, read length, sequencing platform, or any other laboratory technique-related information. The aim is to succinctly convey what the sample is, biologically, without any reference to how it was sequenced or any other procedural data.*

An example of this conversion is shown in Table 1 for the sample with SRA ID SRX386384.

VALIDATION SET GENERATION

For the validation set, we discarded the read counts provided by ARCHS[4] and computed the read count matrix ourselves with the *fetchngs* pipeline (Ewels et al., 2020), in order to provide a realistic validation scenario, as transcriptome data are commonly analyzed with diverse methods. The validation dataset was generated by querying the public MetaSRA database for primary tissue samples associated with a disease state. Similar to the training set generation, we mapped the retrieved NCBI accessions to their SRA UIDs and then used these UIDs to retrieve raw metadata from the SRA database as well as associated GEO metadata and linked PubMed IDs. We then used this compendium of information to manually curate the retrieved data, only retaining those that are primary tissue samples The final dataset contains metadata of primary tissue generated either from healthy controls or patients with a given pathology.

Structured Annotation	**Compressed Sentence**
geo_title: RACA-18 rep2 geo_metadata: cell line: Jurkat T cell, ectopic expression: MALT1-R149A, genotype: endogenous MALT1 knock-out, p/i stimulation: 3 hours P/I geo_source_name: GC003254 sample_type: cell_line mapped_ontology_terms: Jurkat, cultured cell, T cell, hematopoietic cell, leukocyte, Jurkat, cell line, lymphocyte, CAT-MT raw_biosample_metadata: source_name: GC003254; cell_line: Jurkat T cell; ectopic expression: MALT1-R149A; genotype: endogenous MALT1 knock-out; p/i stimulation: 3 hours P/I treatment:	The sample is a Jurkat T cell line with ectopic expression of MALT1-R149A, an endogenous MALT1 knock-out, and was stimulated with phorbol 12-myristate 13-acetate (P/I) for 3 hours.

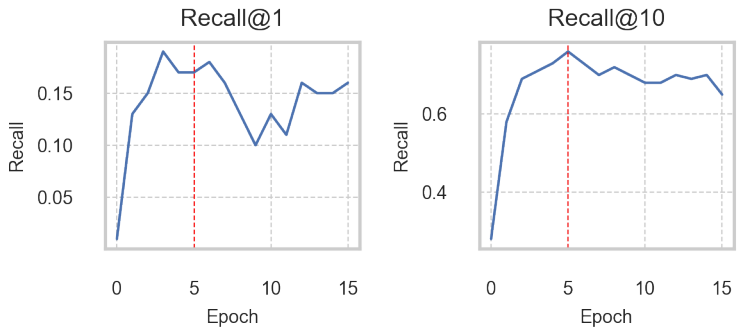
Table 1: Comparison of structured annotations, provided by SRA and preprocessed by MetaSRA on the left, and the LLM-derived natural language representation on the right

We processed the metadata from the validation samples with a different LLM and slightly modified prompt (GPT-4, USD 400 budget), to control for potential artifacts introduced into the training set by this artificial kind of data processing. The following zero-shot prompt was used in combination with the YAML-formatted metadata provided by the MetaSRA database for each sample.

> *You are given a structured (YAML) annotation of a RNA sequencing study with detailed information about a single sample from this study. Your job is to formulate a short and concise formulation of this single sample in natural language.*
>
> *Take special attention to the following points:*
>
> *- The YAML information provides context about the study in which the cellular context of interest was observed. This context may provide predominantly irrelevant information with respect to the cell state, so pay special attention to the sample-specific information in the YAML.*
>
> *- Provide only information that is relevant to the cell state (e.g. cell type, perturbation, ...) in any manner. In other words, focus on biological properties, which are reflected in the cellular phenotype and transcriptome of the sample.*
>
> *- Disregard information that is not reflected in the phenotype or transcriptome of the sample. E.g., discard all study-specific information.*

B Cross-modality retrieval for model selection

To avoid overfitting, we assessed model performance after each training epoch via our validation set and picked the best model for further downstream evaluations. Specifically, following (Radford et al., 2021), we assessed the model's ability to *retrieve* the correct text label for a given transcriptome embedding from paired data points in a deduplicated version of our validation set. For deduplication, we processed our generated natural language annotations (Appendix A) with BioBERT (v1.1 (Lee et al., 2020)) to derive the CLS tokens embeddings, which represent the full input in a single vector. We then used hierarchical clustering ($metric = cosine, linkage = "average"$) to extract 100 clusters from those embeddings. For each cluster, we retained the sample closest to the cluster center as a representative example. This was necessary because the full dataset contained, for many samples, dozens of identical or near-identical annotations that even a perfect model could not reliably distinguish in latent space. We found the $R@10$ (recall at 10) metric, which describes the number of samples of one modality that ranked its partner amongst the top 10 closest hits, to be among the smoothest metrics across epochs in previous runs and chose it to select the best model (see Appendix Fig. 1). The recall at various passing thresholds for our selected checkpoint at epoch 5 is shown in Table 2. Overall, the trained model was able to confidently retrieve pairs from a deduplicated version of the validation set (macro-averaged ROC-AUC: 0.9244, see Appendix A for training details) pointing to the robustness of the model towards diverse subsets of data and preprocessing methods.



Appendix Figure 1: Transcriptome-based text retrieval on deduplicated validation set. Best model was chosen based on Recall@10 metric (epoch=5), to prevent overfitting to the training data.

R@1	R@5	R@10	R@50
0.17	0.53	0.76	1.0

Table 2: Recall of the correct text label for a given transcriptome embedding from paired data points in a deduplicated version of our validation set. A pair is correctly recalled if it is among the top n hits.

C ACCESSING TRANSCRIPTOMES WITH NATURAL LANGUAGE USING CELLWHISPERER

THE HUMAN TRANSCRIPTOME LANDSCAPE MAPPED TO NATURAL LANGUAGE

Leveraging the multimodality of our model, we created an annotated atlas of the measured human transcriptomics landscape. We projected CellWhisperer embeddings of human GEO transcriptomes two dimensions using UMAP (Fig. 2) and annotated the landscape after clustering ($n_{Leiden} = 143$).

To generate the annotations with CellWhisperer, we mined a large number (17,220) of biologically relevant terms, based on the gene set names (not the gene names) from Enrichr (Chen et al., 2013) libraries (see below), which we embedded using CellWhisperer's text tower. Transcriptomes were embedded using CellWhisperer's transcriptome tower and compared to the keywords by computing their cosine similarity. Up to five most similar ones (highest cosine similarity, filtering for positive values) per library are used downstream.

GEO cluster annotations (Fig. 2) were then generated using CellWhisperer keywords based on the cluster-level mean transcriptome embedding following an aggregation via GPT-4 with the following prompt:

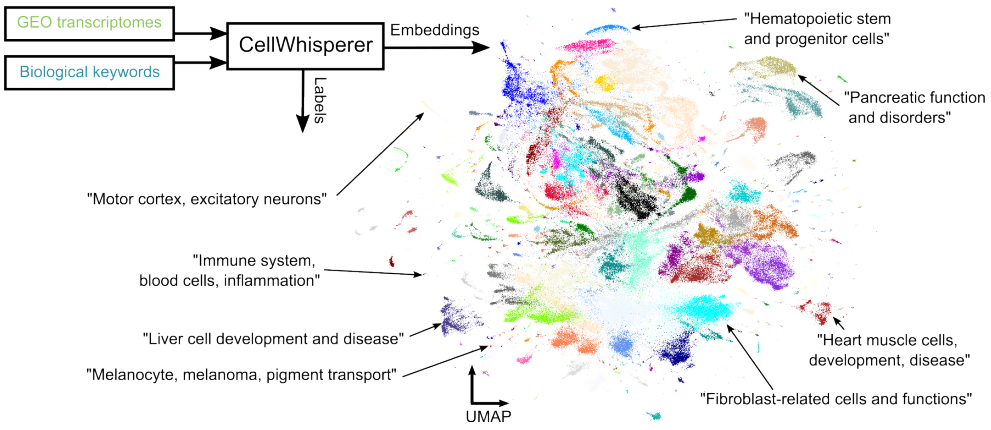
> *I will provide you with a number of entries of terms that describe a cluster of cells. There is a confidence score next to each term - pay more attention to large scores. Please give a very short description of the cells in the cluster based on the term. Note that not all terms will be necessarily relevant for this task - just try to find the common theme of the terms and report that. If possible, describe biological concepts beyond just cell type names. Reply with less than six words!*

Terms from the following gene set libraries were used:

- `Achilles_fitness_decrease`, `Achilles_fitness_increase`
- `Azimuth_2023`
- `Disease_Perturbations_from_GEO_down`, `Disease_Perturbations_from_GEO_up`
- `GO_Biological_Process_2023`, `GO_Cellular_Component_2023`, `GO_Molecular_Function_2023`
- `Gene_Perturbations_from_GEO_down`, `Gene_Perturbations_from_GEO_up`

- `MSigDB_Hallmark_2020`, `MSigDB_Oncogenic_Signatures`
- `PanglaoDB_Augmented_2021`,
- `Tabula_Sapiens`

Additionally, we used annotations from `Tabula Sapiens` (unique values in the columns `organ_tissue`, `anatomical_information`, `gender`, `cell_ontology_class`, `free_annotation` and `compartment`).



Appendix Figure 2: Embeddings and cluster labels created with CellWhisperer capture the diversity of human transcriptomics in the GEO database

Thereby, the cluster labels were generated purely based on the transcriptomic state and without direct reliance on the GEO annotations. The map shows clusters reflecting development, tissues and disease states, among others and can be accessed in our provided web demo.

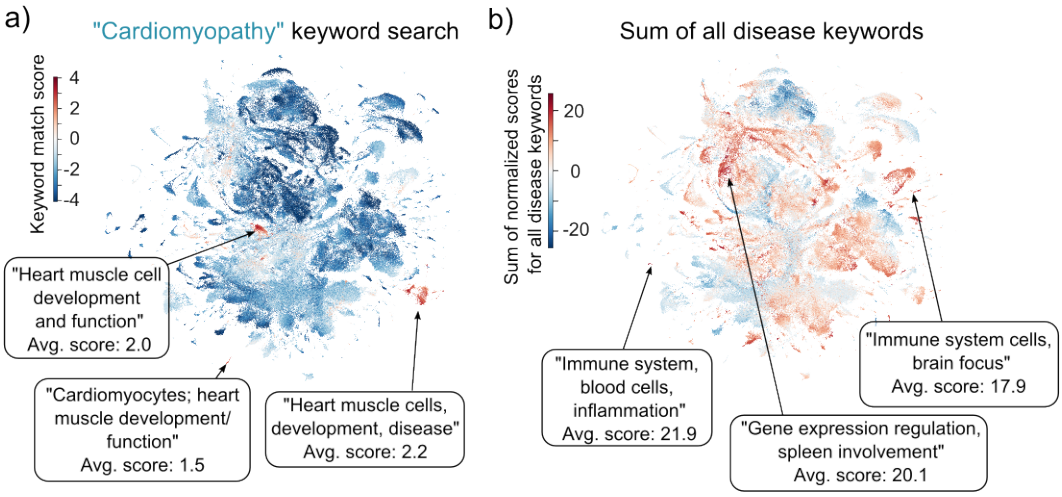
Linking GEO clusters to Diseases

A major interest of human biology is to study disease, so we embedded common disease terms, such as "cardiomyopathy", and compared them to all individual samples, observing how CellWhisperer's label for the top-scoring clusters satisfyingly corresponded (e.g. top hit: "Heart muscle cells, development, disease"; Appendix Fig. 3a). We further validated this observation by confirming that the top-scoring transcriptomes were indeed derived from samples associated with the respective disease. With this insurance, we were curious to see whether some of the biological clusters were more related to studied diseases than others, so for each sample, we calculated the similarities to each of the 187 disease names in the "Expanded OMIM" set and summed them after disease-wise standardization. (see Appendix Fig. 3b). The map showed cluster-coherent disease matching and high disease-scores for clusters with disease-related labels such as "Immune system, blood cells, inflammation". In conclusion, CellWhisperer's allows to flexibly scan all of human GEO by arbitrary biological concepts. The map and labels can be explored in detail on our provided web demo (see category `cluster_label`, details in Appendix F).

D Zero-shot evaluation on published datasets

We performed zero-shot validation in a reference-free manner, i.e. only by comparing CellWhisperer's embeddings of transcriptomes and the dataset-specific cell type labels. Specifically, we embedded the labels as sentences mimicking a user request *A sample of* $\{celltype\}$ *from a healthy individual*. We then embedded all transcriptomes for a given dataset with CellWhisperer and scored transcriptome-text pairs by their cosine similarity, using the softmax over all texts as model predictions to calculate ROC-AUC (macro average) values.

The evaluation datasets used throughout our study, primarily for zero-shot predictions, are described below.



Appendix Figure 3: a) Free-text search of CellWhisperer's GEO embeddings for the keyword "cardiomyopathy". Coloring corresponds to CellWhisperer's computed similarity to the search term. The three clusters with the highest average score are labeled with the CellWhisperer-generated labels. b) Visualization of the sum of similarities to all queried disease-related keywords. The top three clusters with the highest average scores are highlighted.

Evaluation Datasets

The Tabula Sapiens (Tabula Sapiens Consortium et al., 2022) raw read counts were downloaded from [2]. We used the annotation column `cell_ontology_class` as reference cell types, and we capped the number of cells per class at 100 for our quantitative analysis to decrease class imbalance (Appendix Fig. 2b, first bar). For our subset of well-studied cell types, we selected all cells from the top 20 most common cell types from the set of cells found in liver, lung and blood.

For the Immgen dataset, which contains bulk sequencing data of human immune cells (Heng et al., 2008; Unknown author, 2023), raw read counts were downloaded from [3], and the cell type labels were manually curated and simplified.

Pre-processed read counts for the pancreas dataset (Luecken et al., 2022) were downloaded from [4].

E Fine-tuning on single-cell data

To assess whether CellWhisperer's ability to predict cell type labels could be further improved, we fine-tuned it on the Tabula Sapiens dataset, generating natural language annotations via a simple string expansion based on the labels provided by the dataset: *"*$\{free\_annotation\}$ *in the* $\{compartment\}$ *compartment of the* $\{organ\_tissue\}$*"*. We fine-tuned with a reduced maximum learning rate of 1e-5 for 10 epochs to avoid overfitting. Performance on our (GEO-based) validation set dropped notably but remained at a high level (R@10: from 0.76 to 0.53), while predictive performance on Tabula Sapiens increased dramatically (e.g. cell type R@1: from 0.23 to 0.61; R@10: from 0.84 to 0.96). More importantly, performance on the independent Pancreas and Immgen datasets improved notably (macro-averaged ROC-AUC Pancreas 0.83 to 0.86; Immgen 0.95 to 0.97).

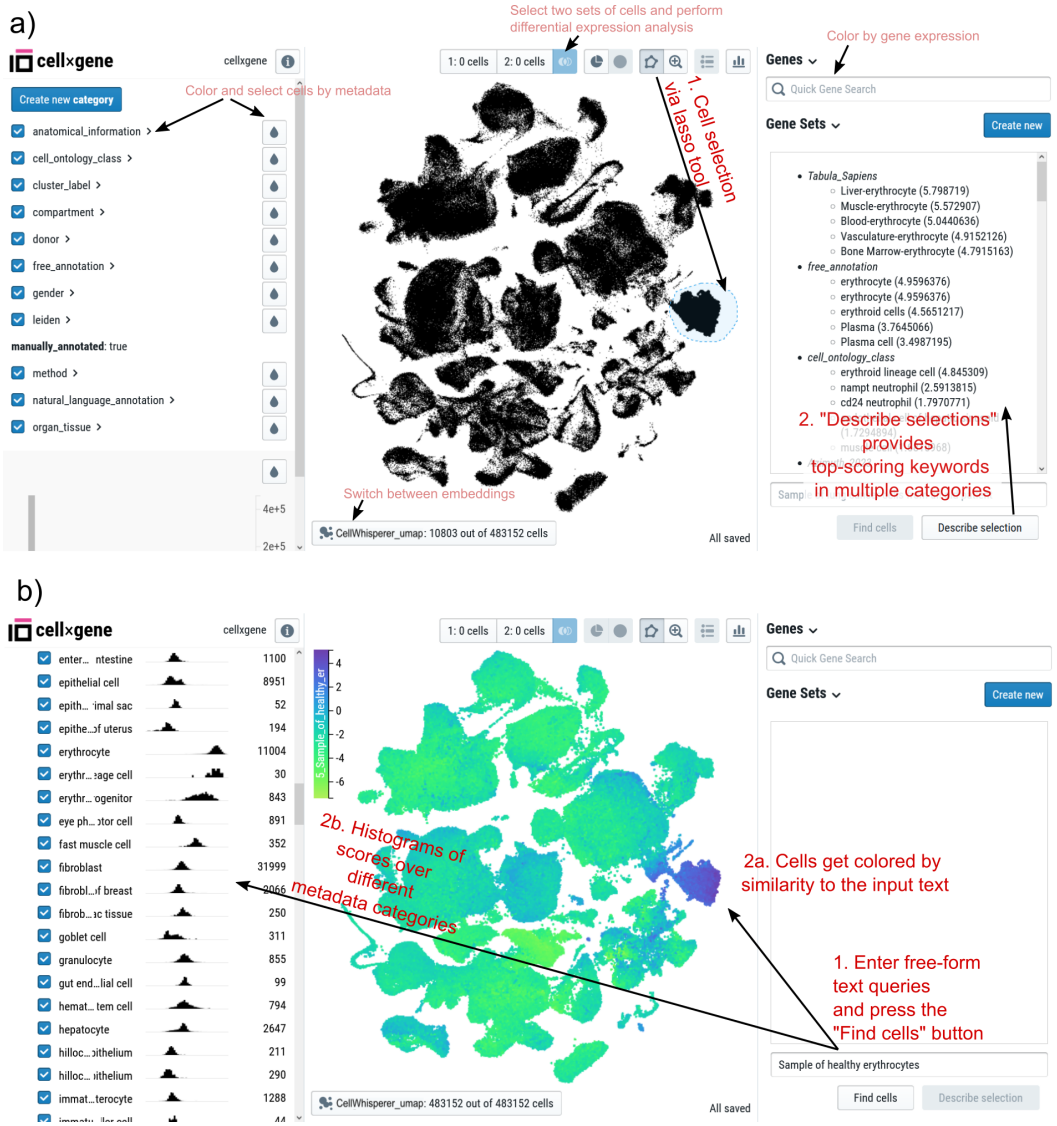
F Interactive genome browser implementation

Building atop the *CELLxGENE Explorer* single cell browser (v1.2.0) (CZI Single-Cell Biology Program et al., 2023), we extend the UI interface by a minimally viable chat interface. We implement

[2] `https://figshare.com/ndownloader/files/40067134`

[3] `https://sharehost.hms.harvard.edu/immgen/GSE227743/GSE227743_Gene_count_table.csv`

[4] `https://figshare.com/ndownloader/files/43480497`



Appendix Figure 4: Interactive usage of CellWhisperer within the CELLxGENE browser. a) Functionality to query matching keywords for a selected subset of cells via CellWhisperer (bright red, larger font), and CELLxGENE core functionality that is additionally available (faded red, smaller font). b) Continuous color annotations per cell for a user-provided free-text query via CellWhisperer.

two API endpoints: (i) to query matching keywords for a selected subset of cells (Appendix Fig. 4a) and (ii) to obtain continuous (color) annotations per cell for a user-provided free-text query (Appendix Fig. 4b). Appendix Fig. 4 shows generation of panel Fig. 2a.

A general introduction to the CELLxGENE Explorer is provided on their project website [5].

A live demo of CellWhisperer can be accessed with username: review and password: iclr2024-mlgenx at [6]

[5] https://cellxgene.cziscience.com/docs/04__AnalyzePublicData/4_1_HostedTutorials

[6] http://cellwhisperer.bocklab.org/

To improve responsiveness to user requests, we precompute embeddings for all single-cells for a given scRNA-seq dataset offline and provide them to the web service via a zipped numpy file (runtime of approximately 1 hour for Tabula Sapiens).