Title

SIMS: A deep-learning label transfer tool for single-cell RNA sequencing analysis

Journal

Cell Genomics, 4(6)

ISSN

2666-979X

Authors

Gonzalez-Ferrer, Jesus

Lehrer, Julian

O'Farrell, Ash

et al.

Publication Date

2024-06-01

DOI

10.1016/j.xgen.2024.100581

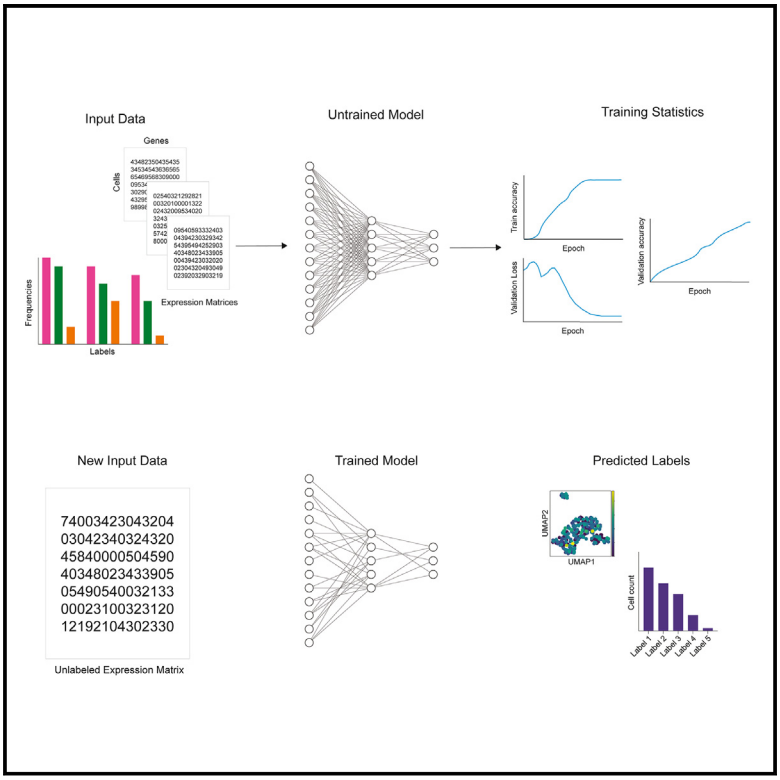
Peer reviewed

Cell Genomics

Technology

SIMS: A deep-learning label transfer tool for single-cell RNA sequencing analysis

Graphical abstract



Highlights

- SIMS is a low-code tool for label transfer in scRNA, removing barriers for adoption
- SIMS generalizes well to different types of tissues and sequencing technologies
- SIMS can identify misannotated cells, including stressed cells in brain organoids
- SIMS facilitates sharing of trained models, streamlining collaboration

Authors

Jesus Gonzalez-Ferrer, Julian Lehrer, Ash O'Farrell, ..., David Haussler, Vanessa D. Jonsson, Mohammed A. Mostajo-Radji

Correspondence

vjonsson@ucsc.edu (V.D.J.), mmostajo@ucsc.edu (M.A.M.-R.)

In brief

Gonzalez-Ferrer et al. introduce SIMS (scalable, interpretable machine learning for single cell), an end-to-end, low-code, data-efficient machine-learning pipeline that accurately classifies single-cell RNA data, including complex and imbalanced datasets. SIMS is then applied to primary developing and adult brain datasets as well as stem cell-derived brain models. SIMS predicts cell identities, unveils genetic variations, and rectifies misannotations, even when cell types are obscured by stress signals, making it a versatile and robust tool for single-cell classification.

Gonzalez-Ferrer et al., 2024, Cell Genomics *4*, 100581
June 12, 2024
https://doi.org/10.1016/j.xgen.2024.100581

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Technology

SIMS: A deep-learning label transfer tool for single-cell RNA sequencing analysis

Jesus Gonzalez-Ferrer,[1,2,3,7] Julian Lehrer,[1,2,4,7] Ash O'Farrell,[1] Benedict Paten,[1,3] Mircea Teodorescu,[1,3,5] David Haussler,[1,3] Vanessa D. Jonsson,[3,4,6,*] and Mohammed A. Mostajo-Radji[1,2,6,8,*]

[1]Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA 95060, USA
[2]Live Cell Biotechnology Discovery Lab, University of California, Santa Cruz, Santa Cruz, CA 95060, USA
[3]Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA 95060, USA
[4]Department of Applied Mathematics, University of California, Santa Cruz, Santa Cruz, CA 95060, USA
[5]Department of Electrical and Computer Engineering, University of California, Santa Cruz, Santa Cruz, CA 95060, USA
[6]Senior author
[7]These authors contributed equally
[8]Lead contact
*Correspondence: vjonsson@ucsc.edu (V.D.J.), mmostajo@ucsc.edu (M.A.M.-R.)
https://doi.org/10.1016/j.xgen.2024.100581

SUMMARY

Cell atlases serve as vital references for automating cell labeling in new samples, yet existing classification algorithms struggle with accuracy. Here we introduce SIMS (scalable, interpretable machine learning for single cell), a low-code data-efficient pipeline for single-cell RNA classification. We benchmark SIMS against datasets from different tissues and species. We demonstrate SIMS's efficacy in classifying cells in the brain, achieving high accuracy even with small training sets (<3,500 cells) and across different samples. SIMS accurately predicts neuronal subtypes in the developing brain, shedding light on genetic changes during neuronal differentiation and postmitotic fate refinement. Finally, we apply SIMS to single-cell RNA datasets of cortical organoids to predict cell identities and uncover genetic variations between cell lines. SIMS identifies cell-line differences and misannotated cell lineages in human cortical organoids derived from different pluripotent stem cell lines. Altogether, we show that SIMS is a versatile and robust tool for cell-type classification from single-cell datasets.

INTRODUCTION

Next-generation sequencing systems have allowed for large-scale collection of transcriptomic data at the resolution of individual cells. Within these data lies variability allowing us to uncover cell-specific features, such as cell type, cell state, and regulatory networks, as well as to infer trajectories of cell differentiation and specification.[1,2] These properties are crucial to understanding biological processes in healthy and diseased tissue. In addition, these properties better inform the development of *in vitro* models, which are often benchmarked against cell atlases of primary tissue.[1]

The lowering costs of sequencing, coupled with several barcoding strategies, have allowed single-cell datasets and atlases to scale with respect to cell and sample numbers as well as data modalities.[3] Yet, despite the increasing size and complexity of datasets, the most popular pipelines for single-cell analysis are based on dimensionality reduction and unsupervised clustering followed by manual interpretation and annotation of each cell cluster.[4] This requires a high level of expertise in understanding the most appropriate cell markers for a given tissue, a major barrier to newcomers to a field. For highly heterogeneous tissues such as the brain, where a consensus in cell-type nomenclature remains challenging,[5] manual cell annotation can introduce additional errors.

Errors in cell annotation may be driven by the following common assumptions. (1) It is assumed that marker genes are uniformly highly expressed, which is not always the case.[6,7] For instance, while OPALIN and HAPLN2 are considered markers of oligodendrocytes in the brain, their expression is low or undetectable in a large subset of oligodendrocytes at the single-cell level.[8] Indeed, high levels of HAPLN2 have been proposed as a landmark of Parkinson's disease.[9] (2) It is assumed that cell-type marker gene expression is constant throughout development, such that a gene that specifically labels a population of cells at one age would label the same population at a different age. For example, while it is known that PVALB-positive cortical interneurons are born during embryonic development,[10] the expression of this gene is not seen until well after birth.[11] Notably, recent studies have shown that a subset of PVALB interneurons may never express the PVALB gene.[12] (3) It is assumed that gene markers discovered in one species apply to others. In several tissues, including the brain, there are major species-specific differences. For example, HCN1 is a key marker of cortical layer-5 subcerebral projection neurons in the mouse, but it is highly expressed in projection neurons of all

cortical layers in humans.[13,14] In summary, manual annotation of every new dataset based on standard marker genes can lead to compounding error propagation and inconsistent single-cell atlases, potentially reducing their utility.

The development of software to automate single-cell analysis has become an important and popular research topic.[4,15–17] However, the accuracy of these automated classifiers often degrades as the number of cell types increase, and the number of samples per label becomes small.[18] The distribution of cell types is often asymmetric, with a majority class dominating a high percentage of cells. Additionally, technical variability between experiments can make robust classification between multiple tissue samples difficult. There have been efforts to apply statistical modeling to this problem,[19,20] but the high-dimensional nature of transcriptomic data makes analysis statistically and computationally intractable.[21] These conditions make applying classical models such as support vector machines difficult and ineffective.[22] In response, generative neural networks have become a popular framework due to their robustness to technical variability within data, scalability, and ability to capture biological variation in the latent representation of the inputs.[23–25] These include deep-learning models based on variational inference,[26,27] adversarial networks,[28] and attention transformers.[25] Early deep-learning models exhibit a lack of interpretability due to their "black box" architecture.[18] However, explainable artificial intelligence (XAI) research aims to understand model decision making by assigning weight values to the genes based on their influence on cell-type predictions. Despite this, some deep-learning approaches display inherent biases favoring multivariate gene selection that impedes straightforward data interpretation.[25,29] Additionally, the computational demands of certain deep-learning systems may preclude adoption by smaller research groups lacking access to high-performance computing infrastructure. Ongoing work seeks to enhance model interpretability and efficiency to enable broader use across the biological sciences.[25,28]

Here we present SIMS (scalable, interpretable machine learning for single cell), a new framework based on the model architecture found in TabNet.[30] SIMS is implemented in PyTorch Lightning,[31] which allows SIMS to be low-code and easy to use. We take advantage of the fact that TabNet uses a sequential self-attention mechanism, which allows for interpretability of tabular data.[30] Importantly, TabNet does not require any feature preprocessing and has built-in interpretability, which visualizes the contribution of each feature to the model.[30] Given these properties, SIMS is an ideal tool to classify RNA sequencing data. We show that SIMS either outperforms or is on par with state-of-the-art single-cell classifiers. This high performance is evident in complex imbalanced datasets, such as peripheral blood samples, full body atlases, and heart, kidney, and lung datasets. We apply SIMS to datasets of the adult mammalian brain and show a high accuracy even with a small number of cells in the training set (<3,500 cells). In the developing brain, SIMS identifies neurons undergoing postmitotic fate refinement. We further apply SIMS to data generated from *in vitro* models, such as pluripotent stem cell-derived cortical organoids. Using the SIMS pipeline, we were able to reclassify mislabeled cells through the use of label transfer from annotated primary tissue. Moreover, we discovered that in cortical organoids, cell stress impairs the proper specification of early postmitotic excitatory projection neurons, but not inhibitory interneurons, in a cell-line-dependent manner. Altogether, we propose SIMS as a new label transfer tool, capable of robust performance with deep annotation and skewed label distributions, high accuracy with small and large datasets, and direct interpretability from the input features.

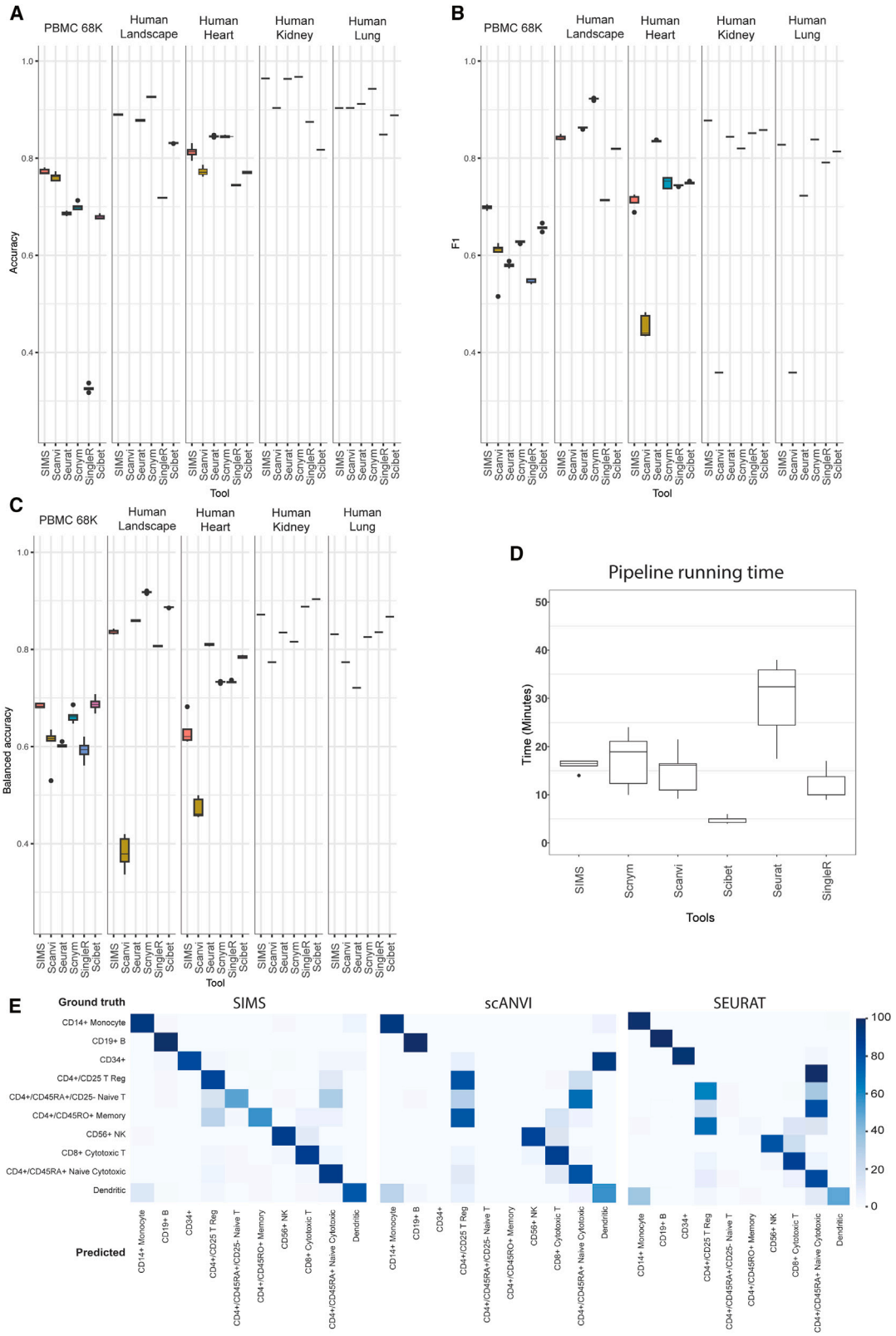
DESIGN

We developed SIMS, a framework for label transfer across single-cell RNA datasets that uses TabNet as the classifier component (Figure S1).[30] TabNet is a transformer-based neural network with sparse feature masks that allow for direct prediction interpretability from the input features.[30] The TabNet model offers several unique features that render it suitable for single-cell data analysis: It employs a sparse attention mechanism that selects only a subset of genes to predict each cell type. This design choice enhances interpretability, particularly valuable for single-cell data given its inherent high dimensionality. The sparse feature attention allows users to quantitatively understand which genes are most critical for prediction. Additionally, TabNet was designed to require minimal preprocessing, as the nonlinearity of the network allows it to capture complex combinations of input features while the sparsity allows for generalization.[30]

To better fit the model for the task of single-cell classification, we added two innovations. First, we included temperature scaling, a postprocessing step of the trained network that provides users with a calibrated probability measure for the classification of each cell in the selected cell type.[32] This feature enables the discovery of cell types not present in the reference sample. We then equipped our pipeline with an automated gene intersection mechanism, allowing the prediction of datasets with a different number of genes than the dataset used for training the model, a common occurrence when different sequencing technologies or experimental protocols are used. This automated intersection serves to ensure that both datasets have the same set of genes, facilitating direct comparison.

In our framework, for each forward pass, batch normalization is applied. The encoder consists of several steps (parameterized by the user) of self-attention layers and learned sparse feature masks. The decoder then takes these encoded features and passes them through a fully connected layer with batch normalization and a generalized linear unit activation.[33] Interpretability by sample is then measured as the sum of feature mask weights across all encoding layers.

SIMS can be trained with either one or several preannotated input datasets, allowing for the integration of atlases generated by the same group or by different groups. For accurate training, the user must input an annotated matrix of gene expression in each cell. After training and production



(legend on next page)

of training statistics, the user can input a new unlabeled dataset. Of note, if the training data were normalized ahead of training, the user must normalize the unlabeled data in a similar manner. For example, one of the most common approaches is to normalize the transcript counts per cell and then do a logarithmic transformation. If the reference dataset underwent this transformation and was then used to train SIMS, the query dataset should undergo the same normalization. The model will then predict the cluster assignment for each cell. SIMS will then output the probability of each cell belonging to each cluster, where the probability is more than 0.

SIMS is accessible through a Python API. The development version can be found on our GitHub repository at the following link: https://github.com/braingeneers/SIMS. Additionally, a pip package is also available for easy installation: https://pypi.org/project/scsims/. SIMS is designed to require minimal input from the users (Data S1). To train the model, the user has to only input the data file of the training dataset and a file with the labels, and define the class label; the user can also choose to load the dataset into Scanpy as an AnnData object (Figure S2). This process will save the learned parameters for each training epoch in a new file.

To perform the label transfer on a new dataset, the user must import the weights from the trained model. The user will then input the new unlabeled dataset (Figure S3).

SIMS takes the cell-by-gene-expression matrix as an input. For newly produced data, we recommend an end-to-end pipeline we have developed within Terra. This pipeline takes raw FASTQ files, runs them through the CellRanger or StarSolo Dockstore workflows[34–36] (Figure S4), and outputs an expression matrix in the h5 format. This file type can then be read as an annotated dataset with the scanpy.read() function. The pipeline then classifies the cell types using a SIMS model trained on the reference dataset of interest. This pipeline can also be used to benchmark new methods in an unbiased manner or to reproduce results obtained from data stored in the Sequence Read Archive with an additional Dockstore workflow step.[37,38]

To extend the reach of SIMS to investigators without coding experience, we developed a web application based on Streamlit. This application allows users to perform predictions based on pretrained SIMS models. To access the web application, the user has to enter the web page at https://sc-sims-app.streamlit.app/. Once there, the user has to upload their dataset of interest in h5ad format, select one of our pretrained models, and perform the predictions. They will be able to download the predictions in csv format and visualize their labeled data as a uniform manifold approximation and projection (UMAP). The user will also be able to obtain the genes selected for the model for cell-type classification. The web application deployed in Streamlit cloud can accept h5ad files up to 1 GB in size. This matches the upload size allowed in Azimuth, a well-known reference-based classifier built on top of Seurat v4.[19] Some of the key differences are the faster inference times as shown in Figure 1 and the ability for the community to upload and rapidly share pretrained models.

RESULTS

Benchmarking SIMS against existing cell classifiers of single-cell RNA data

We conducted benchmark tests in five distinct datasets to evaluate SIMS's performance against other methods built on various theoretical approaches. The first dataset we used was the PBMC68K, also known as Zheng68K, derived from human peripheral blood mononuclear cells.[39] This dataset is particularly valuable due to its complex nature, featuring imbalanced cell clusters and cells with similar molecular identities, making it a robust choice for benchmarking cell-type annotation methods, as it has been extensively employed for this purpose. As a second dataset we included the human heart dataset, also known as Tucker's dataset, comprising 11 cell types and exhibiting imbalanced cell clusters.[40] This dataset shares similarities with PBMC68K but contains a significantly larger number of cells (287,000 cells compared to 68,000 cells).

For the third dataset we used the human lung atlas dataset, also known as Krasnow's dataset.[41] Benchmarking on this dataset showcases the ability of the tools to classify cells independently of sequencing technology, as the dataset comprised cells obtained from two different sequencing technologies: 10× and SmartSeq2. It was also interesting, as it contained 58 different cell types coming from three donors with a similar size to PBMC68K but more cell variety. The fourth dataset we included was a human kidney dataset, also known as Stewart's dataset. This dataset was interesting from the donor variability and batch effect perspective, as it contains 34 different cell types coming from 14 different donors, totaling 40,268 mature kidney cells.

Additionally, we incorporated the human landscape dataset, also known as Han's dataset,[18] into our analysis, primarily for its substantial size (more than 584,000 cells) and the presence of a wide array of different cell types coming from the entire body, totaling 102. Another interesting characteristic of this dataset was the lower dimension from the feature perspective, as it was only sampling around 5,000 genes, in contrast to the

Figure 1. Benchmarking SIMS against other cell classifiers

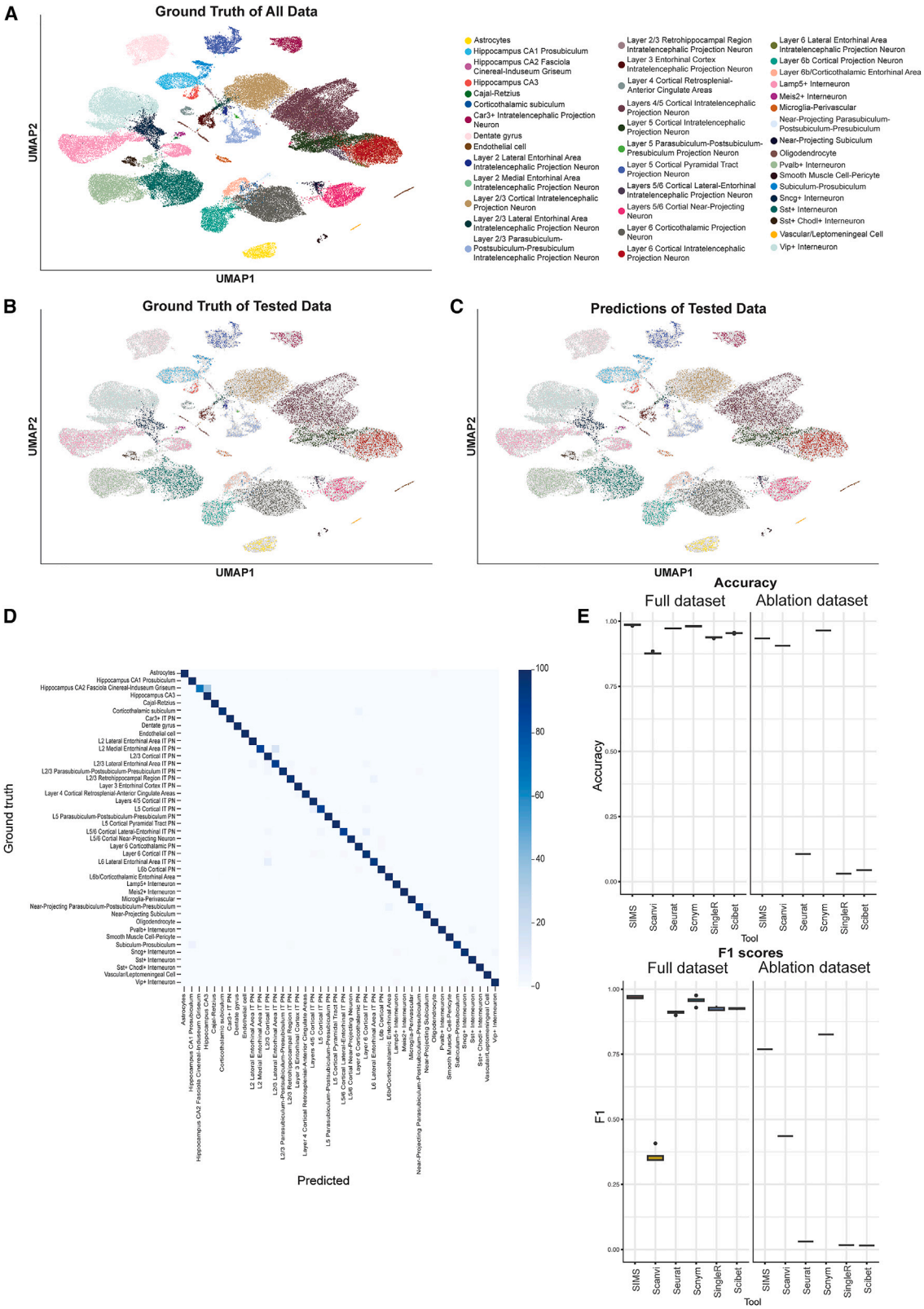
(A) Performance of cell-type annotation methods measured by accuracy in five selected datasets using 5-fold cross-validation. PBMC68K n = 68,450 cells; human landscape n = 584,000 cells; human heart n = 287,369; human kidney n = 40,268 cells; human lung n = 75,400 cells. Box plots show the median (center lines), interquartile range (hinges), and 1.5-times the interquartile range (whiskers).

(B) Performance of cell-type annotation methods measured by macro F1 in five selected datasets using 5-fold cross-validation.

(C) Performance of cell-type annotation methods measured by balanced accuracy in five selected datasets using 5-fold cross-validation.

(D) Performance of cell-type annotation methods measured by pipeline running time in minutes.

(E) Heatmap for PBMC68K comparing ground-truth annotations and predictions by SIMS, scANVI, and Seurat.



(legend on next page)

other datasets comprising from around 20,000 to 45,000 features.

In our benchmarking study, we selected a range of tools that represent diverse methodologies and functionalities within the field of single-cell analysis. The scVI and scANVI pipelines were included owing to their deep-learning foundation, using a variational autoencoder to create cell embeddings.[27] This latent representation serves as the basis for subsequent model building and label transfer, making scVI and scANVI essential benchmarks for evaluating deep-learning-based approaches in single-cell analysis, illustrating the scArches package.[24] Another deep-learning-based tool, Scnym, adopts another two-step process. Beginning with adversarial pretraining, the network is refined through fine-tuning for classification, offering a unique perspective on how deep-learning models can be optimized for single-cell RNA data analysis.[28] In contrast, Scibet adopts a non-deep-learning approach by fitting multinomial models to the mean expression of marker genes. Scibet was benchmarked primarily for its inference speed, a crucial aspect considering its real-time web-enabled inference capabilities.[42] Seurat, a well-established framework in the field, was included due to its versatility in preprocessing, visualization, and analysis of single-cell data. Additionally, Seurat provides label transfer functionality through the identification of anchors, establishing pairwise correspondences between cells in different datasets.[19] Another reason behind the choice to benchmark against this tool is that Seurat is the main engine behind Azimuth, a well-known web application for no-code single-cell RNA label transfer. We also wanted to evaluate a model with a simpler paradigm behind it, SingleR, which employs a correlation-based method, focusing on variable genes in the reference dataset for calculating differences between cell types. Additionally, an attempt was made to benchmark against scBERT, a large transformer-based model.[25] However, due to its computational complexity, we faced limitations. Despite experimenting with an A10 GPU, scBERT's demands were such that we were unable to train or evaluate it on any dataset, even with a minimal batch size of 1. These carefully chosen tools enabled a comprehensive evaluation on considering various approaches and methodologies in the realm of single-cell analysis.

To ensure the robustness of our findings and mitigate the influence of randomness, we employed a 5-fold cross-validation strategy. Notably, SIMS consistently outperformed the majority of label transfer methods in terms of accuracy and Macro F1 score (Figures 1 and S5; Table S1) across these diverse datasets. This compelling evidence underscores SIMS as a highly accurate and robust classifier, demonstrating its proficiency and its ability to generalize across diverse tissue types. Additionally, SIMS exhibits scalability to accommodate a large number of cells and showcases its ability to effectively classify datasets with imbalanced cell types. This ability is important, as imbalanced datasets have been noted to heavily impact downstream analysis and are known to be difficult to annotate.[43]

We also conducted a consistent evaluation of pipeline running times by employing 5-fold cross-validation to assess the speed of the benchmarked tools in minutes, using the same comparison methodology (Figure 1E). This analysis was carried out within the National Research Platform clusters,[44] leveraging user-accessible GPUs. Whenever feasible, training and inference processes were executed on GPUs; otherwise, they were performed on CPUs.

SIMS accurately performs label transfer in highly complex single-cell data: Mouse adult cerebral cortex and hippocampus

Given that SIMS outperforms most state-of-the-art label transfer methods in different datasets, we then asked whether it could perform accurately in a highly complex tissue, such as the brain. We focused on adult mouse cortical and hippocampal data generated by the Allen Brain Institute.[45–47]

The cerebral cortex is among the most complex tissues due to its cellular diversity, the variety and scope of its functions, and its transcriptional regulation.[48] The cerebral cortex is organized in six layers and several cortical areas, each with different composition and proportions of excitatory projection neurons (PNs), inhibitory interneurons (INs), glial cells, and other non-neuronal cell types.[48] The hippocampus, on the other hand, is part of the archicortex (also known as the allocortex).[49] It is further subdivided into cornu ammonis, dentate gyrus, subiculum, and entorhinal area.[49] While the hippocampus also has a layered structure made of three layers, the cell-type composition and numbers vary greatly from those in the cerebral cortex.[49] The great diversity of cell types, the close relationship between some of those subtypes, and the anatomical separation between these regions make cerebral cortex and hippocampal datasets complex but attractive benchmarking models to test SIMS.

The dataset contained 42 cell types, including PNs, INs, and endothelial and glial cells. Training in 80% of the cells selected at random and testing on the remaining 20%, we find that SIMS performs at an accuracy of 97.6% and a Macro F1 score of 0.983 (Figures 2 and S6).

We then performed ablation studies to investigate the performance of SIMS. We find that training in as little as 7% of the dataset (3,285 cells) is sufficient to obtain a label transfer accuracy of over 95% and Median F1 score of over 0.95 (Figure S7). The Macro F1 after training in 7% of the data is 0.90 (Figure S7).

Figure 2. Application of SIMS to single-cell RNA sequencing: Adult mouse cerebral cortex and hippocampus

(A) Ground-truth UMAP representation for the dataset (n = 73,347 cells).

(B) Ground-truth UMAP representation for the subset of cells used for testing the algorithm in the train-test split.

(C) Predictions made by SIMS in that subset of data.

(D) Confusion matrix for the test split. L, layer; IT, intratelencephalic; PN, projection neuron.

(E) Performance of cell-type annotation methods measured by accuracy and Macro F1 in the full Allen mouse dataset and its ablation, using 5-fold cross-validation. Box plots show the median (center lines), interquartile range (hinges), and 1.5-times the interquartile range (whiskers).

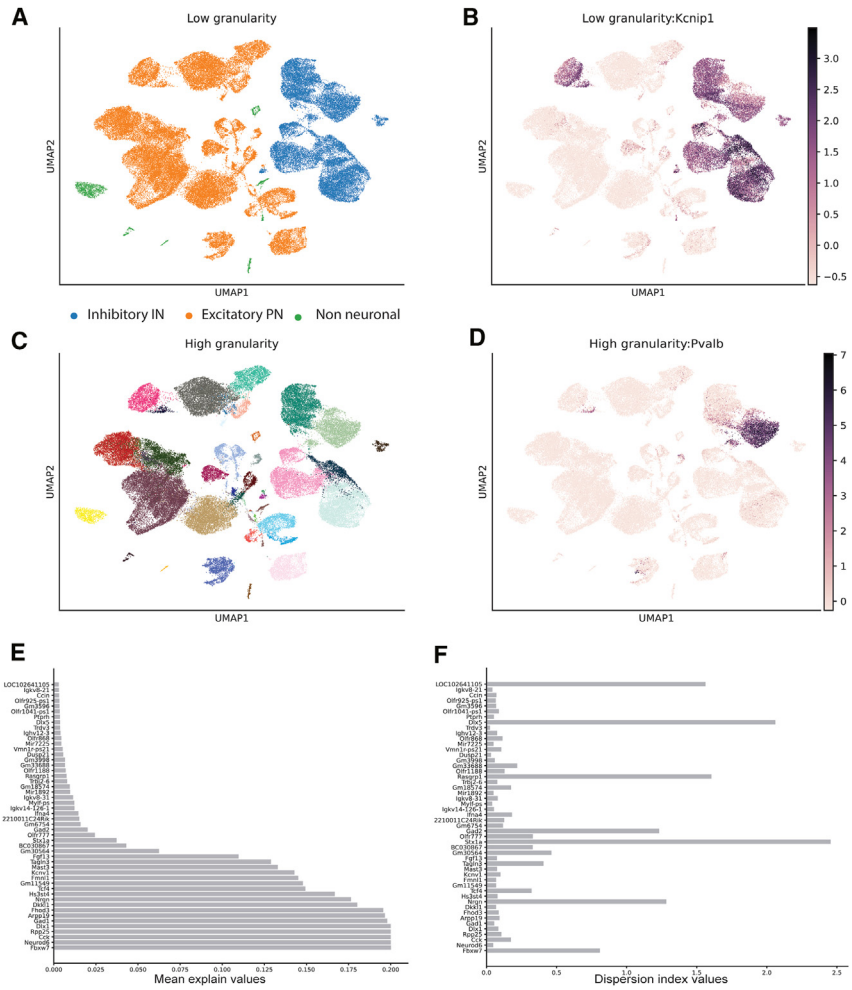


Figure 3. SIMS explainability

(A) UMAP representation of the Allen mouse dataset colored by macro cell type.

(B) UMAP representation of the Allen mouse dataset colored by expression of the selected gene by SIMS for the GABAergic group.

(C) UMAP representation of the Allen mouse dataset colored by cell type. Same naming convention as used for Figure 2A.

(D) UMAP representation of the Allen mouse dataset colored by expression of the selected gene by SIMS for the Pvalb^{+} interneuron group.

(E) Mean explain value for the top 50 genes across 300 runs.

(F) Dispersion index value for the top 50 genes across 300 runs.

We then used the same dataset splits to benchmark the other computational methods (scANVI, Seurat, Scnym, SingleR, and SciBet). We find that while SIMS maintains high accuracy and Macro F1 scores, most of the other methods perform poorly with a reduced number of cells in the training data (Figure 2). We conclude that SIMS is a data-efficient machine-learning model.

SIMS provides interpretability by computing weights for sparse feature masks in the encoding layer. These weights indicate the most influential genes in the network's decision making for assigning cell types. To assess this interpretability, we generated three dataset partitions with varying levels of granularity. Our aim was to observe whether the network could accurately select pertinent genes to distinguish the groups formed at each resolution level. To analyze the results, we focused in the Pvalb^{+} INs, a group of inhibitory neurons born in the medial ganglionic eminence (MGE). For the lowest level of granularity, which limits the cell options to INs, PNs, and non-neuronal cells, we find that for the INs group some important genes selected by the model were Kcnip and Igf1 (Figures 3A and 3B), both of which have been previously shown to be important IN genes.[50–52] For the medium level of granularity (medial ganglionic eminence, nonmedial ganglionic eminence) and consistent with previous literature, we find that for the MGE-derived INs the genes selected were Rpp25, Dlx1, Dlx5, Gad1, Ffg13, and Cck[53–55] (Figure S8). For the highest level of granularity (Pvalb^{+} INs), some of the selected genes were Satb1, Pvalb, Lypd6, Dlx6os-1, and Bmp3[55] (Figures 3C and 3D).

To confirm that the selection of the most important genes was consistent across different runs, we performed the experiment with the highest level of granularity 300 times. For each experiment, we normalized each gene weight against the highest weight gene measured in that run and measured the mean weight and dispersion index for each gene across all runs (Figures 3E and 3F). Given the explainability matrix $E \in R^{n \times m}$ composed of m genes measured across n cells, we select all rows representing cells with the same predicted label and compute

$$e_i = \frac{1}{n_l} \sum_{j=1}^{n_i} E_{ij}, i = 1, 2, \ldots, n_l.$$

We then average e_i across all 300 runs. To calculate the dispersion index, we first measure the average importance of each gene across all 300 runs:

$$g = \frac{1}{m} \sum_{i=1}^{m} E_{i,j}\, i = 1, 2, \ldots, n$$

and then compute the dispersion index as

$$disp_{gene} = e_{gene} / g_{gene}.$$

In the top ten genes more important for classification, we can find excitatory PN markers (Neurod6), inhibitory IN markers (Cck, Rpp25, Dlx1, Gad1), neural progenitor-related genes (Fbxw7), and genes related to different neuropsychiatric

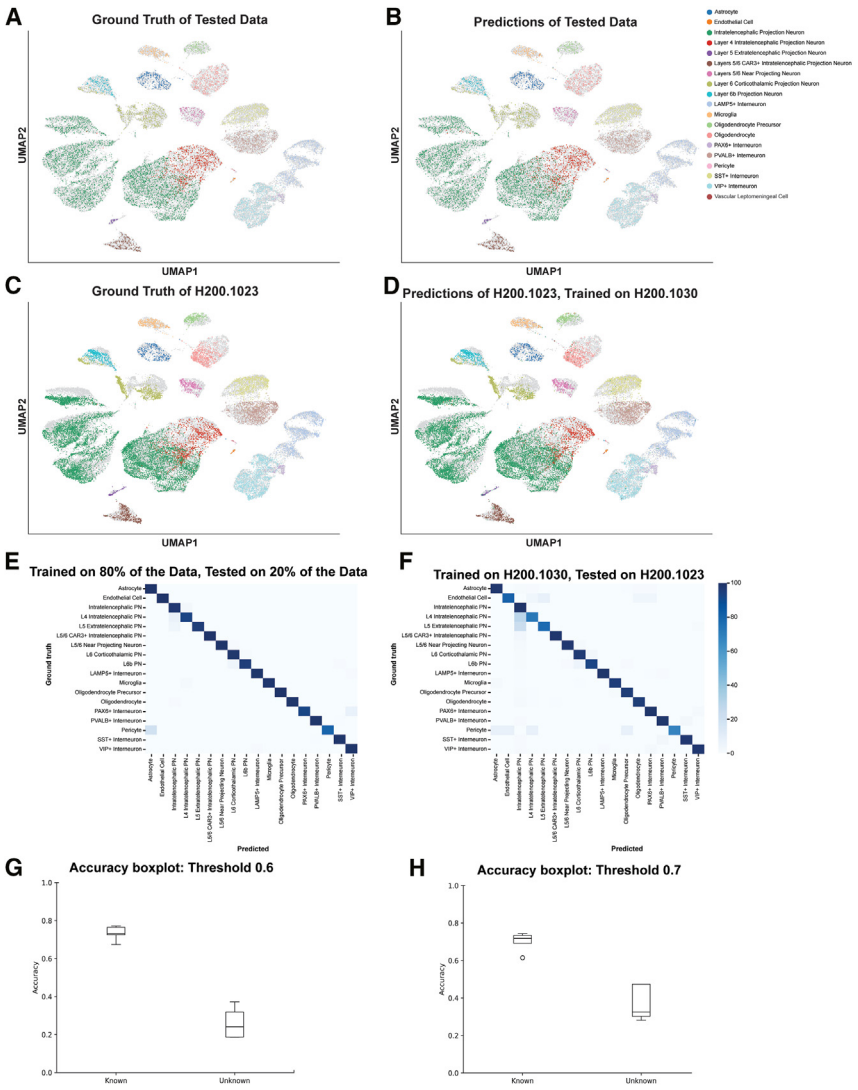


Figure 4. Application of SIMS to trans-sample predictions of single-nuclei RNA sequencing: Adult human cerebral cortex

(A) Ground truth for the test-split data (n = 49,495 cells).

(B) Predictions for the test-split data.

(C) Ground truth for the H200.1023 sample (n = 18,511 cells).

(D) Prediction for the H200.1023 sample after training on the H200.1030 sample.

(E) Confusion matrix for the test split (data percentage).

(F) Confusion matrix for the test split (H200.1023).

(G) Accuracy box plot for the known and unknown cell classification with a confidence threshold of 0.6.

(H) Accuracy box plot for the known and unknown cell classification with a confidence threshold of 0.7.

L, cortical layer; PN, projection neuron. Additional examples are shown in Figure S12.

disorders (Arpp19, Fhod3, Nrgn). Top genes show mean explain values of around 0.2 (Figure 3E); for comparison, the mean explain value for the median gene is approximately 10^{-6} (Figure S9). This showcases the consistency of gene selection by SIMS and how it could be used to find clinically relevant genes overlooked by conventional methods.

SIMS accurately performs trans-sample label transfer in highly complex single-nuclei data: Human adult cerebral cortex

Single-nuclei RNA sequencing has become an important emerging tool in the generation of atlases, particularly for tissues from which obtaining single cells is difficult. Cell nuclei are used in neuroscience because live adult neurons are difficult to isolate, due to their high connectivity, sensitivity to dissociation enzymes, and high fragility, often resulting in datasets with abundant cell death, low neuronal representation, and low-quality RNA.[56] Importantly, single-nuclei sequencing is compatible with cryopreserved banked tissue.[57] Yet the data generated in single-nuclei RNA sequencing is not necessarily similar to the data generated in single-cell RNA sequencing. For instance, a recent study comparing the abundance of cell-activation-related genes in microglia sequenced using single-cell and single-nuclei technologies showed significant differences between both datasets.[58] Moreover, single-nuclei datasets are more prone to ambient RNA contamination from the lysed cells.[59] In the case of the brain, it has been observed that neuronal ambient RNA has masked the transcriptomic signature of glial cells, leading to incorrect classification of glia subclasses in existing atlases.[59]

Given the high label transfer accuracy of SIMS in single-cell data, we then tested its performance in single-nuclei datasets. As a proof of principle, we selected the human adult cerebral cortex dataset generated by the Allen Brain Institute.[45,46] We trained on 80% of the data and tested the model in the remaining 20%. Overall, we obtained an accuracy of 98.0% and a Macro F1 score of 0.974 (Figures 4 and S10; Table S2).

We then performed a data-ablation study and observed that we obtained over 95% accuracy using as little as 7% of data for training (2,124 cells). Similarly, we obtained a Macro F1 score of over 0.95 with 9% (2,731 cells) of the data and a median F1 of over 0.95 with 8% of the data (2,428 cells) for training (Figure S11).

We then asked how SIMS performs in trans-sample predictions. This dataset is made up of three different postmortem samples, namely: H200.1023, a 43-year-old Iranian-descent woman; H200.1025, a 50-year-old Caucasian male; and H200.1030, a 57-year-old Caucasian male. We trained the model on one sample and tested it on the other two samples. We performed this experiment in each possible combination, obtaining accuracies ranging from 93.1% to 95.8% (Figures 4 and S12; Tables S2–S8). As shown, SIMS predicts the label

accurately for most cell types across samples. SIMS shows a decrease in performance when trying to classify pericytes, as sometimes it labels them as astrocytes (Tables S3–S8). This is consistent with recent work showing that previously annotated single-nuclei atlases of the brain often mask non-neuronal cell types.[59] In addition, we observed that layer-4 intratelencephalic neurons often get classified as generic intratelencephalic neurons (Tables S3–S8). This is in agreement with the fact that layer-4 intratelencephalic neurons are a subset of intratelencephalic neurons.[60] We also employed this dataset to assess the capacity of SIMS to differentiate between recognized cell types and those not included in the training dataset. This capability holds significance, as it can function as a surrogate metric for identifying cells in new datasets that were absent from the reference dataset used for training. In this particular scenario, we implemented a leave-one-out methodology, whereby we excluded one cell type from the training dataset and then made predictions on the test set, encompassing all of its cell types. Subsequent to temperature scaling, we used the model's probability outputs as a measure of confidence, such that a probability of 0.5 approximately measures that the model possesses a 50% level of confidence in the predicted cell type's accuracy. Following this, we established a user-adjustable threshold to determine whether the cell type should be labeled as the predicted cell type or categorized as an unknown cell type (Figures 4G and 4H). Altogether, we conclude that SIMS is a powerful approach to perform intra-sample and trans-sample label transfer in complex and highly diverse tissues such as the adult brain.

SIMS can accurately classify cells during neuronal specification

Having established that SIMS can accurately predict cell labels in complex tissues, we then asked how our model performed in predicting cells of different ages. Classifying cells during development is challenging, as several spatiotemporal dynamics can mask the biological cell identities.[61] During cortical development, gene networks of competing neuronal identities first colocalize within the same cells and are further segregated postmitotically,[48,62,63] likely through activity-dependent mechanisms.[64,65]

To test the accuracy of SIMS at classifying developing tissue, we focused on mouse cortical development due to its short timeline.[66] In the mouse cortex, neurogenesis starts at embryonic day 11.5 (E11.5), and it is mostly completed by E15.5.[66] Common C57BL/6 laboratory mice are born at E18.5.[67] Neonatal mice are timed based on the postnatal day.[67] We took advantage of a cell atlas of mouse cortical development that contains two samples of E18 mouse embryos and two samples of postnatal day 1 (P1) mice.[62] These timed samples, which are around 1 day apart from each other, represent time points at which all mouse neurogenesis is completed.[66] At these time points, neurons may still be undergoing fate refinement[68] and consequently retain fate plasticity, albeit limited.[69–71]

First, we trained a model on one E18 and one P1 sample and tested the accuracy of label transfer in two samples, one of each age (Figures S13A and S13B). Across 17 cell types, we find that the model predicts the labels with an accuracy of 84.2% with a Macro F1 score of 0.791 (Figure 5A and Table S9).

We then tested SIMS by training on two P1 samples and testing the label transfer in two E18 samples (Figures S13C and S13D). We find that in this experiment, the label transfer accuracy drops to 73.6% and the Macro F1 score to 0.674 (Figure 5B and Table S10). Interestingly, however, this drop in accuracy is not random, but either follows the developmental trajectories of the misclassified cells or misclassifies cells as transcriptomically similar cell types. For example, astrocytes are a subtype of glial cells that retain the ability to divide throughout life.[72] Indeed, the major source of astrocytes in the cerebral cortex is other dividing astrocytes.[72] Consequently, the "cycling glia cells" cluster is often predicted as astrocytes (Figure S13). In the neuronal lineage, we find that SIMS can accurately predict most cell types. Going back to the combined-ages model, we focused on layer-4 neurons, which is one of the neuronal subtypes with the lowest accuracy in label transfer (24.31%). We find that these neurons are often classified as upper-layer callosal PNs and rarely as callosal PNs of the deep layers (Figures 6B–6E). While morphologically distinct, layer-4 neurons share transcriptional homology with callosal PNs.[62,73] Indeed, recent work has shown that layer-4 neurons transiently have a callosal-projecting axon, which is postmitotically eliminated during circuit maturation, well after P1.[60] In agreement, layer-4 neurons that are mislocalized to the upper cortical layers retain an upper-layer callosal PN identity and fail to refine their identity.[74] By comparing the gene expression of upper-layer callosal PNs, the correctly classified layer-4 neurons, and the misclassified layer-4 neurons, we observe that while upper-layer callosal PNs and correctly classified layer-4 neurons have the gene-expression patterns appropriate to their identity, misclassified layer-4 neurons have an intermediate expression of genes that define the identity of the other two cell types, such as Rorb[75] (Figure 5). Notably, most (90.1%) of the misclassified layer-4 neurons belong to E18, likely representing neurons undergoing fate refinement. Altogether, this example highlights the difficulty that cell classifiers face when trying to discretely label cells during development.

Together, we conclude that SIMS can accurately predict cell labels of specified neurons. However, when applying SIMS during periods of differentiation and fate refinement, it uncovers similar identities in the developmental trajectories. This is likely caused by transcriptomic similarities that can often mask proper identification. Alternatively, SIMS may identify subtle differences in fate transitions that cannot be accurately pinpointed by traditional clustering methods in the reference atlases.

SIMS identifies cell-line differences in gene expression in human cortical organoids

Cortical organoids are a powerful tool to study brain development, evolution, and disease.[13,76,77] However, like many pluripotent stem cell-derived models, cortical organoids are affected by cell-line variability and culture conditions that can affect the reproducibility of the protocols.[78] Moreover, transcriptomic

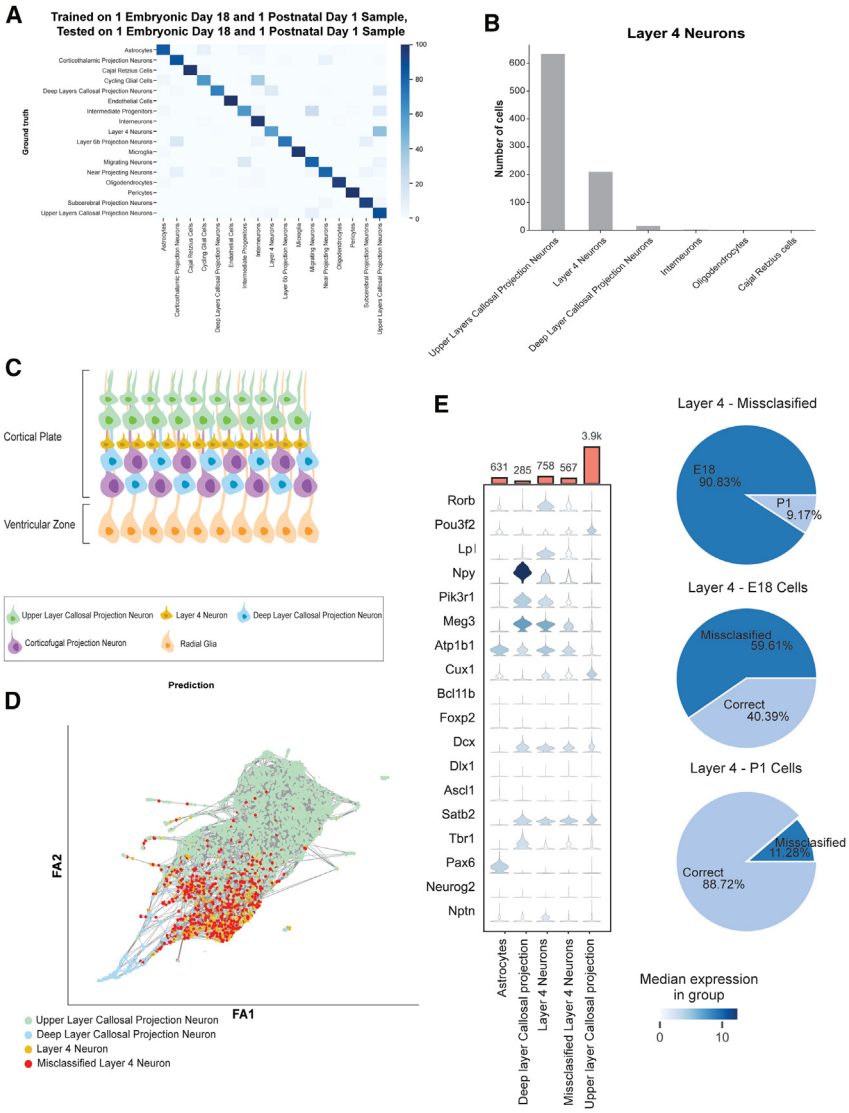


Figure 5. Application of SIMS to developing tissue: Mouse cerebral cortex

(A) Confusion matrix for E18P1 split, where we trained on sample 1 E18 and sample 1 P1 and predicted on sample 2 E18 and sample 2 P1 (n = 20,209 cells).

(B) Bar plot showing the number of layer-4 cells that get predicted as the different cell types.

(C) Diagram of the mouse cerebral cortex after neurogenesis.

(D) Force atlas representation of layer-4 neurons.

(E) Violin plot showing gene expression in the misclassified layer-4 group compared to the groups that is classified as layer 4.

analysis of cortical organoids has revealed strong signatures of cell stress,[79–81] which can impair proper cell-type specification.[82] In addition, *in vitro* conditions generate cell types of uncharacterized identity that do not have an *in vivo* counterpart.[80,83] While some have argued that these cells should be removed from further analysis,[83] the most common approach is to annotate them as "unknown" cell clusters.[76]

To understand whether SIMS could be used to uncover cell-line differences and identify different trajectories, we used a dataset from 6-month-old human cortical organoids derived from three different cell lines (three organoids per batch), each with their own idiosyncrasy.[76] Specifically, this dataset contained: (1) one batch of cortical organoids derived from the 11A cell line, in which all cells had been identified and no cell was labeled as "unknown"; (2) one batch of cortical organoids derived from the GM8330 cell line, which contained a small number of "unknown" cells and a large proportion of immature INs; and (3) two batches of cortical organoids derived from the PGP1 cell line, which contained major batch effects. One of those batches had a large number of "unknown" cells and cells of poor quality and was therefore dropped from further analysis (Figures 6A, 6B, and S14).

We performed label transfers between organoids generated from the three cell lines. We first performed an intra-cell-line label transfer using the 11A organoids. We trained on two organoids and predicted the cells on a third organoid. We find an accuracy of 86.0% and a Macro F1 score of 0.794 (Figure S15). We then performed trans-cell-line predictions training on 11A and predicting the cell types of the other lines. We obtained an accuracy of 71.3% and a Macro F1 score of 0.564 when predicting cells from PGP1 organoids and an accuracy of 67.4% and a Macro F1 score of 0.570 when predicting cells from GM8330 organoids. We observe a high degree of accuracy for most cell types tested, including cycling cells, intermediate progenitor cells, outer radial glia/astroglia, immature INs, ventral precursors, and callosal PNs (Table S11). Interestingly, radial glial cells (RGs) from both PGP1 and GM8330 cell lines often were classified as immature PNs. Specifically, we find that 82% of the PGP1 and 42% of the GM8330 RGs are predicted as immature PNs when the data are trained on the 11A cell line (Figures 6C and 6D). Strikingly, only 1.9% of PGP1 RGs and 3.9% of GM8330 RGs are predicted as RGs. These results suggest major differences in gene expression between the RG annotated cells across cortical organoids derived from different cell lines.

Previous work has shown that cell stress in organoids impairs proper fate acquisition of PNs.[82] We therefore took advantage of Gruffi, a recently developed tool to annotate stressed cells in human neuronal tissue.[83] Overall, we find that organoids derived from the GM8330 cell line showed the biggest percentage of stressed cells (16.67%), while organoids derived from the PGP1 and 11A cell lines had 6.6% and 4.9% of stressed cells,

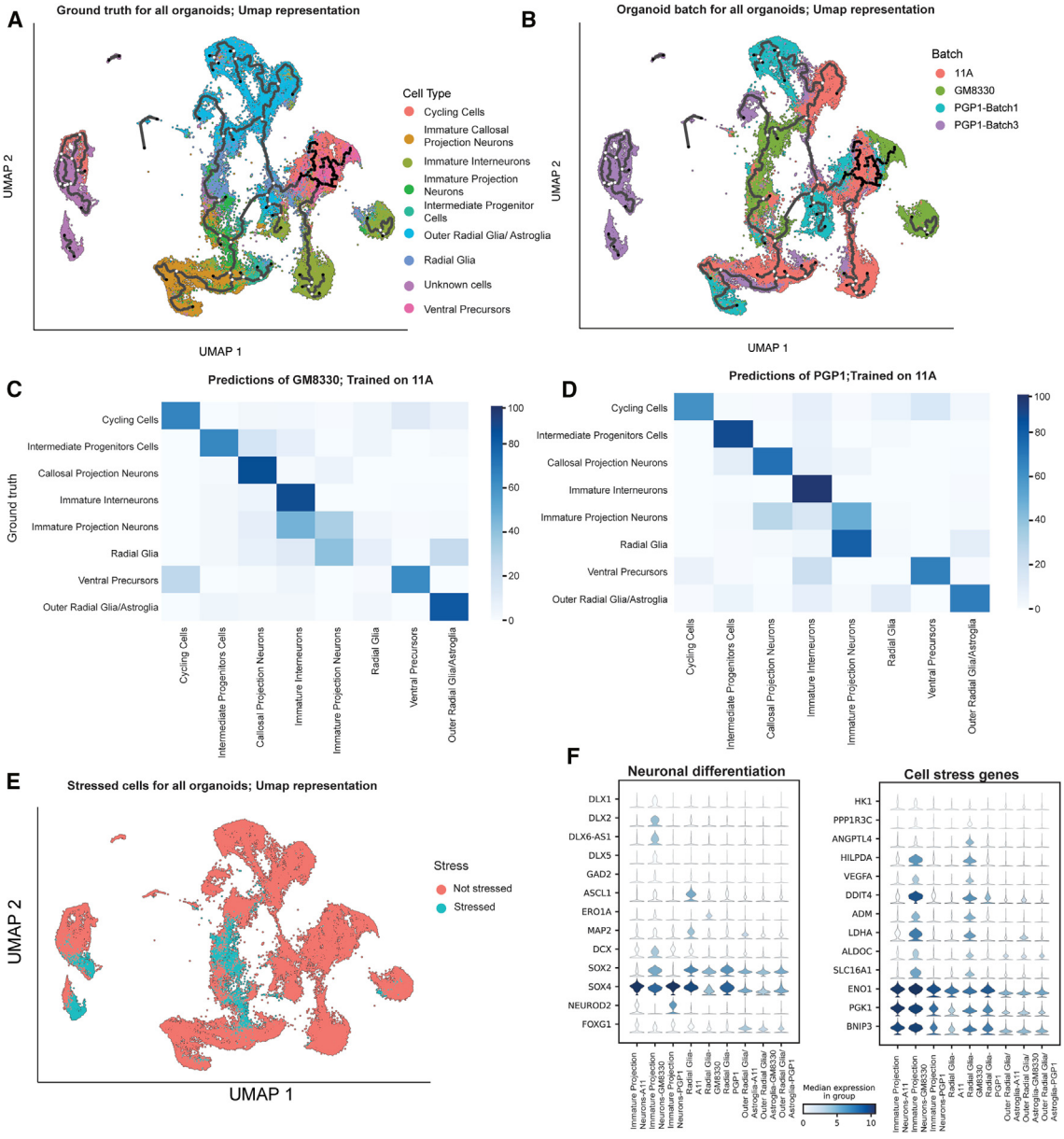


Figure 6. Application of SIMS to *in vitro* generated models: Human cortical organoids

(A) UMAP representation of the ground-truth cell type for all cell lines. (n = 87,863 cells).

(B) UMAP representation of the batch and cell line for all cell lines.

(C) Confusion matrix for GM3880-derived organoids, model trained on 11A-derived organoids.

(D) Confusion matrix for PGP1-derived organoids, model trained on 11A-derived organoids.

(E) UMAP representation of stressed cells as annotated by Gruffi in all organoids.

(F) Violin plots for neuronal differentiation and cell-stress genes showing differences among cell lines.

respectively. (Figure 6E). To understand whether the stressed cells were responsible for the misclassification, we removed these cells from the 11A training set. We then performed a new round of label transfers. Using this approach, we find that 56% of PGP1-derived RGs and 27% of GM8330-derived RGs continue to be classified as immature PNs. Importantly, only 7.2% of PGP1-derived and 14% of GM8330-derived RGs are predicted as RGs.

We then removed the stressed cells from both the training and the predicted datasets and found that 44% of PGP1-derived and 14% of GM8330-derived RGs are classified as immature PNs. Notably, the number of RGs that are properly classified as such remains similar, with only 6.9% of PGP1-derived and 19% of GM8330-derived RGs properly predicted. Altogether, these results suggest that cell stress alone cannot explain the differences in cell expression between RGs of cell lines.

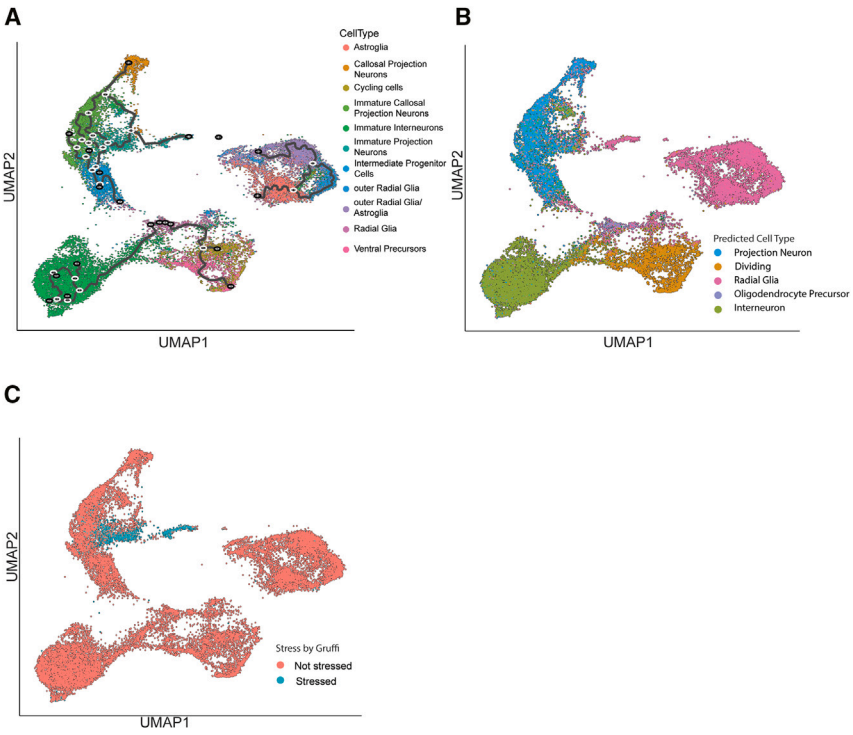


Figure 7. Application of SIMS to *in vitro* generated models: Human cortical organoids

(A) UMAP representation of the ground-truth cell type for 11A organoids (n = 25,618 cells).

(B) UMAP representation of the label transfer from fetal tissue for 11A organoids.

(C) UMAP representation of stressed cells as annotated by Gruffi in 11A organoids.

SIMS identifies improperly annotated cell lineages in human cortical organoid atlases

Given that label transfer between human cortical organoids derived from different cell lines poorly predicted the RG cell type, we then focused on assessing the most common predictions for this cell type after stressed cells were removed from both the training and the prediction datasets. While in the PGP1 line the majority of the misclassified RGs are immature PNs, the second most common cell prediction is the closely related outer radial glia/astroglia cell type. On the other hand, for the GM8330 cell line the most commonly predicted cell type is immature INs. Unlike RGs, outer radial glia/astroglia, and immature PNs that belong to the dorsal telencephalic lineage, INs are derived from the distinct and distant ventral telencephalon.[48] A deeper analysis of the GM8330 cell line reveals that 65% of the immature PNs also get predicted as immature INs (Figure 6C), indicating a consistent misclassification between neuronal lineages in the GM8330 cell line. We then performed a Wilcoxon test rank for differential expression analysis between the three cell lines. We found that, unlike the other cell lines, immature PNs derived from GM8330 organoids expressed genes from the DLX family, present in INs and not in the PN lineage[84] (Figure S16). Together, these results suggest an off-target ventralization of organoids derived from the GM8330 cell line.

To confirm this discovery, we performed a label transfer experiment training on fetal tissue derived from gestational weeks 14–25 (GW14–25) human embryos.[85] Most cell types, such as cycling cells and ventral precursors, get classified as expected. Focusing on neuronal cell types, the majority of callosal PNs get classified as excitatory PNs (80% PGP1, 60% GM8330, 74% 11A), and immature INs are properly classified as INs (93% PGP1, 86% GM8330, 86% 11A). However, immature PNs have clear differences between the cell lines. For the 11A line, 34% of immature PNs get classified as excitatory PNs and 38% as RGs. Similarly, in the PGP1 line, 57% of immature PNs are classified as excitatory PNs and 20% as RGs. On the other hand, only 7% of the GM8330 immature PNs are classified as excitatory PNs, and 21% are classified as RGs. Importantly 44% of these cells are predicted as INs (Figure S17), further suggesting a ventralization of the organoids derived from the GM8330 line.

We then performed a pseudotime analysis using Monocle 3.[86] In the 11A and PGP1 lines, we observe a clear differentiation trajectory from RG to the excitatory PN lineage (immature PNs and callosal PNs). In these lines, the IN lineage follows a separate path (Figures 7A and S18). Focusing on the GM8330 cell line, we observe that a large subset of immature PNs unexpectedly appear together with the IN lineage (Figure S18). Altogether, the data suggest that SIMS has correctly identified that a large subset of cells labeled as immature PNs in the GM8330 are, in fact, INs.

Leveraging *in vivo* data refines cell-type prediction in brain organoids

Visualization methods based on dimensionality reduction, such as principal component analysis and t-distributed stochastic neighbor embedding, often miss the global structure of the data and can lead to misclassification of cells.[87] Given that SIMS identified a ventralization of the GM8330 cell line (Figure 6), we then asked whether it can identify other cells previously misclassified in existing atlases.[76] We analyzed 6-month-old organoids derived from the 11A cell line. We first performed pseudotime analysis and found that a subset of cells labeled as immature PNs cluster in between other immature PNs and glial cells (Figure 7A). Interestingly, all these cells are identified by Gruffi as stressed cells (Figure 7B). To test whether these cells were mistakenly classified in previous atlases, we performed a label transfer from GW14–25 primary fetal tissue.[85] We find that SIMS assigns the entirety of this cell cluster as RGs and not PNs (Figure 7C). Gene-expression analysis of molecular markers of RGs, such as SOX2 and PAX6 (Figure S19), confirm that the SIMS label is correct. In addition, these cells lack expression of PN subtype markers such TBR1, SATB2, CUX1, and CUX2, as well as Pan-PN markers EMX1, DCX, NEUROD2, and NEUROD6 (Figure S19). Altogether, these results

suggest that the stressed cells previously labeled as immature PNs in the 11A cell line are indeed RGs.

We asked how correcting the cell-type annotation in 11A affected the label transfer between organoids derived from different cell lines. We trained SIMS in the newly annotated 11A dataset and made predictions in both the PGP1 and the GM8330 cells. We found that for the new model trained on the 11A cell line there is an accuracy of 75.7% and a Macro F1 score of 0.583 for PGP1 organoids and an accuracy of 76.3% and a Macro F1 score of 0.603 for GM8330 organoids (Tables S12 and S13), representing a significant improvement from label transfer experiments before the reclassification (Tables S11 and S14). Furthermore, we find that RGs now get predicted at an accuracy of 43.0% for PGP1 and 32.0% GM8330, as compared to the original predictions of 1.9% and 3.9% for the respective cell lines. Together, we show that proper identification of cell types through label transfer from primary tissue can help systematize multisample cell atlases.

DISCUSSION

Currently, over 1.5 million cells per month are sequenced and archived through the different cell atlas projects.[88] With the lowering trends in sequencing costs, the number of cells sequenced is increasing exponentially,[3,88] yet cell annotation remains a highly manual process, which is limiting the reproducibility and introducing biases in the data. Several open access solutions have emerged to streamline the process, albeit with different accuracies.[2] Deep-learning approaches that apply transformer-based architectures to gene-expression data have been shown to outperform other commonly used methods.[25] However, these approaches require a large number of cells for their unsupervised pretraining step and advanced computational knowledge and resources to further train their models.[25] SIMS, on the other hand, can be trained efficiently with a supervised training regime, therefore avoiding large data files and increasing its versatility. This allows the users to run SIMS in their local computers.

We designed SIMS as a low-code tool for both training and performing label transfer across single-cell datasets (Figure 1). SIMS can be used on user-specified datasets rather than reference datasets that are usually a prerequisite in popular tools. This is meant to remove barriers in adoption by new labs, medical practitioners, students, and non-experts alike. Unlike other deep-learning models,[25] SIMS can use genes that are defined by the user, allowing the label transfer in novel genomes, or use annotated genomes without standard nomenclature. Other deep-learning approaches, such as scBERT,[25] have been shown to work well with datasets of up to 16,000 genes. SIMS, being based on TabNet, and therefore optimized for tabular data,[30] can work well with over 45,000 features (Figure 2). This property would allow, in principle, SIMS to be trained simultaneously on references of multiple species and species with large genomes such as the axolotl,[89] as well as multimodal data including combined single-cell gene expression and gene accessibility sequencing datasets.[90]

When it comes to interpretability, SIMS is able to output a sparse selection of the most important genes, which can then be easily plotted in the Python ecosystem of Scanpy, while other tools[25] rely on external cross-platform packages. This can hamper the adoption of new users, including non-bioinformaticians.[91] Indeed, non-experts could greatly benefit from intuitive and low-effort tools that can streamline the analysis and integration of their newly generated data with existing knowledge.[91] To facilitate its adoption, we created a web application and a Terra pipeline that can be easily adopted with minimal coding knowledge and low infrastructural resources, offering accessible cloud computing. Furthermore, our approaches facilitate the sharing of trained models that can streamline collaboration between multiple groups.

We have shown that SIMS is applicable to a variety of species and tissues including blood, heart, kidney, lung, and the whole body (Figures 1, 2, and 4). We then focused on applying this tool to data generated from the brain. The brain is a complex tissue, where the great diversity of neurons is generated over a relatively short time period and identities are refined throughout life.[48,68] Several efforts, such as the BRAIN Initiative, the SSPsyGene consortium, and others, exist to sequence neurons across different ages, species, experimental models, and diseases.[92,93] While the neuroscience community has started efforts to agree on naming conventions across the increasing number of datasets,[5,94] there are still significant ontological inconsistencies in existing publications. We believe that SIMS could become an important tool to streamline these community-driven efforts. It is important to mention that while we focused our work in the brain, SIMS can easily be applied to single-cell RNA sequencing data of any other organ.

When performing label transfer in fully differentiated neuronal cell types, SIMS performed remarkably well, with accuracies above 97%, even with a low number of cells in the training set. Unlike many other tools, which define cells by the strong expression of marker genes,[7,95] the SIMS model takes advantage of lack of expression and fluctuations of expression levels of the whole transcriptome to learn and identify cell labels. Consistent with this, we observed that in developing tissue, where gene expression is fluctuating and identities are being refined, SIMS was able to classify most cell types and identify maturation differences in cell types undergoing fate refinement (Figure 5).[75] To date, the differentiation of layer-4 neurons through postmitotic refinement of upper-layer callosal PNs in late embryonic and early postnatal development has been hypothesized in several experiments.[60,74] Strikingly, SIMS is able to pinpoint neurons with a mixed identity between layer-4 neurons and upper-layer callosal PNs, which were missed by traditional clustering approaches. This identification opens the possibility of further understanding the molecular changes underpinning neuronal fate acquisition and plasticity.[64,65]

When applied to cortical organoids, SIMS identified previously misannotated cells in existing atlases.[76] These errors in annotation were caused by traditional clustering followed by differential gene-expression analysis and marker identification.[76] Notably, stressed cells were often misannotated, which is a common issue in the organoid development field.[82,83] Revisiting and reannotating existing atlases will greatly increase the accuracy of label transfer and improve the development of future protocols. Furthermore, annotating stem cell-derived atlases using primary fetal samples as reference can be used as a gold standard in the field and to discover cell types underrepresented in the

existing protocols.[76,93] Special attention should be paid to fate transitions under cell-stress conditions.[82] It has been postulated that cell stress can inhibit proper neuronal specification in brain organoids.[82] By analyzing 6-month-old cortical organoids from multiple cell lines, we showed that this phenomenon is dependent on the genetic background of the cell line. Importantly, we observed that cell stress impairs fate specification of the excitatory PNs but not inhibitory INs of the same organoids (Figures 6 and 7).

Applying SIMS to developing brain tissue including primary samples and organoids allowed us to identify subtle differences in developmental trajectories between cell types generated. We therefore believe that SIMS can be of great value in studying developmental disorders, such as autism, where existing models have already shown cell-type-dependent asynchronous developmental trajectories in different neuronal lineages.[96] Hybrid pipelines that integrate pseudotime-focused tools, such as Monocle or BOMA,[7,86] could become complementary to SIMS and have the potential to provide more comprehensive insights into these questions.

Limitations of the study

While we have shown that SIMS can accurately predict trans-sample labels and perform label transfer across different methodologies (single-cell and single-nuclei RNA sequencing) and models (primary tissue and cortical organoids), we have limited our work to samples within the same species. This is because neuronal subtypes diverge significantly between species[46] and at the individual level, gene orthologs can show different expression levels in different species.[97] However, some neuronal subtypes, such as MGE-derived INs, are transcriptomically more conserved across evolution than other primary neurons, including cortical PNs.[13,46] In the future, these IN subtypes could be used as a way to validate SIMS to perform trans-species predictions.[98] Additional modifications, such as gene module extraction, could provide increased accuracy for label transfer, as meta-modules could prove to be more conserved between evolutionarily distant species than gene orthologs.[94,99,100]

SIMS is a model trained in a supervised fashion, meaning that it relies on existing annotations to learn the mapping from gene-expression counts to cell type. Building these initial annotations requires data normalization, data clustering, differential expression, and expert annotations.[4,101] Technical factors such as sequencing technology or incorrect normalization may affect the downstream differential expression results, leading to misnamed clusters.

Although SIMS has a consistent runtime, it is not the fastest of the methods we benchmarked. For future work, we will implement model distillation via student-teacher methods, where a smaller, faster, and more efficient "student" model learns to mimic a larger "teacher" model.[102] This will reduce both the memory requirements and inference speed of the SIMS network. In future iterations we will consider implementing neuron-wise model pruning, in which individual weights are quantized to 0, and layer-wise model pruning whereby entire layers can be removed. Both methods can improve generalizability, while the latter will also reduce computational complexity. Finally, we would like to add metadata such as sequencing technology and cell lines to the model. This can be done by embedding categorical features into one-dimensional vectors and treating these as features that are integrated with the sequencing input, potentially via addition or concatenation. This would allow us to fingerprint which cells come from which sequencing technology and may allow the model to use this information for more robust prediction.

When applying SIMS to neuroscience, one of the main drawbacks is the lack of use of naming conventions in the field. This makes combining datasets coming from different labs to train the algorithm a manually intensive work. There are efforts under way to reach agreements in naming.[5] In future applications, we would like to explore ontology-based approaches for cell-naming harmonization, allowing the user to input datasets with different naming conventions while the pipeline reannotates the cells to a common nomenclature.

In conclusion, we propose SIMS as a novel, accurate, and easy-to-use tool to facilitate label transfer in single-cell data with a direct application in the neuroscience community.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - The SIMS pipeline
  - Model architecture
  - Interpretability
  - Code library details
  - Web application
  - Training details
  - Datasets details
  - Benchmarking against cell type classification models
  - Pseudotime analysis
  - Cell stress analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2024.100581.

ACKNOWLEDGMENTS

We would like to thank Tomasz Nowakowski, Maximilian Haeussler, and Hunter Schweiger for their valuable feedback on this article. This work was supported by Schmidt Futures (SF857) to M.T. and D.H.; the National Human Genome Research Institute (1RM1HG011543) to M.T. and D.H.; the National Science Foundation (NSF2134955 to M.T. and D.H., NSF2034037 to M.T.); and the National Institute of Mental Health (1U24MH132628) to B.P., D.H., and M.A.M.-R. We are thankful to the Pacific Research Platform, supported by the National Science Foundation under award numbers CNS-1730158, ACI-1540112, ACI-1541349, and OAC-1826967, the University of California Office of the President, and the University of California San Diego's California Institute for Telecommunications and Information Technology/QualcommInstitute.

AUTHOR CONTRIBUTIONS

B.P., M.T., D.H., V.D.J., and M.A.M.-R. conceived the project. J.G.-F. and J.L. performed the experiments. A.O. provided support working with the Terra

system. J.G.-F., J.L., and M.A.M-R. wrote the paper with contributions from all authors.

DECLARATION OF INTERESTS

J.L., V.D.J., and M.A.M.-R. have submitted patent applications related to the work in this paper.

Received: November 17, 2023
Revised: April 2, 2024
Accepted: May 9, 2024
Published: May 31, 2024

REFERENCES

1. Haque, A., Engel, J., Teichmann, S.A., and Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. *9*, 75. https://doi.org/10.1186/s13073-017-0467-4.
2. Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. Nat. Biotechnol. *37*, 547–554. https://doi.org/10.1038/s41587-019-0071-9.
3. Angerer, P., Simon, L., Tritschler, S., Wolf, F.A., Fischer, D., and Theis, F.J. (2017). Single cells make big data: New challenges and opportunities in transcriptomics. Current Opinion in Systems Biology *4*, 85–91. https://doi.org/10.1016/j.coisb.2017.07.004.
4. Luecken, M.D., and Theis, F.J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. Mol. Syst. Biol. *15*, e8746. https://doi.org/10.15252/msb.20188746.
5. Yuste, R., Hawrylycz, M., Aalling, N., Aguilar-Valles, A., Arendt, D., Armañanzas, R., Ascoli, G.A., Bielza, C., Bokharaie, V., Bergmann, T.B., et al. (2020). A community-based transcriptomics classification and nomenclature of neocortical cell types. Nat. Neurosci. *23*, 1456–1468. https://doi.org/10.1038/s41593-020-0685-8.
6. Grabski, I.N., and Irizarry, R.A. (2022). A probabilistic gene expression barcode for annotation of cell types from single-cell RNA-seq data. Biostatistics *23*, 1150–1164. https://doi.org/10.1093/biostatistics/kxac021.
7. He, C., Kalafut, N.C., Sandoval, S.O., Risgaard, R., Sirois, C.L., Yang, C., Khullar, S., Suzuki, M., Huang, X., Chang, Q., et al. (2023). BOMA, a machine-learning framework for comparative gene expression analysis across brains and organoids. Cell Rep. Methods *3*, 100409. https://doi.org/10.1016/j.crmeth.2023.100409.
8. Guo, H., and Li, J. (2021). scSorter: assigning cells to known cell types according to marker genes. Genome Biol. *22*, 69. https://doi.org/10.1186/s13059-021-02281-7.
9. Wang, Q., Zhou, Q., Zhang, S., Shao, W., Yin, Y., Li, Y., Hou, J., Zhang, X., Guo, Y., Wang, X., et al. (2016). Elevated hapln2 expression contributes to protein aggregation and neurodegeneration in an animal model of Parkinson's disease. Front. Aging Neurosci. *8*, 197. https://doi.org/10.3389/fnagi.2016.00197.
10. Wonders, C.P., and Anderson, S.A. (2006). The origin and specification of cortical interneurons. Nat. Rev. Neurosci. *7*, 687–696. https://doi.org/10.1038/nrn1954.
11. de Lecea, L., del Río, J.A., and Soriano, E. (1995). Developmental expression of parvalbumin mRNA in the cerebral cortex and hippocampus of the rat. Brain Res. Mol. Brain Res. *32*, 1–13. https://doi.org/10.1016/0169-328x(95)00056-x.
12. Lee, B.R., Dalley, R., Miller, J.A., Chartrand, T., Close, J., Mann, R., Mukora, A., Ng, L., Alfiler, L., Baker, K., et al. (2023). Signature morphoelectric properties of diverse GABAergic interneurons in the human neocortex. Science *382*, eadf6484. https://doi.org/10.1126/science.adf6484.
13. Mostajo-Radji, M.A., Schmitz, M.T., Montoya, S.T., and Pollen, A.A. (2020). Reverse engineering human brain evolution using organoid models. Brain Res. *1729*, 146582. https://doi.org/10.1016/j.brainres.2019.146582.
14. Zeng, H., Shen, E.H., Hohmann, J.G., Oh, S.W., Bernard, A., Royall, J.J., Glattfelder, K.J., Sunkin, S.M., Morris, J.A., Guillozet-Bongaarts, A.L., et al. (2012). Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures. Cell *149*, 483–496. https://doi.org/10.1016/j.cell.2012.02.052.
15. Pasquini, G., Rojo Arias, J.E., Schäfer, P., and Busskamp, V. (2021). Automated methods for cell type annotation on scRNA-seq data. Comput. Struct. Biotechnol. J. *19*, 961–969. https://doi.org/10.1016/j.csbj.2021.01.015.
16. Zhang, Y., Aevermann, B., Gala, R., and Scheuermann, R.H. (2022). Cell type matching in single-cell RNA-sequencing data using FR-match. Sci. Rep. *12*, 9996. https://doi.org/10.1038/s41598-022-14192-z.
17. Pliner, H.A., Shendure, J., and Trapnell, C. (2019). Supervised classification enables rapid annotation of cell atlases. Nat. Methods *16*, 983–986. https://doi.org/10.1038/s41592-019-0535-3.
18. Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M.J.T., and Mahfouz, A. (2019). A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biol. *20*, 194. https://doi.org/10.1186/s13059-019-1795-z.
19. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. Cell *177*, 1888–1902.e21. https://doi.org/10.1016/j.cell.2019.05.031.
20. Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J., Abate, A.R., et al. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat. Immunol. *20*, 163–172. https://doi.org/10.1038/s41590-018-0276-y.
21. Kuo, F.Y., and Sloan, I.H. (2005). Lifting the curse of dimensionality. Notices of the AMS *52*, 1320–1328.
22. Akbani, R., Kwek, S., and Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets. In Machine Learning: ECML 2004: 15th European Conference on Machine Learning, Pisa, Italy. Proceedings 15, J.P. Boulicaut, F. Esposito, F. Gionnotti, and D. Pedreschi, eds. (Springer), pp. 39–50. https://doi.org/10.1007/978-3-540-30115-8_7.
23. Wang, T., Johnson, T.S., Shao, W., Lu, Z., Helm, B.R., Zhang, J., and Huang, K. (2019). BERMUDA: a novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes. Genome Biol. *20*, 165. https://doi.org/10.1186/s13059-019-1764-6.
24. Lotfollahi, M., Naghipourfar, M., Luecken, M.D., Khajavi, M., Büttner, M., Wagenstetter, M., Avsec, Ž., Gayoso, A., Yosef, N., Interlandi, M., et al. (2022). Mapping single-cell data to reference atlases by transfer learning. Nat. Biotechnol. *40*, 121–130. https://doi.org/10.1038/s41587-021-01001-7.
25. Yang, F., Wang, W., Wang, F., Fang, Y., Tang, D., Huang, J., Lu, H., and Yao, J. (2022). scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. Nat. Mach. Intell. *4*, 852–866. https://doi.org/10.1038/s42256-022-00534-z.
26. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods *15*, 1053–1058. https://doi.org/10.1038/s41592-018-0229-2.
27. Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M.I., and Yosef, N. (2021). Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. Mol. Syst. Biol. *17*, e9620. https://doi.org/10.15252/msb.20209620.
28. Kimmel, J.C., and Kelley, D.R. (2021). Semisupervised adversarial neural networks for single-cell classification. Genome Res. *31*, 1781–1793. https://doi.org/10.1101/gr.268581.120.

29. Cheng, C., Chen, W., Jin, H., and Chen, X. (2023). A review of single-cell RNA-seq annotation, integration, and cell–cell communication. Cells *12*, 1970. https://doi.org/10.3390/cells12151970.

30. Arik, S.Ö., and Pfister, T. (2021). TabNet: Attentive interpretable tabular learning. Proc. AAAI Conf. Artif. Intell. *35*, 6679–6687. https://doi.org/10.48550/arXiv.1908.07442.

31. Falcon, W.; The PyTorch Lightning Team (2020). PyTorch Lightning: The lightweight PyTorch wrapper for high-performance AI research. Scale your models, not the boilerplate. https://doi.org/10.5281/zenodo.3828935. https://zenodo.org/records/7545285.

32. Guo, C., Pleiss, G., Sun, Y., and Weinberger, K.Q. (2017). On calibration of modern neural networks. International Conference on Machine Learning *2017*, 1321–1330. https://doi.org/10.48550/arXiv.1706.04599.

33. Shazeer, N. (2020). Glu variants improve transformer. arXiv. https://doi.org/10.48550/arXiv.2002.05202.

34. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049. https://doi.org/10.1038/ncomms14049.

35. Kaminow, B., Yunusov, D., and Dobin, A. (2021). STARsolo: Accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. bioRxiv. https://doi.org/10.1101/2021.05.05.442755.

36. Cumulus Team (2023). Cumulus Cellranger workflow version 2.4.1. Dockstore. Oct 19, 2023. https://dockstore.org/workflows/github.com/lilab-bcb/cumulus/Cellranger:2.4.1?tab=info. https://dockstore.org/workflows/github.com/lilab-bcb/cumulus/Cellranger:2.4.1?tab=info.

37. Leinonen, R., Sugawara, H., and Shumway, M.; International Nucleotide Sequence Database Collaboration (2011). The Sequence Read Archive. Nucleic Acids Res. *39*, D19–D21. https://doi.org/10.1093/nar/gkq1019.

38. Farrell, A.O. (2023). Sranwrp: Pull Fastqs from Sra by Run. https://doi.org/10.5281/zenodo.11237529. https://dockstore.org/workflows/github.com/aofarrel/SRANWRP/pull_FASTQs_from_SRA_by_run:v1.1.17?tab=files.

39. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049. https://doi.org/10.1038/ncomms14049.

40. Tucker, N.R., Chaffin, M., Fleming, S.J., Hall, A.W., Parsons, V.A., Bedi, K.C., Jr., Akkad, A.-D., Herndon, C.N., Arduini, A., Papangeli, I., et al. (2020). Transcriptional and cellular diversity of the human heart. Circulation *142*, 466–482. https://doi.org/10.1161/CIRCULATIONAHA.119.045401.

41. Travaglini, K.J., Nabhan, A.N., Penland, L., Sinha, R., Gillich, A., Sit, R.V., Chang, S., Conley, S.D., Mori, Y., Seita, J., et al. (2020). A molecular cell atlas of the human lung from single-cell RNA sequencing. Nature *587*, 619–625. https://doi.org/10.1038/s41586-020-2922-4.

42. Li, C., Liu, B., Kang, B., Liu, Z., Liu, Y., Chen, C., Ren, X., and Zhang, Z. (2020). SciBet as a portable and fast single cell type identifier. Nat. Commun. *11*, 1818. https://doi.org/10.1038/s41467-020-15523-2.

43. Maan, H., Zhang, L., Yu, C., Geuenich, M.J., Campbell, K.R., and Wang, B. (2024). Characterizing the impacts of dataset imbalance on single-cell data integration. Nat. Biotechnol. https://doi.org/10.1038/s41587-023-02097-9.

44. Smarr L., Crittenden C., DeFanti T., Graham J., Mishin D., Moore R., Papadopoulos P., Würthwein F. (2018). The Pacific Research Platform: Making high-speed networking a reality for the scientist. In Proceedings of the Practice and Experience on Advanced Research Computing. S. Sanieleveci, ed. (Assocation for Computing Machinery). pp. 1–8. https://doi.org/10.1145/3219104.3219108.

45. Tasic, B., Yao, Z., Graybuck, L.T., Smith, K.A., Nguyen, T.N., Bertagnolli, D., Goldy, J., Garren, E., Economo, M.N., Viswanathan, S., et al. (2018). Shared and distinct transcriptomic cell types across neocortical areas. Nature *563*, 72–78. https://doi.org/10.1038/s41586-018-0654-5.

46. Hodge, R.D., Bakken, T.E., Miller, J.A., Smith, K.A., Barkan, E.R., Graybuck, L.T., Close, J.L., Long, B., Johansen, N., Penn, O., et al. (2019). Conserved cell types with divergent features in human versus mouse cortex. Nature *573*, 61–68. https://doi.org/10.1038/s41586-019-1506-7.

47. Yao, Z., van Velthoven, C.T.J., Nguyen, T.N., Goldy, J., Sedeno-Cortes, A.E., Baftizadeh, F., Bertagnolli, D., Casper, T., Chiang, M., Crichton, K., et al. (2021). A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. Cell *184*, 3222–3241.e26. https://doi.org/10.1016/j.cell.2021.04.021.

48. Cadwell, C.R., Bhaduri, A., Mostajo-Radji, M.A., Keefe, M.G., and Nowakowski, T.J. (2019). Development and arealization of the cerebral cortex. Neuron *103*, 980–1004. https://doi.org/10.1016/j.neuron.2019.07.009.

49. Anand, K.S., and Dhikav, V. (2012). Hippocampus in health and disease: An overview. Ann. Indian Acad. Neurol. *15*, 239–246. https://doi.org/10.4103/0972-2327.104323.

50. Xiong, H., Kovacs, I., and Zhang, Z. (2004). Differential distribution of KChIPs mRNAs in adult mouse brain. Brain Res. Mol. Brain Res. *128*, 103–111. https://doi.org/10.1016/j.molbrainres.2004.06.024.

51. Xiong, H., Xia, K., Li, B., Zhao, G., and Zhang, Z. (2009). KChIP1: A potential modulator to GABAergic system. Acta Biochim. Biophys. Sin. *41*, 295–300. https://doi.org/10.1093/abbs/gmp013.

52. Fukumoto, K., Tamada, K., Toya, T., Nishino, T., Yanagawa, Y., and Takumi, T. (2018). Identification of genes regulating GABAergic interneuron maturation. Neurosci. Res. *134*, 18–29. https://doi.org/10.1016/j.neures.2017.11.010.

53. Miyoshi, G., Young, A., Petros, T., Karayannis, T., McKenzie Chang, M., Lavado, A., Iwano, T., Nakajima, M., Taniguchi, H., Huang, Z.J., et al. (2015). Prox1 regulates the subtype-specific development of caudal ganglionic eminence-derived GABAergic cortical interneurons. J. Neurosci. *35*, 12869–12889. https://doi.org/10.1523/JNEUROSCI.1164-15.2015.

54. Herring, C.A., Simmons, R.K., Freytag, S., Poppe, D., Moffet, J.J.D., Pflueger, J., Buckberry, S., Vargas-Landin, D.B., Clément, O., Echeverría, E.G., et al. (2022). Human prefrontal cortex gene regulatory dynamics from gestation to adulthood at single-cell resolution. Cell *185*, 4428–4447.e28. https://doi.org/10.1016/j.cell.2022.09.039.

55. Kawaguchi, Y., and Kondo, S. (2002). Parvalbumin, somatostatin and cholecystokinin as chemical markers for specific GABAergic interneuron types in the rat frontal cortex. J. Neurocytol. *31*, 277–287. https://doi.org/10.1023/a:1024126110356.

56. Joseph, D.J., Von Deimling, M., Hasegawa, Y., Cristancho, A.G., Risbud, R., McCoy, A.J., and Marsh, E.D. (2021). Protocol for isolating young adult parvalbumin interneurons from the mouse brain for extraction of high-quality RNA. STAR Protoc. *2*, 100714. https://doi.org/10.1016/j.xpro.2021.100714.

57. Larson, A., and Chin, M.T. (2021). A method for cryopreservation and single nucleus RNA-sequencing of normal adult human interventricular septum heart tissue reveals cellular diversity and function. BMC Med. Genomics *14*, 161. https://doi.org/10.1186/s12920-021-01011-z.

58. Thrupp, N., Sala Frigerio, C., Wolfs, L., Skene, N.G., Fattorelli, N., Poovathingal, S., Fourne, Y., Matthews, P.M., Theys, T., Mancuso, R., et al. (2020). Single-nucleus RNA-seq is not suitable for detection of microglial activation genes in humans. Cell Rep. *32*, 108189. https://doi.org/10.1016/j.celrep.2020.108189.

59. Caglayan, E., Liu, Y., and Konopka, G. (2022). Neuronal ambient RNA contamination causes misinterpreted and masked cell types in brain single-nuclei datasets. Neuron *110*, P4043–P4056.e5. https://doi.org/10.1016/j.neuron.2022.09.010.

60. De León Reyes, N.S., Mederos, S., Varela, I., Weiss, L.A., Perea, G., Galazo, M.J., and Nieto, M. (2019). Transient callosal projections of L4 neurons are eliminated for the acquisition of local connectivity. Nat. Commun. *10*, 4549. https://doi.org/10.1038/s41467-019-12495-w.

61. Kiselev, V.Y., Andrews, T.S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. Nat. Rev. Genet. *20*, 273–282. https://doi.org/10.1038/s41576-018-0088-9.

62. Di Bella, D.J., Habibi, E., Stickels, R.R., Scalia, G., Brown, J., Yadollahpour, P., Yang, S.M., Abbate, C., Biancalani, T., Macosko, E.Z., et al. (2021). Molecular logic of cellular diversification in the mouse cerebral cortex. Nature *595*, 554–559. https://doi.org/10.1038/s41586-021-03670-5.

63. Nowakowski, T.J., Bhaduri, A., Pollen, A.A., Alvarado, B., Mostajo-Radji, M.A., Di Lullo, E., Haeussler, M., Sandoval-Espinosa, C., Liu, S.J., Velmeshev, D., et al. (2017). Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. Science *358*, 1318–1323. https://doi.org/10.1126/science.aap8809.

64. Ozair, M.Z., Kirst, C., van den Berg, B.L., Ruzo, A., Rito, T., and Brivanlou, A.H. (2018). hPSC modeling reveals that fate selection of cortical deep projection neurons occurs in the subplate. Cell Stem Cell *23*, 60–73.e6. https://doi.org/10.1016/j.stem.2018.05.024.

65. Mostajo-Radji, M.A., and Pollen, A.A. (2018). Postmitotic fate refinement in the subplate. Cell Stem Cell *23*, 7–9. https://doi.org/10.1016/j.stem.2018.06.017.

66. Greig, L.C., Woodworth, M.B., Galazo, M.J., Padmanabhan, H., and Macklis, J.D. (2013). Molecular logic of neocortical projection neuron specification, development and diversity. Nat. Rev. Neurosci. *14*, 755–769. https://doi.org/10.1038/nrn3586.

67. Ciemerych, M.A., and Sicinski, P. (2005). Cell cycle in mouse development. Oncogene *24*, 2877–2898. https://doi.org/10.1038/sj.onc.1208608.

68. Lodato, S., and Arlotta, P. (2015). Generating neuronal diversity in the mammalian cerebral cortex. Annu. Rev. Cell Dev. Biol. *31*, 699–720. https://doi.org/10.1146/annurev-cellbio-100814-125353.

69. Rouaux, C., and Arlotta, P. (2013). Direct lineage reprogramming of postmitotic callosal neurons into corticofugal neurons in vivo. Nat. Cell Biol. *15*, 214–221. https://doi.org/10.1038/ncb2660.

70. Ye, Z., Mostajo-Radji, M.A., Brown, J.R., Rouaux, C., Tomassy, G.S., Hensch, T.K., and Arlotta, P. (2015). Instructing perisomatic inhibition by direct lineage reprogramming of neocortical projection neurons. Neuron *88*, 475–483. https://doi.org/10.1016/j.neuron.2015.10.006.

71. De la Rossa, A., Bellone, C., Golding, B., Vitali, I., Moss, J., Toni, N., Lüscher, C., and Jabaudon, D. (2013). In vivo reprogramming of circuit connectivity in postmitotic neocortical neurons. Nat. Neurosci. *16*, 193–200. https://doi.org/10.1038/nn.3299.

72. Ge, W.-P., Miyawaki, A., Gage, F.H., Jan, Y.N., and Jan, L.Y. (2012). Local generation of glia is a major astrocyte source in postnatal cortex. Nature *484*, 376–380. https://doi.org/10.1038/nature10959.

73. Leone, D.P., Srinivasan, K., Chen, B., Alcamo, E., and McConnell, S.K. (2008). The determination of projection neuron identity in the developing cerebral cortex. Curr. Opin. Neurobiol. *18*, 28–35. https://doi.org/10.1016/j.conb.2008.05.006.

74. Oishi, K., Nakagawa, N., Tachikawa, K., Sasaki, S., Aramaki, M., Hirano, S., Yamamoto, N., Yoshimura, Y., and Nakajima, K. (2016). Identity of neocortical layer 4 neurons is specified through correct positioning into the cortex. Elife *5*, e10907. https://doi.org/10.7554/eLife.10907.

75. Clark, E.A., Rutlin, M., Capano, L.S., Aviles, S., Saadon, J.R., Taneja, P., Zhang, Q., Bullis, J.B., Lauer, T., Myers, E., et al. (2020). Cortical rorβ is required for layer 4 transcriptional identity and barrel integrity. Elife *9*, e52370. https://doi.org/10.7554/eLife.52370.

76. Velasco, S., Kedaigle, A.J., Simmons, S.K., Nash, A., Rocha, M., Quadrato, G., Paulsen, B., Nguyen, L., Adiconis, X., Regev, A., et al. (2019). Individual brain organoids reproducibly form cell diversity of the human cerebral cortex. Nature *570*, 523–527. https://doi.org/10.1038/s41586-019-1289-x.

77. Velasco, S., Paulsen, B., and Arlotta, P. (2020). 3D brain organoids: studying brain development and disease outside the embryo. Annu. Rev. Neurosci. *43*, 375–389. https://doi.org/10.1146/annurev-neuro-070918-050154.

78. Hernández, D., Rooney, L.A., Daniszewski, M., Gulluyan, L., Liang, H.H., Cook, A.L., Hewitt, A.W., and Pébay, A. (2021). Culture variabilities of human IPSC-derived cerebral organoids are a major issue for the modelling of phenotypes observed in Alzheimer's disease. Stem Cell Reviews and Reports *18*, 718–731. https://doi.org/10.1007/s12015-021-10147-5.

79. Pollen, A.A., Bhaduri, A., Andrews, M.G., Nowakowski, T.J., Meyerson, O.S., Mostajo-Radji, M.A., Di Lullo, E., Alvarado, B., Bedolli, M., Dougherty, M.L., et al. (2019). Establishing cerebral organoids as models of human-specific brain evolution. Cell *176*, 743–756.e17. https://doi.org/10.1016/j.cell.2019.01.017.

80. Uzquiano, A., Kedaigle, A.J., Pigoni, M., Paulsen, B., Adiconis, X., Kim, K., Faits, T., Nagaraja, S., Antón-Bolaños, N., Gerhardinger, C., et al. (2022). Proper acquisition of cell class identity in organoids allows definition of fate specification programs of the human cerebral cortex. Cell *185*, 3770–3788.e27. https://doi.org/10.1016/j.cell.2022.09.010.

81. Seiler, S.T., Mantalas, G.L., Selberg, J., Cordero, S., Torres-Montoya, S., Baudin, P.V., Ly, V.T., Amend, F., Tran, L., Hoffman, R.N., et al. (2022). Modular automated microfluidic cell culture platform reduces glycolytic stress in cerebral cortex organoids. Sci. Rep. *12*, 20173. https://doi.org/10.1038/s41598-022-20096-9.

82. Bhaduri, A., Andrews, M.G., Mancia Leon, W., Jung, D., Shin, D., Allen, D., Jung, D., Schmunk, G., Haeussler, M., Salma, J., et al. (2020). Cell stress in cortical organoids impairs molecular subtype specification. Nature *578*, 142–148. https://doi.org/10.1038/s41586-020-1962-0.

83. Vértesy, Á., Eichmüller, O.L., Naas, J., Novatchkova, M., Esk, C., Balmaña, M., Ladstaetter, S., Bock, C., von Haeseler, A., and Knoblich, J.A. (2022). Gruffi: an algorithm for computational removal of stressed cells from brain organoid transcriptomic datasets. EMBO J. *41*, e111118. https://doi.org/10.15252/embj.2022111118.

84. Anderson, S.A., Eisenstat, D.D., Shi, L., and Rubenstein, J.L. (1997). Interneuron Migration from Basal Forebrain to Neocortex: Dependence on Dlx Genes. Science *278*, 474–476. https://doi.org/10.1126/science.278.5337.474.

85. Bhaduri, A., Sandoval-Espinosa, C., Otero-Garcia, M., Oh, I., Yin, R., Eze, U.C., Nowakowski, T.J., and Kriegstein, A.R. (2021). An atlas of cortical arealization identifies dynamic molecular signatures. Nature *598*, 200–204. https://doi.org/10.1126/science.278.5337.474.

86. Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature *566*, 496–502. https://doi.org/10.1038/s41586-019-0969-x.

87. Wang, H.Y., Zhao, J.P., Su, Y.S., and Zheng, C.H. (2022). scCDG: a method based on DAE and GCN for scRNA-seq data analysis. IEEE/ACM Trans. Comput. Biol. Bioinform. *19*, 3685–3694. https://doi.org/10.1109/TCBB.2021.3126641.

88. Svensson, V., da Veiga Beltrame, E., and Pachter, L. (2020). A curated database reveals trends in single-cell transcriptomics. Database *2020*, baaa073. https://doi.org/10.1093/database/baaa073.

89. Nowoshilow, S., Schloissnig, S., Fei, J.F., Dahl, A., Pang, A.W.C., Pippel, M., Winkler, S., Hastie, A.R., Young, G., Roscito, J.G., et al. (2018). The axolotl genome and the evolution of key tissue formation regulators. Nature *554*, 50–55. https://doi.org/10.1038/s41586-018-0141-z.

90. Jiang, F., Zhou, X., Qian, Y., Zhu, M., Wang, L., Li, Z., Shen, Q., Wang, M., Qu, F., Cui, G., et al. (2023). Simultaneous profiling of spatial gene expression and chromatin accessibility during mouse brain development. Nat. Methods *20*, 1048–1057. https://doi.org/10.1038/s41592-023-01884-1.

91. Krampis, K. (2022). Democratizing bioinformatics through easily accessible software platforms for non-experts in the field. Biotechniques *72*, 36–38. https://doi.org/10.2144/btn-2021-0060.

92. Maitra, M., Nagy, C., and Turecki, G. (2019). Sequencing the human brain at single-cell resolution. Curr. Behav. Neurosci. Rep. *6*, 197–208. https://doi.org/10.1007/s40473-019-00192-3.

93. He, Z., Dony, L., Fleck, J.S., Szałata, A., Li, K.X., Slišković, I., Lin, H.C., Santel, M., Atamian, A., Quadrato, G., et al. (2023). An integrated transcriptomic cell atlas of human neural organoids. bioRxiv. https://doi.org/10.1101/2023.10.05.561097.

94. Song, Y., Miao, Z., Brazma, A., and Papatheodorou, I. (2023). Benchmarking strategies for cross-species integration of single-cell RNA sequencing data. Nat. Commun. *14*, 6495. https://doi.org/10.1038/s41467-023-41855-w.

95. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573–3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.

96. Paulsen, B., Velasco, S., Kedaigle, A.J., Pigoni, M., Quadrato, G., Deo, A.J., Adiconis, X., Uzquiano, A., Sartore, R., Yang, S.M., et al. (2022). Autism genes converge on asynchronous development of shared neuron classes. Nature *602*, 268–273. https://doi.org/10.1038/s41586-021-04358-6.

97. Liao, B.Y., and Zhang, J. (2006). Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol. Biol. Evol. *23*, 530–540. https://doi.org/10.1093/molbev/msj054.

98. Liu, X., Shen, Q., and Zhang, S. (2023). Cross-species cell-type assignment from single-cell RNA-seq data by a heterogeneous graph neural network. Genome Res. *33*, 96–111. https://doi.org/10.1101/gr.276868.122.

99. Nano, P.R., Fazzari, E., Azizad, D., Nguyen, C.V., Wang, S., Kan, R.L., Wick, B., Haeussler, M., and Bhaduri, A. (2023). A meta-atlas of the developing human cortex identifies modules driving cell subtype specification. bioRxiv. https://doi.org/10.1101/2023.09.12.557406.

100. Suresh, H., Crow, M., Jorstad, N., Hodge, R., Lein, E., Dobin, A., Bakken, T., and Gillis, J. (2023). Comparative single-cell transcriptomic analysis of primate brains highlights human-specific regulatory evolution. Nat. Ecol. Evol. *7*, 1930–1943. https://doi.org/10.1038/s41559-023-02186-7.

101. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). Scanpy: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15. https://doi.org/10.1186/s13059-017-1382-0.

102. Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. arXiv arXiv:1503.02531. https://doi.org/10.48550/arXiv.1503.02531.

103. Han, X., Zhou, Z., Fei, L., Sun, H., Wang, R., Chen, Y., Chen, H., Wang, J., Tang, H., Ge, W., et al. (2020). Construction of a human cell landscape at single-cell level. Nature *581*, 303–309. https://doi.org/10.1038/s41586-020-2157-4.

104. Stewart, B.J., Ferdinand, J.R., Young, M.D., Mitchell, T.J., Loudon, K.W., Riding, A.M., Richoz, N., Frazer, G.L., Staniforth, J.U.L., Vieira Braga, F.A., et al. (2019). Spatiotemporal immune zonation of the human kidney. Science *365*, 1461–1466. https://doi.org/10.1126/science.aat5031.

105. Santurkar, S., Tsipras, D., Ilyas, A., and Madry, A. (2018). How does batch normalization help optimization? Adv. Neural Inf. Process. Syst. *31*. https://doi.org/10.48550/arXiv.1805.11604.

106. Martins, A., and Astudillo, R. (2016). From softmax to sparsemax: A sparse model of attention and multi-label classification. In International conference on machine learning (PMLR), pp. 1614–1623. https://doi.org/10.48550/arXiv.1602.02068.

107. Alquicira-Hernandez, J., Sathe, A., Ji, H.P., Nguyen, Q., and Powell, J.E. (2019). scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. Genome Biol. *20*, 264. https://doi.org/10.1186/s13059-019-1862-5.

108. Kingma, D.P., and Ba, J. (2014). Adam: A method for stochastic optimization. arXiv. https://doi.org/10.48550/arXiv.1412.6980.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
PBMC 68K dataset	Zheng et al.,[34]	https://www.10xgenomics.com/resources/datasets/fresh-68-k-pbm-cs-donor-a-1-standard-1-1-0
Human Landscape dataset	Han et al.,[103]	https://cells.ucsc.edu/?ds=human-cellular-landscape
Human Heart dataset	Tucker et al.,[40]	https://singlecell.broadinstitute.org/single_cell/study/SCP498/transcriptional-and-cellular-diversity-of-the-human-heart#study-summary
Human Kidney dataset	Stewart et al.,[104]	https://cells.ucsc.edu/?bp=kidney&ds=kidney-atlas
Human Lung dataset	Travaglini et al.,[41]	https://cells.ucsc.edu/?bp=lung&ds=stanford-czb-hlca
Adult mouse cortical and hippocampal dataset	Tasic et al.,[45]	https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-whole-cortex-and-hippocampus-10x
Adult human cortical dataset	Hodge et al.,[46]	https://portal.brain-map.org/atlases-and-data/rnaseq/human-multiple-cortical-areas-smart-seq
Developing mouse cortical dataset	Di Bella et al.,[62]	https://singlecell.broadinstitute.org/single_cell/study/SCP1290/molecular-logic-of-cellular-diversification-in-the-mammalian-cerebral-cortex%20%20Human%20cortical%20organoids%20dataset
Human cortical organoids dataset	Velasco et al.,[76]	https://singlecell.broadinstitute.org/single_cell/study/SCP282/reproducible-brain-organoids#study-summary
Human fetal brain development	Bhaduri et al.,[85]	https://cells.ucsc.edu/?bp=brain&ds=dev-brain-regions
Software and algorithms		
SIMS 1.0.0	This manuscript	https://github.com/braingeneers/SIMS and https://doi.org/10.5281/zenodo.11095105
SingleR 1.6.1	Aran et al.,[20]	SingleR (RRID:SCR_023120)
Scanvi 1.0.2	Xu et al.,[27]	https://github.com/scverse/scvi-tools
Seurat 4.0.3	Stuart et al.,[19]	Seurat (RRID:SCR_016341)
Scnym 0.3.2	Kimmel and Kelley.,[28]	https://github.com/calico/scnym
scBERT 1.0	Yang et al.,[25]	https://github.com/TencentAILabHealthcare/scBERT
Scibet 1.0	Li et al.,[42]	Scibet (RRID:SCR_024743)
Monocle 3.1	Cao et al.,[86]	Monocle 3 (RRID: SCR_018685)
Gruffi 1.0	Vértesy et al.,[83]	https://github.com/jn-goe/gruffi
Scanpy 1.9.3	Wolf et al.,[101]	scanpy (RRID:SCR_018139)

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and code should be directed to and will be fulfilled by the lead contact, Mohammed A. Mostajo-Radji (mmostajo@ucsc.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Peripheral blood mononuclear cells: https://www.10xgenomics.com/resources/datasets/fresh-68-k-pbm-cs-donor-a-1-standard-1-1-0.

Human landscape: https://cells.ucsc.edu/?ds=human-cellular-landscape.

Tucker's heart dataset: https://singlecell.broadinstitute.org/single_cell/study/SCP498/transcriptional-and-cellular-diversity-of-the-human-heart.

Stewart's kidney dataset: https://cells.ucsc.edu/?bp=kidney&ds=kidney-atlas.
Krasnow's lung dataset: https://cells.ucsc.edu/?bp=lung&ds=stanford-czb-hlca.
Human adult cerebral cortex: https://portal.brain-map.org/atlases-and-data/rnaseq/human-multiple-cortical-areas-smart-seq.
Mouse adult cerebral cortex and hippocampus: https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-whole-cortex-and-hippocampus-10x.
Developing mouse cerebral cortex (E12-P1): https://singlecell.broadinstitute.org/single_cell/study/SCP1290/molecular-logic-of-cellular-diversification-in-the-mammalian-cerebral-cortex.
Human cortical organoids:https://singlecell.broadinstitute.org/single_cell/study/SCP282/reproducible-brain-organoids#study-summary.
Human fetal brain development:https://cells.ucsc.edu/?bp=brain&ds=dev-brain-regions.
All code for SIMS has been deposited in https://github.com/braingeneers/SIMS and in the Zenodo archive https://doi.org/10.5281/zenodo.11095105.

METHOD DETAILS

The SIMS pipeline

The classifier component of the SIMS framework is TabNet,[30] a transformer-based neural network with sparse feature masks that allow for direct prediction interpretability from the input features. For each forward pass, batch-normalization is applied. The encoder is several steps (parameterized by the user) of self-attention layers and learned sparse feature masks, we offer some preset configurations that depend on the size and complexity of the reference dataset. The decoder then takes these encoded features and passes them through a fully-connected layer with batch-normalization and a generalized linear unit activation.[33] Interpretability by sample is then measured as the sum of feature mask weights across all encoding layers. For our visualization, we average all feature masks across all cells to understand the average contribution of each gene to the classification. You could also average the feature masks by cell type.

Model architecture

The encoder architecture consists of three components: a feature transformer, an attentive transformer, and a feature mask. The raw features are used as inputs, and while no global normalization is applied internally, batch normalization is used during training to improve convergence and stability.[105] This has been shown to be important for keeping the model training stable.[30,105] This is separate from single-cell batch normalization, a technique that can refer to removing technical variation from sequencing technology while retaining biological signal. The same p dimensional inputs are passed to each decision step of the encoder which has N_{steps} decision steps. For feature selection at the ith step, an element-wise multiplicative learnable mask Mi is used. This mask is learned via the attentive transformer, and sparsemax normalization[106] is used to induce sparsity in the feature mask. These sequential feature masks are first normalized via batch normalization with a gated linear unit[33] for the activation then passed to fully-connected layers for the classification head. We use the raw output of the fully connected classification layer for the optimization process, as[31] the implementation of the cross entropy loss in PyTorch uses unnormalized probabilities. During inference, we apply temperature scaling to return calibrated probabilities for each cell type.

Interpretability

In SIMS the input features correspond to the genes used for cell type prediction by the classifier. Unlike other machine learning models in which computational restrictions force reduced input data representation,[42,107] SIMS can be trained on the entire transcriptome for each cell.

TabNet, which serves as the foundation for SIMS, enables interpretability through the calculation of the weights of the sparse feature masks in the encoding layer. This allows for an understanding of which input features were used in the prediction process at the level of an individual cell. Furthermore, by averaging the sum of the attention weights across all samples for a given cell type, it is possible to determine the features used per class, while averaging across all cells in a sample shows the total features used when classifying the entire dataset. Similar to other deep learning models,[25] in SIMS the weights do not represent differential gene expression but a measure of the relevance (positive or negative signal) of said gene in the distinction between cell types. Additionally, the sparsity introduced in the sequential attention layers via the sparsemax prior acts as a form of model regularization,[30] allowing us to categorize a cell type via only a small number of genes.

Code library details

The SIMS pipeline was designed with an easy-to-use application programming interface (API) to support a streamlined analysis with minimal code. To achieve this goal, the pipeline was constructed primarily using PyTorch Lightning, a high-level library that aims to improve reproducibility, modularity, and simplicity in PyTorch deep learning code. We used Weights and Biases to visualize training metrics, including accuracy, F1 score, and loss, to facilitate the assessment of model performance. To accommodate the large data formats used by SIMS, we implemented two methods for data loading: a distributed h5 backend for training on h5ad files and a custom parser for csv and delimited files that allows for the incremental loading of individual samples during training. These same

methods are also used for inference. In addition, cell-type inference can be performed directly on an h5ad file that has been loaded into memory. This allows for efficient handling of datasets that may exceed the available memory capacity. We strongly support the use of h5ad files as they are faster and more efficient than plain text files and allow for more straightforward data sharing in the Python-scanpy environment. All the code and instructions to use SIMS are available in the Braingeneers GitHub repository: https://github.com/braingeneers/SIMS and in the Zenodo archive: https://doi.org/10.5281/zenodo.11095105.

Web application

In parallel to the API we also developed a web application in Streamlit. In this case the web application allows for quick and easy inference based on pretrained models. The user only needs to input the single cell RNA dataset in the h5ad format, select the pretrained model they want to use and perform the predictions. The application is hosted in the streamlit developer cloud, allowing access from anywhere without the need of institutional credentials.

Laboratories interested in sharing models created with their data with the public can request to include their pretrained models in our repository for easy hosting with a pull request to our repository https://github.com/JesusGF1/sims_app.

Training details

For all models benchmarked, the Adam optimizer[108] was used. The learning rate varied but was generally between 0.003 and 0.01, while the weight decay (L2 regularization) was between 0 and 0.1. To numerically encode the vectors, we used a standard one-hot encoding, where for K labels we have that the kth label is given by the standard basis vector e_k of all zeros except a 1 in the kth position. We define the loss function as

$$L(X,Y) = -\frac{1}{M}\sum_{i=1}^{M} w_i y_i \log(f(x_i))$$

where ix represents the transcriptome vector for the ith sample, y_i is the encoded label, w_i is the weight and M is the size of the batch. For our model, we defined w_i as the inverse frequency of the ith label, in order to incentivize the model to learn the transcriptomic structure of rarer cell types. The final signal to update the model weights was calculated as the average across all entries in the loss vector.

A learning rate optimizer was used such that $l \leftarrow 0.75l$ when the validation loss did not improve for twenty epochs. In all cases, models reached convergence by the early stopping criterion on validation accuracy before the maximum number of epochs (500) was reached. Gradient clipping was used to avoid exploding gradient values, which was required to avoid bad batches exploding the loss and stopping convergence. Although we used a train, validation and test split for reducing overfitting via hyperparameter tuning bias, the only hyperparameters tuned were the learning rate to avoid divergence in the loss and weight decay to avoid overfitting in the smaller datasets. Convergence took around 20–100 epochs for all models. For all models, we found model training to be consistent and had few cases of suboptimal convergence due to poor initialization. The train, validation and test sets were stratified, meaning the distribution of labels is the same in all three (up to an error of one sample, when the number of samples for a given class was not divisible by three), except for the ablation study, where there were not enough samples to stratify across all three splits.

For all benchmarks, models were trained using the most granular annotation available. When F1 score is mentioned in benchmarks it refers to the Macro F1-score.

Datasets details

Peripheral blood mononuclear cells (PBMC68K) dataset

Also known as Zheng68K is the PBMC dataset described in.[39] The dataset was generated using 10X Genomics technologies and sequenced using Illumina NextSeq500. It contains about 68,450 cells within eleven subtypes of cells. The distribution of cell types is imbalanced and transcriptomic similarities between cell types make classification a difficult task. Due to these properties, the PBMC68K dataset is widely used for cell type annotation performance assessment.

Human landscape: Han's dataset

The Human cellular landscape dataset described in.[103] The dataset was generated using Microwell-seq technology. It contains 584,000 cells with 102 different cell types across all major human organs and different developmental time points from more than 50 different donors.

Human heart: Tucker's dataset

The Tucker dataset described in[40] is a single nuclei RNA-sequencing dataset comprised of 287,269 cells representing 9 different cell types (20 cell subtypes) from 7 different donors.

Human kidney: Stewart's dataset

The kidney dataset described in[104] is a single cell RNA-sequencing dataset comprised of 40,268 mature human kidney cells representing 34 different cell types from 14 different donors.

Human lung: Krasnow's dataset

Krasnow's dataset described in[41] is a single cell RNA-sequencing dataset comprised of 75,400 cells representing 58 different cell types from 3 different donors and corresponding to two different sequencing technologies Smart-seq 2 and 10X chromium.

Adult mouse cortical and hippocampal dataset

This dataset was generated by the Allen Brain Institute and described in.[45–47] The dataset was generated from male and female 8-week-old mice labeled using pan-neuronal transgenic lines. The dataset includes micro-dissected cortical and hippocampal regions. It contains 42 cell types including excitatory projection neurons, interneurons and non-neuronal cells.

Adult human cortical dataset

This dataset was generated from postmortem samples by the Allen Brain Institute.[45,46] It includes single-nucleus transcriptomes from 49,495 nuclei across multiple human cortical areas. The large majority of nuclei are contributed from 3 donors: 1) H200.1023 was a female Iranian-descent donor who was 43 years old at the time of death. The cause of death was mitral valve collapse. 2) H200.1025 was a male Caucasian donor who was 50 years old at the time of death. The cause of death was cardiovascular. 3) H200.1030 was a male Caucasian donor who was 57 years old at the time of death. The cause of death was cardiovascular. For sampling, individual cortical layers were dissected from the middle temporal gyrus, anterior cingulate cortex, primary visual cortex, primary motor cortex, primary somatosensory cortex and primary auditory cortex. All samples were dissected from the left hemisphere. As part of the purification processes, nuclei were isolated and sorted using Fluorescently Activated Cell Sorting (FACS) using NeuN as a marker. For statistics, we only used cell types that were common between all samples.

Developing mouse cortical dataset

This dataset was described in.[62] It contains microdissected cortices from mice ranging from embryonic day 10 to postnatal day 4. For this study we used data from mice at embryonic day 12 (1 batch, 9,348 cells), 13 (1 batch, 8,907 cells), 14 (1 batch, 5249 cells) and 18 (2 batches, 7,137 cells), as well as postnatal day 1 (2 batches, 13,072 cells). Of note, only postnatal day 1 samples had Ependymocytes, and as such, they were removed for inter-age testing.

Human cortical organoids dataset

We used 6-months old organoids described in.[76] The dataset contained cells derived from 3 cell lines: GM8330 (3 organoids, 1 batch, 15,256 cells), 11A (3 organoids, 1 batch, 25,618 cells) and PGP1 (6 organoids 2 batches, 46,989 cells). PGP1 has a strong batch effect which is almost entirely caused by one organoid in batch 3. The dataset was generated using Chromium Single Cell 3′ Library and Gel Bead Kit v.2 (10x Genomics, PN-120237) and sequenced using the Illumina NextSeq 500 instrument. Of note, one of the cell lines had a cell cluster named "Callosal Projection Neurons" while others had "Immature Callosal Projection Neurons. Given the naming inconsistency, we aggregated both clusters as "Callosal Projection Neurons".

Human fetal brain development

We used fetal tissue representative of the second trimester of human development, specifically focusing our analysis on data sourced exclusively from the neocortex. This study encompassed the sampling of six distinct neocortical regions. The dataset contained samples from gestational weeks 14, 17, 18, 19, 20, 22, and 25. The number of cells contained in this dataset was around 404,000.[85]

Benchmarking against cell type classification models

We compared our model to:

scBERT 1.0. scBERT is a transformer architecture based on the deep learning model BERT. It has been adapted to work with single cell data and it offers interpretability as the attention weights for each gene.[25]

scNym 0.3.2. scNym is a neural network model for predicting cell types from single cell profiling data and deriving cell type representations from these models. These models can map single cell profiles to arbitrary output classes.[28]

scANVI 1.0.2 scANVI (single-cell ANnotation using Variational Inference) represents a semi-supervised approach designed specifically for single cell transcriptomics data. It relies on the use of variational autoencoders as the foundational component of its model architecture.[27]

SciBet 1.0. SciBet is a supervised classification tool, consisting of 4 steps: preprocessing, feature selection, model training and cell type assignment, that selects genes using E-test for multinomial model building.[42]

Seurat 4.0.3. We used Seurat's reference-based mapping, with the Transfer anchor settings, where very transcriptomically similar cells from the reference and query datasets are used to create a shared space for the two datasets.[19]

SingleR 1.6.1. SingleR is a reference-based method that requires transcriptomic datasets of pure cell types to infer the cell of origin of each of the single cells independently. It uses the Spearman coefficient on variable genes and aggregates the coefficients to score the cell for each cell type.[20]

Pseudotime analysis

The human cortical organoid dataset was parsed into R (v. 4.2.1) using Seurat and its dependencies (v. 4.3.0) and converted into a CellDataSet (CDS) for further analysis using Monocle 3 Beta (v. 3.1.2.9; https://cole-trapnell-lab.github.io/monocle3/).[86] Cell clusters and trajectories were visualized using the conventional Monocle workflow, as detailed in https://cole-trapnell-lab.github.io/monocle3/docs/trajectories/.

Cell stress analysis

We performed cell stress analysis using Gruffi. Gruffi is a computational algorithm that identifies and removes stressed cells from brain organoid transcriptomic datasets in an unbiased manner.[83] It uses granular functional filtering to isolate stressed cells based on stress pathway activity scoring.[83] Gruffi integrates into a typical single-cell analysis workflow using Seurat.[83] In this paper we followed the default implementation shown in the GitHub repository to obtain a dataframe containing what cells were stressed based on Gruffi's default analysis https://github.com/jn-goe/gruffi.