SCVI-HUB: A FLEXIBLE FRAMEWORK FOR REFERENCE ENABLED DATA ANALYSIS

Can Ergen[1,2,*,#], Valeh Valiollah Pour Amiri[1,*], Martin Kim[1], Aaron Streets[1,3,4], Adam Gayoso[1,#], Nir Yosef[1,2,5,#]
[1]Center for Computational Biology, University of California, Berkeley, Berkeley, CA, USA
[2]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA, USA
[3]Department of Bioengineering, University of California, Berkeley, Berkeley, CA, USA
[4]Chan Zuckerberg Biohub, San Francisco, CA, USA
[5]Department of Systems Immunology, Weizmann Institute of Science, Rehovot, Israel
[*]Equal contribution
[#]Correspondence to: cergen@berkeley.edu, adamgayoso@berkeley.edu, nir.yosef@weizmann.ac.il

ABSTRACT

The accumulation of single-cell omics datasets in the public domain has opened new opportunities to reuse and leverage the vast amount of information they contain. Such uses, however, are complicated by the need for complex and resource-consuming procedures for data transfer, normalization, and integration that must be addressed prior to any analysis. Here we present scvi-hub: a platform for evaluating, sharing, and accessing probabilistic models that were trained on single-cell omics datasets. We demonstrate that these pre-trained models allow immediate access to a slew of fundamental tasks like visualization, imputation, annotation, outlier detection, and deconvolution of new (query) datasets with a much lower requirement for compute resources. We also show that pretrained models can help drive new discoveries with the existing (reference) datasets through rapid, model-based analyses. Scvi-hub is built within scvi-tools and integrated into scverse. Scvi-hub is publicly available to enable efficient sharing of single-cell omic studies, and also to put advanced capabilities for transfer learning at the fingertips of a broad community of users. We provide an extended journal version on bioRxiv.

1 INTRODUCTION

Machine learning models have been central to efforts to catalog cell states in health and disease with single-cell omics technologies Kharchenko (2021); Wagner et al. (2016); Heumos et al. (2023). These models are capable of performing a variety of analysis tasks including dimensionality reduction, differential expression comparison, automated cell type annotation, denoising, deconvolution of spatial data, and modality imputation Heumos et al. (2023). With the growth of single-cell data corpora, transfer learning will be an essential technique for accomplishing such tasks by leveraging large-scale datasets as reference atlases in a computationally efficient and performant manner. At present, transfer learning is primarily used for projecting cells onto a common low-dimensional representation that is used for tasks such as annotation or trajectory inference Lotfollahi et al. (2021); Kang et al. (2021); Hao et al. (2022). However, moving forward, additional applications will become more prevalent, such as the interpretation of spatial data Lopez et al. (2022) the prediction of the outcome of a genetic perturbation Roohani et al. (2023), the prediction of multi-modal information from single-modality data Ashuach et al. (2022), the detection of abnormal cellular subsets Dann et al. (2022), or a more robust analysis of differential expression Boyeau et al. (2023).

Methods for transfer learning in single-cell omics broadly fall into two categories: nonparametric and parametric. In the non-parametric case, the algorithm uses the reference data directly to remove unwanted sources of variation. For example, Seurat and FastMNN integration utilize mutual nearest neighbors between reference and query data to remove query-specific effects Hao et al. (2022); Haghverdi et al. (2018). In the second case, the algorithm uses a parametrized model like a con-

ditional variational autoencoder (cVAE) to reduce the dimension of the reference data and remove unwanted variation. The same model can then be leveraged to project new data onto the same low-dimensional space. This approach is used by methods like scVI+scArches Lotfollahi et al. (2021); Lopez et al. (2018); Gayoso et al. (2022) and can be efficiently designed so that at query time only the query dataset is used, and not the original, potentially large, reference dataset. In many cases, these models also offer the benefit of a neural network decoder that can regenerate normalized raw count data with high fidelity from the low-dimensional representation.

Although parametric approaches have been used successfully in large-scale analyses Suo et al. (2022); Jones et al. (2022), there are still challenges to realize the power of reusing trained models. First, models can be trained using a variety of machine learning libraries and frameworks or using different versions of the same library, making them difficult to quickly ingest and use. Second, there is no standard for how pre-trained models should be deposited, which limits reuse to knowledge of a particular publication or model deposition. Finally, it can be difficult to assess the quality of a pre-trained model without standardized summary statistics of key performance metrics.

To address these issues, we introduce scvi-hub, a platform for sharing and reusing single cell machine learning models that are implemented in the scvi-tools codebase Gayoso et al. (2022). As a new component of scvi-tools, scvi-hub provides access to various popular single-cell model architectures spanning the core data analysis tasks. Scvi-hub facilitates model sharing through the Hugging Face Model Hub. For model consumers, it offers streamlined access to a variety of downstream analysis tasks using the downloaded models while using a minified version of the data to lower the data storage and download bandwidth requirements, thus increasing accessibility and inclusion within the data analysis community 3. For model developer, it allows sharing trained models in a streamlined way and offers posterior predictive checks to evaluate quality of fit of a trained model. To demonstrate its utility, we have seeded scvi-hub with a collection of more than 90 models pretrained on a variety of tissues and experimental conditions from the Tabula Sapiens consortium Jones et al. (2022) and other projects (https://huggingface.co/scvi-tools).

2 RESULTS

2.1 SCVI-HUB FACILITATES REUSE OF MACHINE LEARNING MODELS PRE-TRAINED ON SINGLE-CELL DATASETS.

For the community of contributors, scvi-hub provides features that facilitate both evaluation and sharing of models. Scvi-hub uses posterior predictive checks Gelman et al. (1996) with model-simulated data to evaluate models. Here, we implemented previously described metrics for single-cell omics data, including the coefficient of variation Levitin et al. (2019); Gayoso et al. (2021b), as well as new metrics based on differential expression (see Methods). Importantly, these metrics are dataset-agnostic, in that they do not require dataset-specific information (such as cell-type labels or sample metadata) and are therefore broadly and immediately applicable.

The provided Hugging Face Model Hub has features that make it ideal for single-cell genomics. Models can be discovered, enabled via an advanced search interface and a uniform documentation and presentation with Model Cards (description files that accompany and provide information on the uploaded model). Additionally, the Hugging Face Model Hub provides backward compatibility through built-in git-based version control. Model contributors have the option to upload and share the data behind their model, allowing for a wide variety of uses (examples in Figures 1, 2). Data can be uploaded to Hugging Face in its raw form (count matrix) or in a substantially reduced form, using a new feature, which we refer to as ”data minification”, in which we only store the posterior parameters of the data, which can then be converted into an approximated and normalized form of the original data, using the generative part of the model (see Methods). Since the models occupy orders of magnitude less space than the raw data, this feature makes available a compressed representation of the reference data set, while the generative part still allows efficient downstream analysis, such as differential expression, feature correlation, and missing data imputation Lopez et al. (2018); Gayoso et al. (2021a); Ashuach et al. (2022); Boyeau et al. (2023); Steier et al. (2023). Using minified data also provides more than 50% speed improvement over standard datasets by improving speed of the dataloaders.

The second set of users to which scvi-hub caters is the model consumer who wishes to analyze existing (reference) datasets or leverage reference datasets to analyze their own (query) data. Since scvi-hub is part of the scvi-tools package and the scverse ecosystem Virshup et al. (2021), which includes Scanpy Wolf et al. (2018), consumers can seamlessly integrate the downloaded models into existing analysis workflows. There are extensive tutorials on working with scvi-hub in the R ecosystem as well, thus supporting downstream analysis with Seurat Butler et al. (2018) and other popular environments.

2.2 SCVI-HUB ENABLES EFFICIENT EXPLORATION AND RE-ANALYSIS OF LARGE REFERENCE DATASETS.

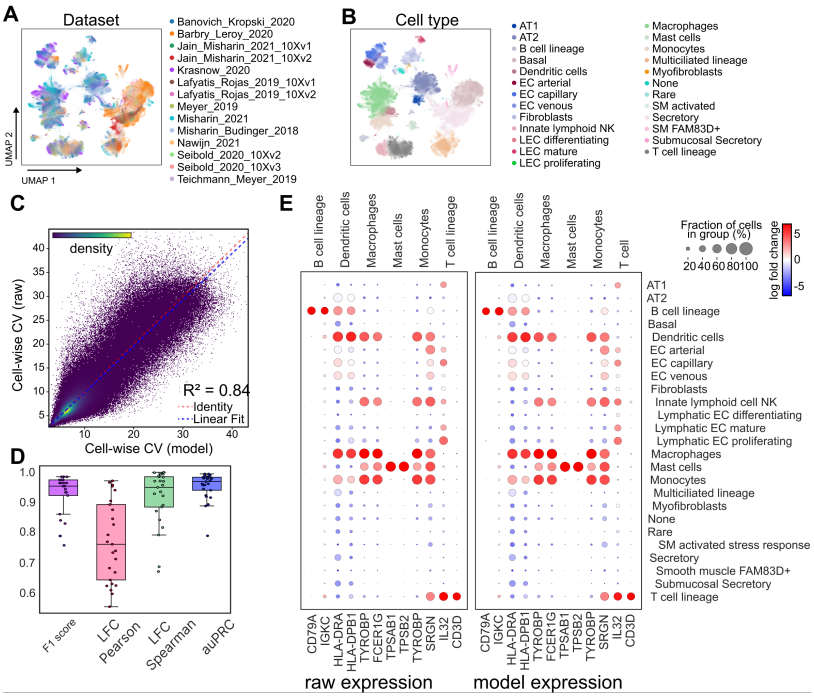


Figure 1: Figure 1: Reference-only tasks enabled by scvi-hub. Using the Human Lung Cell Atlas (HLCA) Sikkema et al. (2023) as an example of a reference data set in scvi-hub. (A-B). Cells colored by the original studies and cell type. (C). Coefficient of variation (CV) of the HLCA scANVI model. Each dot is a cell. y-axis shows CV computed on raw data, while x-axis shows the CV computed for generated data. (D). Posterior predictive checks using Scanpy differential expression (DE) between the cell-types in panel B (one vs. all DE); each dot represents metrics for one cell-type. Reported are: F1 accuracy, evaluated by overlap of the top 100 genes, Pearson and Spearman correlation coefficients between the log-2 fold-changes (LFC) evaluated with raw vs. generated counts, and area under the precision-recall curve (auPRC) with genes identified by analysis of the raw data (adjusted p-value below 0.2) as the set of true hits and gene ranking defined by p-values evaluated with the generated counts. (E). Comparison of top marker genes for all immune cell-types (taking the top two for each type). Entries are colored by LFC of the respective one-vs-all comparison and sized by the number of cells with non-zero values for the respective gene in the raw data (left) or model-estimated proportion of expressing cells (right; see methods).

The first set of analyzes facilitated by scvi-hub is centered on the reference datasets, without inclusion of any additional (query) information. This approach allows for rapid exploration and reanalysis of reference atlases (without requiring time-consuming and resource-heavy model training procedures). After searching for a model that fits the user's specification in Hugging Face (i.e., tissue, cell types, and a modeling scheme, such as scVI, DestVI etc.), model consumers can pull (download) the model and its corresponding data, which can be in raw or minified format (see Methods). Analysis of reference datasets includes operations that concern the low-dimensional (latent) representation of cells, such as visualization, clustering, trajectory inference, and differential abundance analysis.

Additionally, model consumers can perform analyses at the high-dimensional omics measurement level by accessing the raw data or generating count data using the minified format. These high-dimensional representations can then be readily analyzed in Scanpy Wolf et al. (2018) or Seurat Butler et al. (2018), through a slew of procedures.

Figure 1 highlights the use of *scvi.criticism* coefficient of variation and differential expression metric and highlights that minified data gives well correlated estimates of gene expression (further discussion in the Appendix).

2.3 SCVI-HUB ENABLES EFFICIENT REFERENCE-BASED ANALYSIS OF QUERY DATA BY TRANSFER LEARNING.

The most fundamental part in analyzing a query data set given the appropriate reference is to represent the query data using the reference model. In the context of scvi-tools, this is done by calculating the coordinates of each query cell in the latent space of cell states that is encoded by the model and that is initially populated by the reference cells (Fig. 1B). While the encoder networks in scvi-tools can readily provide latent representation for any query cells, these representations might be distorted by batch effects. As retraining the model anew on the reference and query data together is a resource-consuming procedure, an efficient way to address this is to perform minimal training only to capture these batch effects and leaving the model otherwise unchanged. This procedure, which is implemented by scArches Lotfollahi et al. (2021), effectively produces a joint embedding of the reference and query data set while quickly removing unwanted variation. Here, we demonstrate the capabilities of scvi-hub to analyze query data with transfer learning through four such fundamental tasks: visualization, annotation, anomaly detection, comparative analysis.

As a first test case, we used a query dataset of three healthy individuals and three emphysema patients Wang et al. (2023) and designated the HLCA data set as a reference. The two datasets (query and reference) are well integrated with scANVI as reference model and with use of scArches to add the query data, leading to an informative and reference-informed visualization of the query data (Figs. 2 A-B). To annotate the cells states in our query, we transferred labels from HLCA using a simple KNN classifier (using the reference embedding in the joint latent space a neighborhood index; Fig. 2b). We found that the reference-based annotation is consistent with the labels provided by the original study of the query dataset, yet adds a great deal of resolution. For example, cells that were originally labeled Endothelial Cells (EC) were now divided into several subgroups of that lineage, including venous systemic, venous capillary, arterial, and aerocyte capillary.

The reference-based probabilistic representation computed for each query cell can also facilitate comparative analysis within the query data set. To explore this, we first used Milo (which relies on the integrated embedding space) to compare the composition of cell states in the three healthy query samples versus the three emphysema-affected samples. We found a significant increase (FDR $< 10\%$) in the abundance of certain states of macrophages, fibroblasts, and epithelial cells in the emphysema samples (Fig. 2D). To gain more insight into the change in cell states, we next used the differential expression function built into scvi-tools (which uses the reference-based scVI model) to explore disease-associated gene expression changes in fibroblasts of the query data (Fig. 2E). We find that fibroblasts in patients with emphysema strongly upregulate pro-inflammatory chemokines that attract neutrophils (*CXCL1*/*CXCL2*/*CXCL8*), monocytes (*CCL2*, *CSF3*) and T cells (*CCL19*, *CCL20*). We were unable to generate equally insightful differentially expressed genes by pseudobulk differential expression analysis or training models from scratch, but improved performance is restricted to using the reference-based model (Supplementary Fig. 3, 4). We confirmed these findings using raw gene expression (Fig. 2F)

The original publication highlighted the role of fibroblasts in inducing a niche for resident memory T cells, further corroborated by the fact that depletion of fibroblasts in a mouse model led to a decrease in resident Th17 cells. Therefore, our reference-powered analysis highlights additional putative mechanisms associated with neutrophils and monocytes. In fact, neutrophils release granule proteins such as neutrophil elastase and myeloperoxidase and have thus been associated with emphysema Gernez et al. (2010). Similarly, monocyte-derived macrophages produce metalloproteinases that lead to tissue remodeling in emphysema Shapiro (1999).

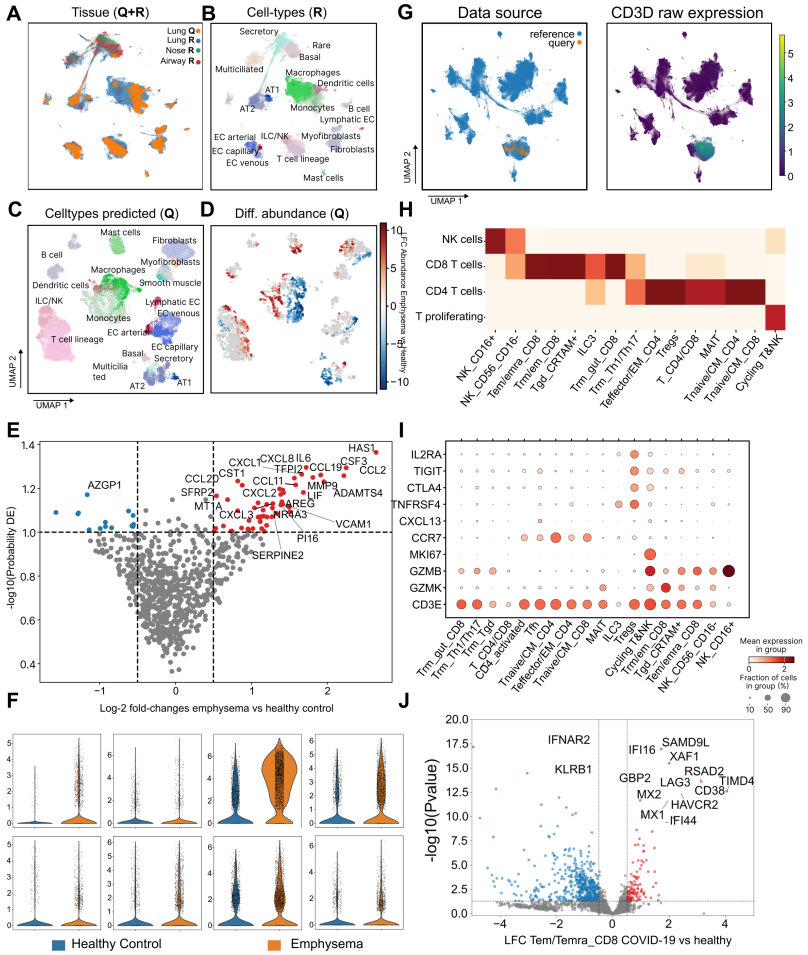


Figure 2: Figure 2: Query-to-Reference mapping tasks enabled by scvi-hub. (A). Joint embedding and UMAP visualization of the emphysema dataset (as query) and HLCA (as reference); cells are colored by their respective tissue and dataset. Q denotes the query dataset, while R denotes the reference data set. (B). Visualization of the reference cells in the same UMAP coordinates, colored by coarse-level cell type (ann_level_3 in HLCA). For display purposes, the different annotations for smooth muscle cells and lymphatic endothelial cells were summarized into one label. (C). UMAP computed on query cells only (using their coordinates in the joint query and reference latent space). Cells are colored by their summarized transferred cell-type. (D). Differential abundance computed using Milo between cells from healthy donors vs. patients with emphysema. Colored by log-2 fold change for neighborhoods with FDR $< 10\%$. (E). Model-based differential gene expression, comparing fibroblasts from healthy vs. diseased samples. The mean log-2 fold change is displayed on the x-axis and probability for non-DE on the y-axis. All genes displayed are significant with an $FDR < 10\%$. (F). Violin plots of key differentially expressed genes based (library-size normalized and log1p transformed raw data). (G). Embedding of T cells from the cross-tissue immune cell dataset (as query) integrated with HLCA reference dataset. CD3D expression is library-size normalized and log1p transformed. (H). Confusion matrix between the finest annotation scheme in HLCA (rows) and the labels infused using the query annotations (columns). (I). Canonical marker genes for cell-types are displayed on normalized raw data. (J). Pseudo-bulk differential expression analysis between Tem/Temra_CD8 cells from COVID-19 infected samples vs. healthy controls using PyDESeq2.

2.4 SCVI-HUB ENABLES EFFICIENT RE-ANALYSIS OF REFERENCE DATA SETS BY INFUSING NOVEL INSIGHTS FROM QUERY DATA SETS.

The joint embedding of query and reference datasets can also be used for an additional, less prevalent procedure, which we term label infusion. Although reference atlases tend to include large numbers of samples and cells, their levels of annotations can be limited in granularity. By re-annotating reference datasets based on labels in more finely annotated datasets, fine cellular subsets of interest can be identified and their function in disease contexts can be studied leveraging the power of the reference dataset.

We demonstrate label infusion by refining the cell-type labeling of NK and T cells in the HLCA dataset (Fig. 2G). We used a recent study of immune cells in different organs Conde et al. (2022), subsetted this study to only NK and T cells across all different organs, and integrated the resulting "query" data set with the reference HLCA, using its scANVI model. The joint embedding then helped us transfer knowledge from the query to the reference (label infusion; Fig. 2H). We further validated cell-type labels infused by checking their canonical cell-type marker genes (Fig. 2I) and found agreement.

The infused labels can be used to gain new insight from the reference data. To demonstrate this, we focus on a subset of cells that we reannotated as CD8+ resident memory T cells (labeled as Tem/Temra_CD8 in Fig. 2). Since the HLCA includes samples from COVID patients in addition to healthy donors, we were now able to examine the specific effects of infection on this more narrowly defined immune subset. This analysis finds an up-regulation of markers of exhaustion of CD8+ T cells (*LAG3*, *CD38*, *TIMD4*, and *HAVCR2*) as well as an increase in interferon-regulated genes (*GBP2*, *MX1*, *MX2*, *IFI16*, *XAF1*, *SAMD9L*, and *IFI44*), both of which are consistent with recent, more targeted studies of COVID infection Szabo et al. (2021); Rha & Shin (2021).

3 DISCUSSION

Given the growing corpus of single-cell omics datasets (both individual studies and atlas-level efforts), transfer learning techniques are becoming pivotal in enabling studies of both new and revisited datasets. Despite the promise of these techniques, their use has been limited owing to two major reasons. First, the use of large reference datasets can be prohibitive both in terms of the required compute resources (e.g. to access and process the data) and expertise (e.g. for integrating it with query data). Second, there is a need for an appropriate platform to facilitate communication between data providers and consumers and to provide the infrastructure for quality control, access, and downstream analysis. Scvi-hub provides a way to alleviate these problems by establishing a platform for sharing and reusing single-cell omics data. Sharing of models helps reduce the need for technical expertise and expensive compute resources (e.g., to process atlas-scale references and integrate them with query data), and the respective API was designed to facilitate a wide variety of analyzes.

The decentralized nature of scvi-hub enables community access through friendly and easy-to-use interfaces. As such, we envision scvi-hub to serve several types of users. First, we expect it to become a platform for individual researchers who wish to make their analyses accessible and reproducible. Second, we expect it to be used in efforts to generate tissue atlases and large-scale single-cell census datasets as a way of facilitating advanced use of these atlases. Third, we expect the research community to leverage the models in scvi-hub as an actionable resource for an array of use cases, from annotation of new scRNA-seq samples to deconvolution of ST samples. To demonstrate this, we presented several case studies showing that the incorporation of external references can improve and enrich the analysis of individual datasets and provide novel insights into disease mechanisms.

The model-centric approach of scvi-hub enables representation of large reference datasets in a minified format, which enables access with limited memory resources or download bandwidth, and thereby accelerates access to those valuable resources. It is our hope that this will help democratize single-cell data analysis and expand it to communities with low compute resources. Validating the findings in the original expression space is key. Scvi-hub can serve as a central gateway to data repositories, such as CELLxGENE Discover CZI Single-Cell Biology Program et al. (2023), which allows users to access selected portions of very large datasets (e.g., genes or subpopulations highlighted by the model-based analysis) for close inspection.

An important part of the interface between data providers and consumers is the ability to criticize and make informed decisions about the merits of a given reference data set or model for an application of interest. To that end, we developed a new suite scvi-criticism, which can serve to validate models prior to upload as well as for evaluating how well a query dataset fits a reference model. We believe both are essential for effective transfer learning.

The development of scvi-hub aims to foster a model-driven paradigm in the single-cell data analysis community, one where models are easy to find, access, develop and share, and can be efficiently leveraged to analyze various aspects of new and existing datasets. We expect it to become a growing resource, catering for new types of analyzes, use cases, and data modalities.

A Appendix

A.1 Scvi-hub provides evaluation procedures to select models and test their suitability for query datasets.

Evaluation of model quality is an essential part of scvi-hub. It allows contributors to scrutinize their models before upload, and consumers to verify that the models that they download are relevant and of sufficient quality. To achieve this, we developed *scvi.criticism* - a new module for evaluating models that were trained with scvi-tools. *scvi.criticism* implements posterior predictive checks (PPCs), which compare the distribution generated by the fitted model and the actual observed data Lopez et al. (2018); Gelman et al. (1996); Gayoso et al. (2021a). To perform PPC, we first sample from the distribution of the data as predicted by the fitted model and then compute summary statistics per gene (coefficient of variation and differential expression between prespecified groups of cells) in this distribution and in the raw data. The PPC consists of measuring the closeness between the resulting pairs of statistics (raw vs. generated). Close similarity is a standard measure that defines a well-trained model that is representative of the original data. To demonstrate this, we computed PPC on the Human Lung Cell Atlas dataset (HLCA), Sikkema et al. (2023) using the scANVI model provided by the author Xu et al. (2021) that was trained on this data (Fig. 2A-B). We reported the coefficient of variation, as well as a set of metrics computed based on the results of differential expression between the author-provided cell-type labels, performed on the predicted data and on the raw data (Fig. 2D-F). We observed that the data generated by the model fit well to the raw data, e.g., identifying similar sets of differentially expressed genes in a one- vs.-all-cell-type comparison.

Model contributors can use scvi.criticism to evaluate the goodness-of-fit of their models, select among several candidate models (e.g., with different hyperparameters) and optionally include these evaluations as part of the respective Model Card on Hugging Face. This is particularly useful in the case where the data are minified, as it provides more confidence in the reliability of the counts generated by the model in the absence of the full raw counts. To demonstrate this, we used *scvi.criticism* to evaluate the goodness of fit of one well-trained and two poorly trained scVI models on the Heart Cell Atlas dataset Litviňuková et al. (2020) (Supplementary Fig. 1 and Methods). We compared the cellwise coefficient of variation results computed on the raw data and on the estimated data from each of the three models. We observe that the well-trained model performs better (higher Pearson correlation with the raw coefficients of variation) than both poorly trained models. We also report similarities between the differential expression results calculated on the raw and estimated data from each of the three models (Supplementary Fig. 1B). We observed, once again, that the well-trained model performed consistently better than either of the two poorly trained models in most metrics.

Another important use case of *scvi.criticism* is to evaluate the extent to which a reference model is apt to analyze a query dataset. To demonstrate this, we created a query dataset consisting of all epithelial cells from the Tabula Sapiens Jones et al. (2022) project (spanning different tissues) and used the HLCA pre-trained scANVI model as our reference (i.e., many cell types, but airway only). We projected the query data set onto the reference model, using the scArches functionality. We then fed the obtained latent representation of the query data to the generative part of the model after transfer learning to test whether it is capable of generating gene expression profiles that are similar to the raw query data (evaluated using *scvi.criticism*). Reassuringly, we see very good performance in lung epithelial cells, airway epithelial cells, and epithelial cells from the back of the tongue (arguably similar to the upper airway epithelium). All of these organs are included in the HLCA reference dataset. For tissues that are not represented in the HLCA, we find a substantially worse correlation, indicating that the model does not capture their underlying distribution of gene expression well. Criticism, therefore, allows us to detect reference models that are well suited for the query dataset at hand and rely on those for downstream analysis.

B Methods

B.1 HubModel

Scvi-hub is implemented as a lightweight submodule within scvi-tools (*scvi.hub*), which provides an API for uploading and downloading pre-trained models to the Hugging Face Model Hub. To do so, scvi-hub uses the *huggingface_hub* Python API. The main construct within *scvi.hub*

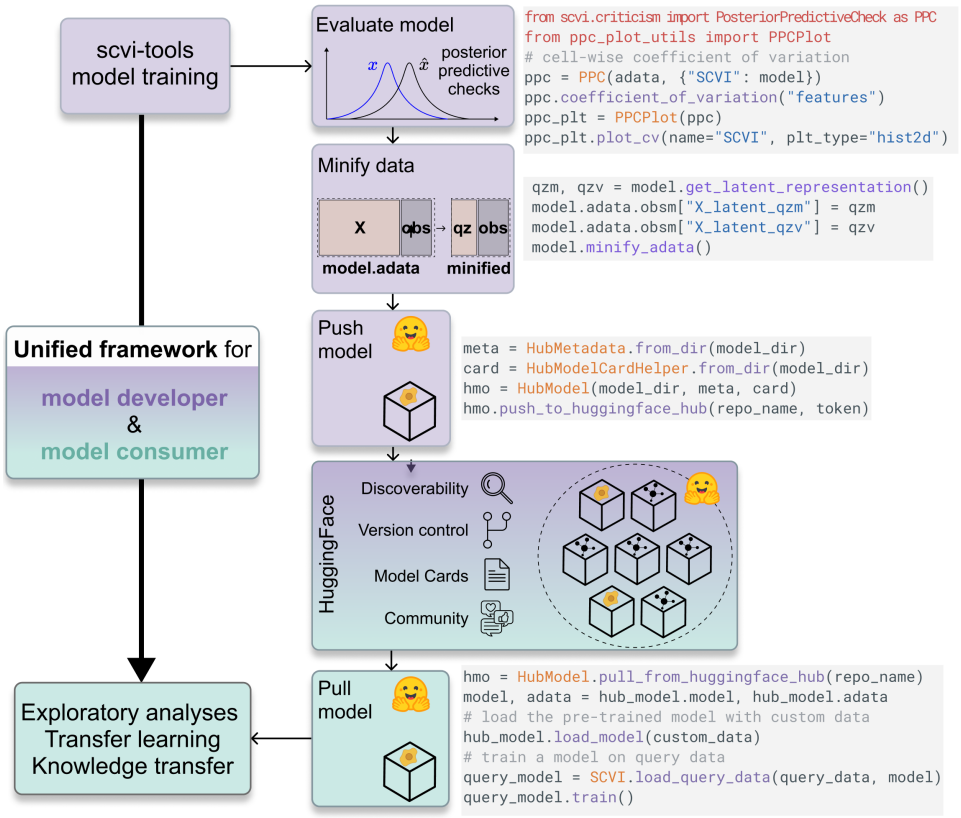


Figure 3: Figure 1: Overview of scvi-hub, depicting the new functionality for model contribution (purple) and retrieval (teal) that is implemented in scvi-hub, and its placement in an analysis workflow. A code snippet is presented for each of the tasks, highlighting the simplicity of this workflow. The workflow for model deposition (for model contributors) consists of training the model on reference data, evaluating model fit, minifying the reference data, and uploading the model to the scvi-hub. The training data (minified or raw) can be uploaded as well, either to Hugging Face or using any other online storage system (e.g., Zenodo) that can provide a direct link to access the data. This link is included in the Model card (visible through Hugging Face). The workflow for model retrieval (for the model consumer) consists of selecting the desired reference on scvi-hub and using the scvi-hub API for download and analysis. Note that the model consumers do not need to train a model on the reference data, and can also use the minified reference data instead of the raw form, thus lowering the current resource barriers (compute power, time and memory) for analysis. A single line of code switches from the minified version of the data to the full uploaded raw data to allow further downstream analysis.

is the *scvi.hub.HubModel* class, which represents a pre-trained scvi-tools model hosted on the Hugging Face Model Hub. An instance of this class has properties that can be used to load the model (*HubModel.model*) and data (*HubModel.adata*) into memory on demand (i.e. only when/if the property is invoked). To help with new model creation and upload, we also provide a *scvi.hub.HubModelCardHelper* class – which can be used to autogenerate a template Model Card for a new model to be uploaded to the hub – and a *scvi.hub.HubMetadata* class which encapsulates metadata required to be uploaded with the model to the hub. Model versioning is built into scvi-hub via usage of Hugging Face as our storage platform, since Hugging Face Hub Models are backed by git. Hub users can thus request a specific version of the model at download time via the revision argument to the *pull_from_huggingface_hub* function. We additionally support Hub Models uploaded to an AWS s3 bucket using *scvi.hub.HubModel.pull_from_s3*. This allows interacting with

models that are stored privately inside company buckets or shared with the community without the requirement to register to HuggingFace.

B.2 DATA MINIFICATION

We allow to store the latent posterior parameters (variance and mean in the latent space for each cell). This allows for faster generation of estimated gene expression as the inference function of scvi-tools models is not executed and instead the stored latent posteriors are used for the generative parts of the models. In addition, we can reduce the size of the data set by removing all count data from the AnnData object. This is performed by setting all expression values to zero and storing count objects as a sparse matrix (therefore zero entries in the sparse matrix). We additionally allow storing latent posteriors without removing the associated count data. A few functions in scvi-tools rely on computing the reconstruction loss and therefore require access to the full count data. These functions are not executable after data minification. All code used throughout the manuscript is compatible with data minification.

B.3 EVALUATION USING *scvi.criticism*

scvi.criticism is a new submodule in scvi-tools that can be used to efficiently compute model evaluation metrics for scvi-tools models. Models must inherit from the scvi's BaseModelClass class and implement the posterior_predictive_sample method, which samples from the generative distribution used by the scvi-tools model such as a negative binomial distribution. The main entry point for using *scvi.criticism* is the PPC class, which computes and stores various metrics for the provided collection of models. The PPC class can be initialized with one or a collection of models (for instance for model comparison), the raw counts, and a host of configuration options (such as number of posterior predictive samples to compute). It internally computes and stores a user-provided number of posterior predictive samples for each model (samples from the generative function of the model with user-provided library size). A crucial aspect of this package is integrated support for 3D sparse arrays. As the posterior predictive samples are n_cells x n_features x n_samples data cubes, their size grows linearly with the size of the dataset and the number of posterior predictive samples. We use the sparse (`https://github.com/pydata/sparse`) and xarray (`https://github.com/pydata/xarray`) Python packages to store the raw counts and posterior predictive samples in sparse format in memory, and only hydrate the data one batch at a time (this logic is implemented in the scvi-tools package). Supporting sparse arrays in *scvi.criticism* enables efficient evaluation of models trained on large-scale datasets.

Once initialized, the PPC class can be used to compute various metrics on the per-model posterior predictive samples. The class will store a collection of metrics keyed by metric names where each entry is a Pandas DataFrame holding results for all models that the class was initialized with.

We implemented the coefficient of variation, which is defined as the standard deviation over genes or cells divided by the mean over the same axis. Respectively, the user decides whether the metric is summarized over the different genes or different cells.

To generate Fig. 1C, we instantiated the PPC class with the HLCA raw counts data, the pretrained scANVI model, and a value of 2 for the number of posterior predictive samples. The plot shows the coefficient of variation metric computed across the "features" dimension, using the coefficient_of_variation method of the PPC class. We generated 2D histograms shown along with the best fit and identity lines. The method also prints out correlation measures (such as R2, Pearson and Spearman correlations, and mean absolute error) between raw and approximated results. Results can be stored as an obs column in AnnData.

We implemented the differential expression (DE) metric. For each of the posterior predictive samples differentially expressed genes are computed separately. The samples are normalized to the library size and transformed by log1p. We execute *scanpy.tl.rank_genes_groups* with a user-provided column for the cell-types and the respective method for computation, t-test by default. For each sample, we afterwards compute the F1 score of the top 100 overlapping genes, the mean absolute error (MAE), Pearson correlation and Spearman correlation between all estimated log2-fold changes and the AUC under the ROC curve and the average precision score for all genes with a significant p-value computed on the raw expression (significance threshold - pvalue_thresh by default 0.001). Fig. 1D

shows the differential expression metric computed using the differential_expression method of the PPC class, "ann_level_3" key for cell types and a pvalue_threshold of 0.2. We used the plot_diff_exp method of the ppc_utils function to generate the box plot shown. In Fig. 1E, we subsetted the dotplots to the immune cell subtypes, i.e. those with a cell type label in the following list: "B cell lineage", "Dendritic cells", "Macrophages","Mast cells", "Monocytes", "T cell lineage" and display the log-2 fold-changes of the respective top 2 marker genes. To generate the dotplots, we used one sample of the posterior predictive samples.

In Supplementary Fig. 1, we show the results of coefficient_of_variation (a) and differential_expression metrics (b) computed on a subset of the heart cell atlas dataset Litviňuková et al. (2020). We preprocessed the dataset as presented in an scvi-tools tutorial (`https://docs.scvi-tools.org/en/stable/tutorials/notebooks/api_overview.html`), then trained three models on the preprocessed data as follows: A model was trained with the default maximum number of epochs (400) and latent dimensions (10). The second model was trained with only five epochs (and the default number of latent dimensions), and the third model was trained with only two latent dimensions (and the default max number of epochs). (a) was generated in the same way as the 2D histogram shown in Fig. 1C. (b) was generated in a similar fashion to the Fig. 1D.

In Supplementary Fig. 2, we downloaded all epithelial cells from Tabula Sapiens (https://cellxgene.cziscience.com/collections/e5f58829-1a66-40b5-a624-9046778e74f5) and used those cells as query cells for the pre-trained HLCA scANVI model. We set the batch_key to the respective tissue to learn independent transfer mappings for each organ. We computed the CV over cells separately for each tissue and plotted the results similar to the top plot in Fig. 1C.

B.4 Reference-based analysis

For this analysis, we used the Human Lung Cell Atlas (HLCA) Sikkema et al. (2023) dataset and its pre-trained scANVI model. We used cellxgene_census to download the dataset (implemented in scvi-tools) and used for all analysis the raw, unnormalized counts. We used scvi-hub to download the model from the Hugging Face Model Hub. The author-provided UMAP embeddings are used and Scanpy is used for downstream analysis Wolf et al. (2018). The celltype UMAP is representative of the "ann_level_3" annotations.

B.5 Transfer learning analyses

For all models trained using transfer learning, we used following parameters surgery_epochs=500 (200 for emphysema dataset), early_stopping=True, early_stopping_monitor='elbo_train', early_stopping_patience=10, early_stopping_min_delta=0.001, weight_decay=0.0. For Tabula sapiens epithelial cells, each tissue was treated as a separate batch inside scANVI to allow comparisons between the various organs, while for both the emphysema data set and the cross-organ immune data set all cells were treated as a single batch.

We used the same HLCA reference model as in Fig. 1 as the reference data, and the data set from human lung emphysema as the query data. We used the scArches functionality implemented in scvi-tools to prepare and train a model on the query dataset. We then used the get_latent_representation method of the concatenated reference and query dataset to retrieve coordinates of the query data embeddings in the joint reference/query latent space. To generate Fig. 2A, we computed Scanpy's nearest neighbors in the combined latent space using n_neighbors=30 and UMAP using min_dist=0.3 in the RAPIDS implementation.

We transferred labels to the query dataset by first learning a nearest neighbors index on the latent space of the reference atlas, using the NNDescent class of the Pynndescent package, and then using the index to compute a nearest neighbor graph for the query dataset. We then used this graph to assign to each cell in the query dataset, a predicted cell type based on the reference dataset, along with a prediction uncertainty (we used the "ann_level_3" for the cell type annotation in the reference dataset). To this end, we converted nearest neighbor distances to affinities, and weighted the predictions using these affinities (this follows the approach used in the HLCA). For the emphysema dataset, we only use predictions for the confusion matrices with an uncertainty below 0.4 (4.8% of cells filtered out for emphysema dataset). For transferring labels from the query dataset to the refer-

ence dataset, the same function was used but the role of reference and query dataset were replaced. We used an uncertainty threshold of 0.2 (52.1% of cells filtered out of extended HLCA dataset) for the confusion matrix.

In Fig. 2C, we display summarized cell-type labels after recomputing UMAP using Scanpy with n_neighbors set to default. In Fig. 2D, we compute differential abundance using Milo between healthy and diseased cells from the query dataset using 100 neighbors in the latent embedding space, treating donor_id as the sample columns and not correcting for any covariates. Results are displayed for an FDR ¡ 0.1.

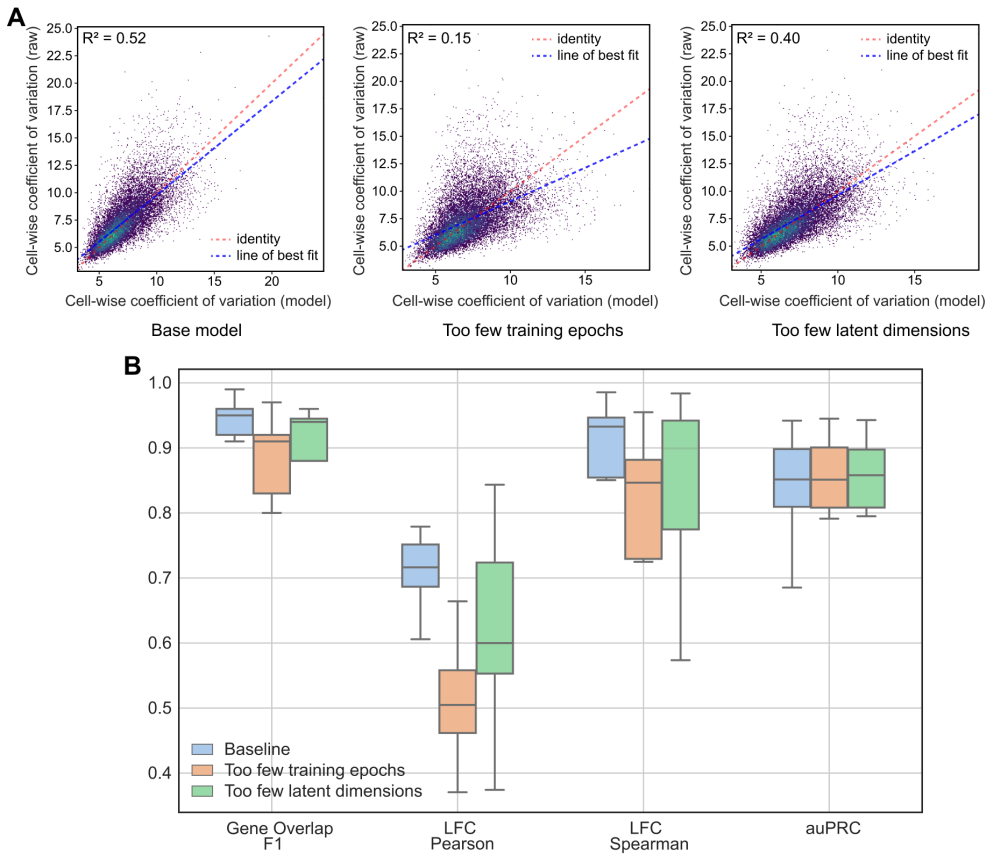
In Fig. 2E, to compute differentially expressed genes, we set idx1 to all fibroblast from diseased individuals and idx2 to fibroblast from healthy individuals. Fibroblasts were identified as all cells labeled as Distal fibroblast, Proximal fibroblast or Peribronchial fibroblast in the original publication. The differential_expression function of the HLCA model was used filtering for outlier cells and correcting the batch. Of note, weighting='Importance' requires access to the full expression values of the reference dataset and was deactivated here. Genes were selected to be expressed with an estimated mean (scale in scvi-tools) of ¿1e-4 in either of both groups. For the volcano plot decoupler-py was used highlighting the top 20 genes. The mean log-2 fold-changes are displayed on the x-axis and proba_not_de on the y-axis. In Fig. 2F, violin plots were created using Scanpy on library-size normalized and log1p-transformed raw gene expression values. In Supplementary Fig. 3, for PyDESeq2, we used decoupler-py and filtered out all genes with min_count=50 and min_total_count=150 grouped by disease status. Pseudobulks were computed per donor. All genes were used in the top panel, while for the bottom run we subsetted all genes to the ones incorporated in the reference model before running PyDESeq2. In Supplementary Fig. 4, we set the batch_key to the donor_id to remove additional batch correction from the reasons for outperforming PyDESeq2, we find similar performance. For the bottom plot, we set *unfreeze* True during query model training to retrain the whole network We find worse interpretability of predicted differentially expressed genes. For Fig. 2G, we downloaded all T cells and innate lymphoid cells from CELLxGENE Discover and subset this object to all cells originating from the lung. Analysis was performed as described above. Labels were infused to the extended HLCA dataset and confusion matrix is displayed in Fig. 2H. For Fig. 2I, marker genes were hand-selected and expression is displayed using library-size normalized and log1p raw expression values. For Fig. 2J, we subset to all cells labeled as Tem/Temra_CD8 and used PyDESeq2 without correcting for covariates.

B.6 Availability of data and materials

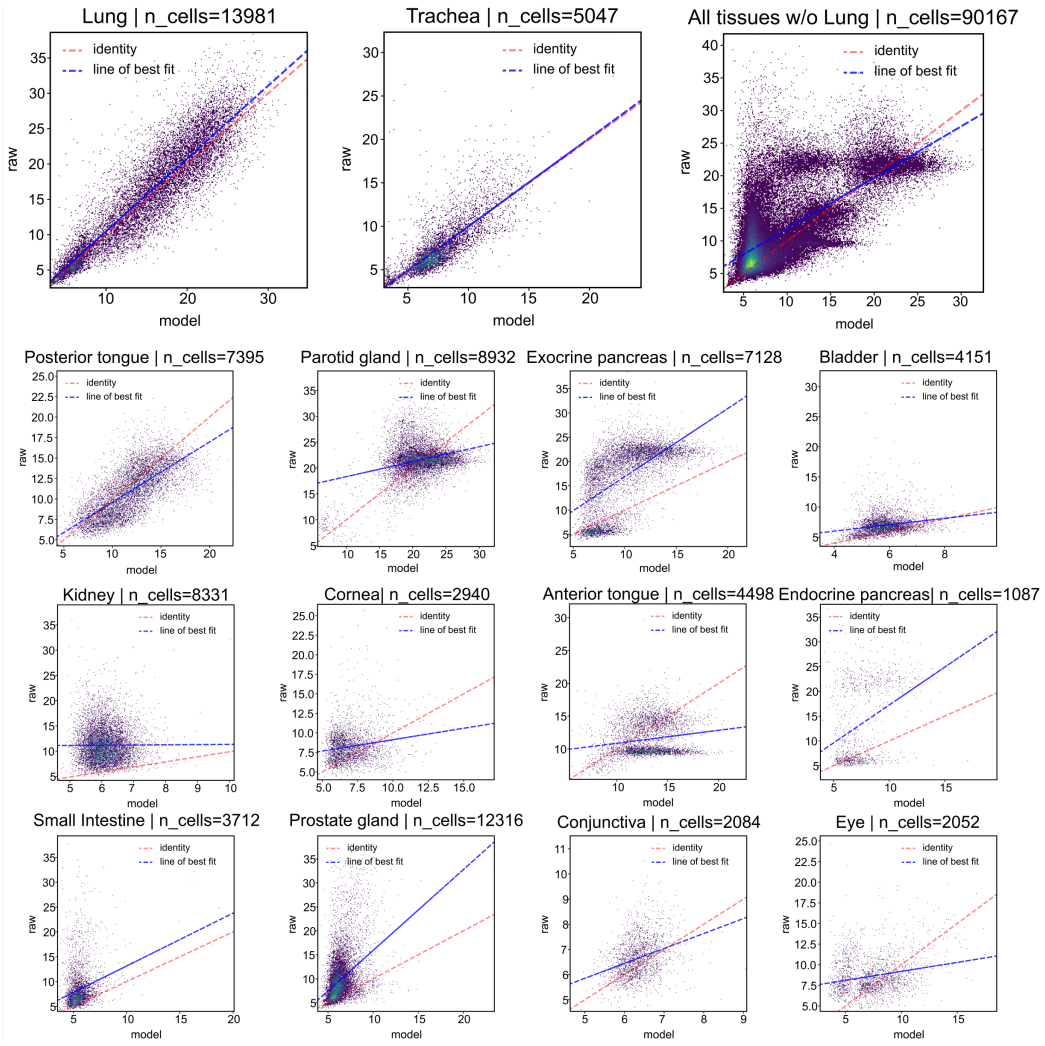
We used the Human Lung Cell Atlas (HLCA) dataset in our analysis, which can be found here (`https://cellxgene.cziscience.com/e/066943a2-fdac-4b29-b348-40cede398e4e.cxg/`) and their pre-trained scANVI model which can be found here (`https://zenodo.org/record/6337966/files/HLCA_reference_model.zip`). We provide tutorials for data minification (`https://docs.scvi-tools.org/en/stable/tutorials/notebooks/minification.html`), as well as how to implement this feature for newly developed latent variable models.

As additional single-cell datasets, we used several datasets from CELLxGENE Discover. Namely we accessed Tabula sapiens data at (`https://cellxgene.cziscience.com/e/53d208b0-2cfd-4366-9866-c3c6114081bc.cxg/`) and epithelial cells at (`https://cellxgene.cziscience.com/e/97a17473-e2b1-4f31-a544-44a60773e2dd.cxg/`). The emphysema dataset was downloaded from (`https://cellxgene.cziscience.com/collections/03cdc7f4-bd08-49d0-a395-4487c0e5a168`). All files were downloaded. The cell-type information of AT2 cells was used and the three separate datasets were concatenated and treated as one dataset for query analysis. The cross-tissue immune-cell dataset was downloaded from (`https://cellxgene.cziscience.com/e/ae29ebd0-1973-40a4-a6af-d15a5f77a80f.cxg/`). The Heart Cell Atlas was downloaded from (`https://cellxgene.cziscience.com/collections/b52eb423-5d0d-4645-b217-e1c6d38b2e72`). The Visium dataset for the prostate was downloaded from 10X (`https://cf.10xgenomics.com/samples/spatial-exp/2.0.0/Visium_FFPE_Human_Prostate_IF/Visium_FFPE_Human_Prostate_IF_spatial.tar.gz`). All datasets were accessed and used in the version of December 1st, 2023.

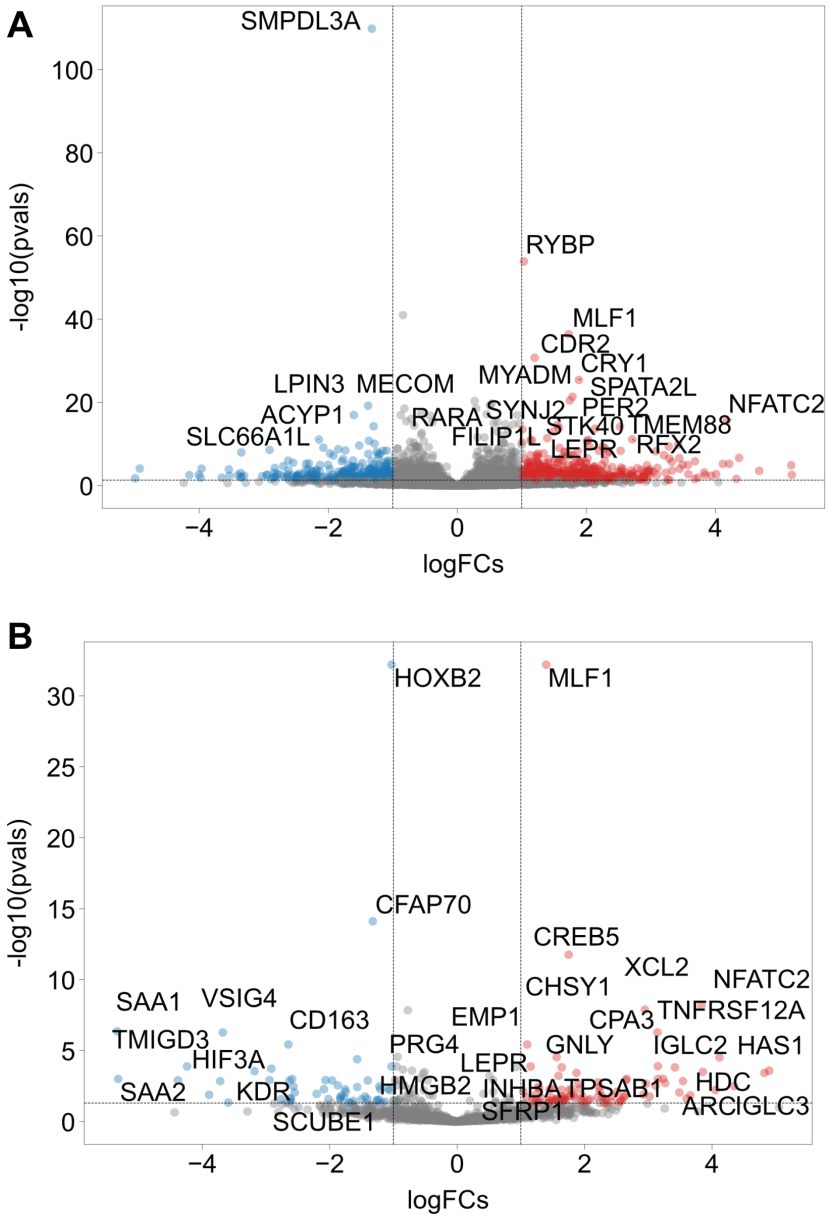
C SUPPLEMENTARY MATERIAL



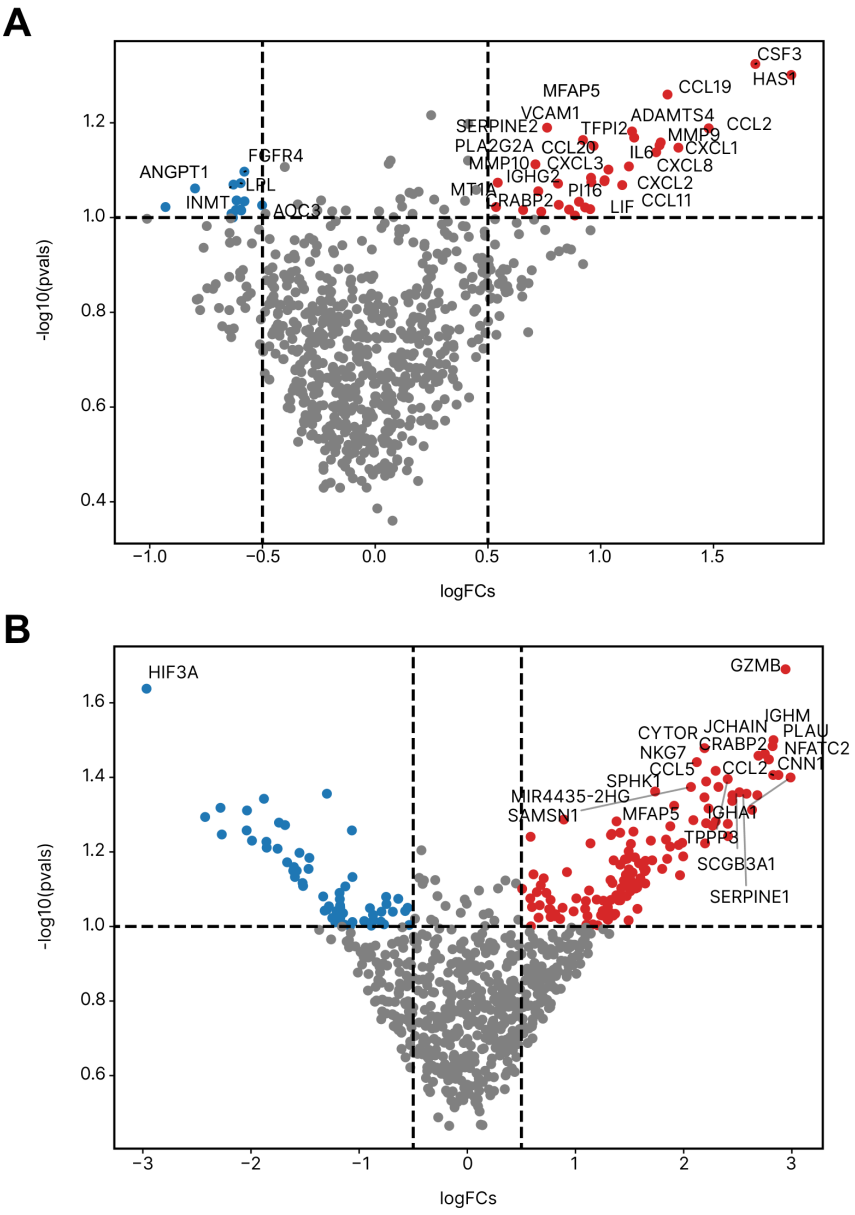
Supplementary Figure 1: (A) Cell-wise coefficient of variation for 3 different models trained on the Heart Cell Atlas data: the left model is the base model trained for 400 epochs with 10 latent dimensions, the middle model is trained for 5 epochs with 10 latent dimensions, the right model is trained for 400 epochs with 2 latent dimensions. The Pearson correlation coefficient is lower for both models trained with corrupted parameters. (B) Differential expression based metric with F1-score of top-100 genes between the different celltypes, second and third Pearson and Spearman correlation coefficient between all estimated log-2 fold-changes, fourth area under the precision-recall curve (auPRC). All metrics despite auPRC display a reduced performance for those corrupted models compared to the default model.



Supplementary Figure 2: Scvi-criticism detects query data not suited for a reference model. We trained a query model based on the HLCA reference model using all epithelial cells from Tabula sapiens. Coefficient of variation is displayed separately for each organ. For lung, trachea and posterior tongue (slightly worse) we find a high correlation of raw and model coefficient of variation. For all other tissues (top right) we find a much worse capturing of the CV by the trained model. For bladder, parotid gland, kidney, corneal, anterior tongue and eye there's no correlation between the model estimated CV and the raw data CV with the model estimated CV overestimating the raw data CV. For other organs like prostate gland and exocrine and endocrine pancreas the raw CV is in general higher than the model estimated CV with a low linear dependency. Of note, highly variable genes were selected initially based on lung tissue and this causes the small CV for cornea and eye epithelial cells.



Supplementary Figure 3: Pseudo-bulk differential expression test does not reveal similarly relevant differentially expressed genes. (A). Pseudo-bulk DE analysis considering all genes after filtering for lowly expressed ones. Genes listed as up-regulated are described to be linked with tumorigenesis. These genes have no specific role in fibroblasts and their role in emphysema remains unclear. (B) We additionally computed differentially expressed genes after subsetting to the genes used to train the HLCA model (i.e., the genes that were considered in the model-based DE in Fig. 3). Differentially expressed genes are enriched in canonical marker genes of other cell-types (*GNLY*, *XCL2* lymphoid cells; *TPSAB1* mast cells; *CD163*, *VSIG4*, *TMIGD3* in macrophages). They also include genes associated with P53 signaling such as *CREB5*, *NFATC2*, *CPA3* and *SFRP1*, which are expected to be upregulated in smoking individuals. These genes were also identified with the scvi-tools model-based analysis.



Supplementary Figure 4: Ablation study for differential expression function in scvi-tools. Both figures highlight differentially expressed genes with exactly the same settings as in Fig. 3E. (A) A query model was trained using the same setting as in Fig. 3E. The batch_key in scANVI was set to the donor ID instead of providing a single batch_key to the query model. In the DE function, we are not using transform_batch setting and instead generate the estimated expression using the original donor IDs. (B) A query model was trained using the same settings as in Fig. 3E. During *scvi.model.SCANVI.load_query_data*, we changed unfrozen=True, meaning that every component in the reference model is updated based on the query data. Bottom plots displays much more differentially expressed genes, with a high number of genes actually expressed by other cell-types than fibroblasts (mainly CD8 T cells and plasma cells).

REFERENCES

Tal Ashuach, Daniel A Reidenbach, Adam Gayoso, and Nir Yosef. PeakVI: A deep generative model for single-cell chromatin accessibility analysis. *Cell Rep Methods*, 2(3):100182, March 2022.

Pierre Boyeau, Jeffrey Regier, Adam Gayoso, Michael I Jordan, Romain Lopez, and Nir Yosef. An empirical bayes method for differential expression analysis of single cells with deep generative models. *Proc. Natl. Acad. Sci. U. S. A.*, 120(21):e2209124120, May 2023.

Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, 36(5):411–420, June 2018.

C Dominguez Conde, C Xu, L B Jarvis, D B Rainbow, S B Wells, T Gomes, S K Howlett, O Suchanek, K Polanski, H W King, L Mamanova, N Huang, P A Szabo, L Richardson, L Bolt, E S Fasouli, K T Mahbubani, M Prete, L Tuck, N Richoz, Z K Tuong, L Campos, H S Mousa, E J Needham, S Pritchard, T Li, R Elmentaite, J Park, E Rahmani, D Chen, D K Menon, O A Bayraktar, L K James, K B Meyer, N Yosef, M R Clatworthy, P A Sims, D L Farber, K Saeb-Parsy, J L Jones, and S A Teichmann. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594):eabl5197, 2022.

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$\times$GENE discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. November 2023.

Emma Dann, Neil C Henderson, Sarah A Teichmann, Michael D Morgan, and John C Marioni. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nat. Biotechnol.*, 40(2):245–253, February 2022.

Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Katherine Wu, Michael Jayasuriya, Edouard Melhman, Maxime Langevin, Yining Liu, Jules Samaran, Gabriel Misrachi, Achille Nazaret, Oscar Clivio, Chenling Xu, Tal Ashuach, Mohammad Lotfollahi, Valentine Svensson, Eduardo da Veiga Beltrame, Carlos Talavera-López, Lior Pachter, Fabian J Theis, Aaron Streets, Michael I Jordan, Jeffrey Regier, and Nir Yosef. scvi-tools: a library for deep probabilistic analysis of single-cell omics data. April 2021a.

Adam Gayoso, Zoë Steier, Romain Lopez, Jeffrey Regier, Kristopher L Nazor, Aaron Streets, and Nir Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat. Methods*, 18(3):272–282, March 2021b.

Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Valeh Valiollah Pour Amiri, Justin Hong, Katherine Wu, Michael Jayasuriya, Edouard Mehlman, Maxime Langevin, Yining Liu, Jules Samaran, Gabriel Misrachi, Achille Nazaret, Oscar Clivio, Chenling Xu, Tal Ashuach, Mariano Gabitto, Mohammad Lotfollahi, Valentine Svensson, Eduardo da Veiga Beltrame, Vitalii Kleshchevnikov, Carlos Talavera-López, Lior Pachter, Fabian J Theis, Aaron Streets, Michael I Jordan, Jeffrey Regier, and Nir Yosef. A python library for probabilistic analysis of single-cell omics data. *Nat. Biotechnol.*, 40(2):163–166, February 2022.

Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Stat. Sin.*, 6(4):733–760, 1996.

Y Gernez, R Tirouvanziam, and P Chanez. Neutrophils in chronic inflammatory airway diseases: can we target them and how? *Eur. Respir. J.*, 35(3):467–469, March 2010.

Laleh Haghverdi, Aaron T L Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.*, 36(5):421–427, June 2018.

Yuhan Hao, Tim Stuart, Madeline Kowalski, Saket Choudhary, Paul Hoffman, Austin Hartman, Avi Srivastava, Gesmira Molla, Shaista Madad, Carlos Fernandez-Granda, and Rahul Satija. Dictionary learning for integrative, multimodal, and scalable single-cell analysis. February 2022.

Lukas Heumos, Anna C Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte D Lücken, Daniel C Strobl, Juan Henao, Fabiola Curion, Single-cell Best Practices Consortium, Herbert B Schiller, and Fabian J Theis. Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.*, 24(8):550–572, August 2023.

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, Angela Oliveira Pisco, 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, Sheela Crasta, 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, Saroja Kolluru, William Kong, Maya E Kumar, Angera H Kuo, Rebecca Leylek, Baoxiang Li, Gabriel B Loeb, Wan-Jin Lu, Sruthi Mantri, Maxim Markovic, Patrick L McAlpine, Antoine de Morree, Maurizio Morri, Karim Mrouj, Shravani Mukherjee, Tyler Muser, Patrick Neuhöfer, Thi D Nguyen, Kimberly Perez, Ragini Phansalkar, Angela Oliveira Pisco, 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, Kyle J Travaglini, 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, Serena Y Tan, Angus Toland, Venkata N P Vemuri, Shaked Afik, Kyle Awayan, Olga Borisovna Botvinnik, Ashley Byrne, Michelle Chen, Roozbeh Dehghannasiri, Angela M Detweiler, Adam Gayoso, Alejandro A Granados, Qiqing Li, Gita Mahmoudabadi, Aaron McGeever, Antoine de Morree, Julia Eve Olivieri, Madeline Park, Angela Oliveira Pisco, Neha Ravikumar, Julia Salzman, Geoff Stanley, Michael Swift, Michelle Tan, Weilun Tan, Alexander J Tarashansky, Rohan Vanheusden, Sevahn K Vorperian, Peter Wang, Sheng Wang, Galen Xing, Chenling Xu, Nir Yosef, Marcela Alcántara-Hernández, Jane Antony, Charles K F Chan, Charles A Chang, Alex Colville, Sheela Crasta, Rebecca Culver, Les Dethlefsen, Camille Ezran, Astrid Gillich, Yan Hang, Po-Yi Ho, Juan C Irwin, Sori Jang, Aaron M Kershner, William Kong, Maya E Kumar, Angera H Kuo, Rebecca Leylek, Shixuan Liu, Gabriel B Loeb, Wan-Jin Lu, Jonathan S Maltzman, Ross J Metzger, Antoine de Morree, Patrick Neuhöfer, Kimberly Perez, Ragini Phansalkar, Zhen Qi, Poorvi Rao, Hayley Raquer-McKay, Koki Sasagawa, Bronwyn Scott, Rahul Sinha, Hanbing Song, Sean P Spencer, Aditi Swarup, Michael Swift, Kyle J Travaglini, Emily Trimm, Stefan Veizades, Sivakamasundari Vijayakumar, Bruce Wang, Wanxin Wang, Juliane Winkler, Jamie Xie, Andrea R Yung, Steven E Artandi, Philip A Beachy, Michael F Clarke, Linda C Giudice, Franklin W Huang, Kerwyn Casey Huang, Juliana Idoyaga, Seung K Kim, Mark Krasnow, Christin S Kuo, Patricia Nguyen, Stephen R Quake, Thomas A Rando, Kristy Red-Horse, Jeremy Reiter, David A Relman, Justin L Sonnenburg, Bruce Wang, 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, 2022.

Joyce B Kang, Aparna Nathan, Kathryn Weinand, Fan Zhang, Nghia Millard, Laurie Rumker, D Branch Moody, Ilya Korsunsky, and Soumya Raychaudhuri. Efficient and precise single-cell reference atlas mapping with symphony. *Nat. Commun.*, 12(1):5890, October 2021.

Peter V Kharchenko. The triumphs and limitations of computational methods for scRNA-seq. *Nat. Methods*, 18(7):723–732, July 2021.

Hanna Mendes Levitin, Jinzhou Yuan, Yim Ling Cheng, Francisco Ruiz, Jr, Erin C Bush, Jeffrey N Bruce, Peter Canoll, Antonio Iavarone, Anna Lasorella, David M Blei, and Peter A Sims. De novo gene signature identification from single-cell RNA-seq with hierarchical poisson factorization. *Mol. Syst. Biol.*, 15(2):e8557, February 2019.

Monika Litviňuková, Carlos Talavera-López, Henrike Maatz, Daniel Reichart, Catherine L Worth, Eric L Lindberg, Masatoshi Kanda, Krzysztof Polanski, Matthias Heinig, Michael Lee, Emily R Nadelmann, Kenny Roberts, Liz Tuck, Eirini S Fasouli, Daniel M DeLaughter, Barbara McDonough, Hiroko Wakimoto, Joshua M Gorham, Sara Samari, Krishnaa T Mahbubani, Kourosh Saeb-Parsy, Giannino Patone, Joseph J Boyle, Hongbo Zhang, Hao Zhang, Anissa Viveiros, Gavin Y Oudit, Omer Ali Bayraktar, J G Seidman, Christine E Seidman, Michela Noseda, Norbert Hubner, and Sarah A Teichmann. Cells of the adult human heart. *Nature*, 588(7838):466–472, December 2020.

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

Romain Lopez, Baoguo Li, Hadas Keren-Shaul, Pierre Boyeau, Merav Kedmi, David Pilzer, Adam Jelinski, Ido Yofe, Eyal David, Allon Wagner, Can Ergen, Yoseph Addadi, Ofra Golani, Franca Ronchese, Michael I Jordan, Ido Amit, and Nir Yosef. DestVI identifies continuums of cell types in spatial transcriptomics data. *Nat. Biotechnol.*, 40(9):1360–1369, September 2022.

Mohammad Lotfollahi, Mohsen Naghipourfar, Malte D Luecken, Matin Khajavi, Maren Büttner, Marco Wagenstetter, Žiga Avsec, Adam Gayoso, Nir Yosef, Marta Interlandi, Sergei Rybakov, Alexander V Misharin, and Fabian J Theis. Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.*, 40(1):121–130, August 2021.

Min-Seok Rha and Eui-Cheol Shin. Activation or exhaustion of CD8+ T cells in patients with COVID-19. *Cell. Mol. Immunol.*, 18(10):2325–2333, October 2021.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nat. Biotechnol.*, August 2023.

S D Shapiro. The macrophage in chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med.*, 160(5 Pt 2):S29–32, November 1999.

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, Marie-Jeanne Arguel, Leonie Apperloo, Martin Banchero, Christophe Bécavin, Marijn Berg, Evgeny Chichelnitskiy, Mei-I Chung, Antoine Collin, Aurore C A Gay, Janine Gote-Schniering, Baharak Hooshiar Kashani, Kemal Inecik, Manu Jain, Theodore S Kapellos, Tessa M Kole, Sylvie Leroy, Christoph H Mayr, Amanda J Oliver, Michael von Papen, Lance Peter, Chase J Taylor, Thomas Walzthoeni, Chuan Xu, Linh T Bui, Carlo De Donno, Leander Dony, Alen Faiz, Minzhe Guo, Austin J Gutierrez, Lukas Heumos, Ni Huang, Ignacio L Ibarra, Nathan D Jackson, Preetish Kadur Lakshminarasimha Murthy, Mohammad Lotfollahi, Tracy Tabib, Carlos Talavera-López, Kyle J Travaglini, Anna Wilbrey-Clark, Kaylee B Worlock, Masahiro Yoshida, Lung Biological Network Consortium, Maarten van den Berge, Yohan Bossé, Tushar J Desai, Oliver Eickelberg, Naftali Kaminski, Mark A Krasnow, Robert Lafyatis, Marko Z Nikolic, Joseph E Powell, Jayaraj Rajagopal, Mauricio Rojas, Orit Rozenblatt-Rosen, Max A Seibold, Dean Sheppard, Douglas P Shepherd, Don D Sin, Wim Timens, Alexander M Tsankov, Jeffrey Whitsett, Yan Xu, Nicholas E Banovich, Pascal Barbry, Thu Elizabeth Duong, Christine S Falk, Kerstin B Meyer, Jonathan A Kropski, Dana Pe'er, Herbert B Schiller, Purushothama Rao Tata, Joachim L Schultze, Sara A Teichmann, Alexander V Misharin, Martijn C Nawijn, Malte D Luecken, and Fabian J Theis. An integrated cell atlas of the lung in health and disease. *Nat. Med.*, 29(6):1563–1577, June 2023.

Zoë Steier, Dominik A Aylard, Laura L McIntyre, Isabel Baldwin, Esther Jeong Yoon Kim, Lydia K Lutes, Can Ergen, Tse-Shun Huang, Ellen A Robey, Nir Yosef, and Aaron Streets. Single-cell multi-omic analysis of thymocyte development reveals drivers of CD4/CD8 lineage commitment. April 2023.

Chenqu Suo, Emma Dann, Issac Goh, Laura Jardine, Vitalii Kleshchevnikov, Jong-Eun Park, Rachel A Botting, Emily Stephenson, Justin Engelbert, Zewen Kelvin Tuong, Krzysztof Polanski, Nadav Yayon, Chuan Xu, Ondrej Suchanek, Rasa Elmentaite, Cecilia Domínguez Conde, Peng He, Sophie Pritchard, Mohi Miah, Corina Moldovan, Alexander S Steemers, Pavel Mazin, Martin Prete, Dave Horsfall, John C Marioni, Menna R Clatworthy, Muzlifah Haniffa, and Sarah A Teichmann. Mapping the developing human immune system across organs. *Science*, 376(6597): eabo0510, June 2022.

Peter A Szabo, Pranay Dogra, Joshua I Gray, Steven B Wells, Thomas J Connors, Stuart P Weisberg, Izabela Krupska, Rei Matsumoto, Maya M L Poon, Emma Idzikowski, Sinead E Morris, Chloé Pasin, Andrew J Yates, Amy Ku, Michael Chait, Julia Davis-Porada, Xinzheng V Guo, Jing Zhou, Matthew Steinle, Sean Mackay, Anjali Saqi, Matthew R Baldwin, Peter A Sims, and Donna L Farber. Longitudinal profiling of respiratory and systemic immune responses reveals myeloid cell-driven lung inflammation in severe COVID-19. *Immunity*, 54(4):797–814.e6, April 2021.

Isaac Virshup, Sergei Rybakov, Fabian J Theis, Philipp Angerer, and F Alexander Wolf. anndata: Annotated data. December 2021.

Allon Wagner, Aviv Regev, and Nir Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.*, 34(11):1145–1160, November 2016.

Chaoqun Wang, Ben Hyams, Nancy C Allen, Kelly Cautivo, Kiara Monahan, Minqi Zhou, Madelene W Dahlgren, Carlos O Lizama, Michael Matthay, Paul Wolters, Ari B Molofsky, and Tien Peng. Dysregulated lung stroma drives emphysema exacerbation by potentiating resident lymphocytes to suppress an epithelial stem cell reservoir. *Immunity*, 56(3):576–591.e10, March 2023.

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, 19(1):15, February 2018.

Chenling Xu, Romain Lopez, Edouard Mehlman, Jeffrey Regier, Michael I Jordan, and Nir Yosef. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.*, 17(1):e9620, January 2021.