A data-efficient deep learning tool for scRNA-Seq label transfer in neuroscience

Julian Lehrer*[1,2,3], Jesus Gonzalez-Ferrer*[1,2,4], David Haussler[1,2,4], Mircea Teodorescu[1,2,4,5], Vanessa D. Jonsson+[1,3,4], and Mohammed A. Mostajo-Radji+[1,2]

[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 Applied Mathematics, University of California Santa Cruz, Santa Cruz, CA, 95060, USA.
[4]Department of Biomolecular Engineering, 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.
*These authors contributed equally to this work.
+Co-senior authors.
-Correspondence to vjonsson@ucsc.edu (V.D.J.) and mmostajo@ucsc.edu (M.A.M.-R.)

ABSTRACT

Large single-cell RNA datasets have contributed to unprecedented biological insight. Often, these take the form of cell atlases and serve as a reference for automating cell labeling of newly sequenced samples. Yet, classification algorithms have lacked the capacity to accurately annotate cells, particularly in complex datasets. Here we present SIMS (Scalable, Interpretable Machine Learning for Single-Cell), an end-to-end data-efficient machine learning pipeline for discrete classification of single-cell data that can be applied to new datasets with minimal coding. We benchmarked SIMS against common single-cell label transfer tools and demonstrated that it performs as well or better than state of the art algorithms. We then use SIMS to classify cells in one of the most complex tissues: the brain. We show that SIMS classifies cells of the adult cerebral cortex and hippocampus at a remarkably higher accuracy than state-of-the-art single cell classifiers. This accuracy is maintained in trans-sample label transfers of the adult human cerebral cortex. We then apply SIMS to classify cells in the developing brain and demonstrate a high level of accuracy at predicting neuronal subtypes, even in periods of fate refinement. Finally, we apply SIMS to single cell datasets of cortical organoids to predict cell identities in previously unclassified cells and to uncover genetic variations in the developmental trajectories of 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.

1 Introduction

Next-generation sequencing systems have allowed for large scale collection of transcriptomic data at the resolution of individual cells. Within this data lies variability allowing us to uncover cell-specific features, such as cell type, state, regulatory networks, as well as inferring trajectories of cell differentiation and specification[1,2]. These properties are crucial to understand biological processes in healthy and diseased tissue. In addition, these properties allow to 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 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 issues 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) 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) 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) 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 sub-cerebral projection neurons in the mouse, but highly expressed in projection neurons of all cortical layers in humans[13,14]. In summary, manual annotation of every new dataset can lead to compounding error propagation and inconsistent single cell atlasses, potentially reducing their utility.

The development of software to automate single cell analysis has become an important and popular research topic[4,15,16]. 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. 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. Finally, the high-dimensional nature of transcriptomic data makes analysis statistically and computationally intractable[17]. These conditions make applying classical models such as support vector machines difficult[18].

In response, generative neural networks have become a popular frame work for their robustness to technical variability within data, scalability, and ability to capture biological variability in the latent representation of the inputs[19–21]. Recently, scBERT, a deep language model for cell annotation has been described[21]. scBERT leverages large unlabeled datasets to learn the landscape followed by a training step where the model is fine-tuned for cell annotations with labeled datasets[21]. The pre-training step limits scBERT's application to other sequencing technologies, species, or reference genomes. Therefore, new and more versatile tools could therefore become an important step in cell annotation.

Here we present SIMS (Scalable, Interpretable Machine Learning for Single-Cell), a new framework based on the model architecture found in TabNet[22]. SIMS is implemented in Pytorch Lightning[23], 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[22]. Importantly, TabNet does not require any feature preprocessing and has built-in interpretability which visualizes the contribution of each feature to the model[22]. 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 in complex datasets, such as peripheral blood samples. We apply SIMS to datasets of the mammalian brain and show a high accuracy in adult and developing tissue. We further apply SIMS to data generated from *in vitro* models, such as pluripotent stem cell-derived cortical organoids, to uncover developmental differences between cell lines, which allows us to assign an identity to previously uncharactherized cells. 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.

2 Results

2.1 Development of a TabNet-based framework for label transfer across single cell RNA datasets

We developed SIMS, a framework for label transfer of across single cell RNA datasets that uses TabNet as the classifier component (Figure 1)[22]. TabNet is a transformer-based neural network with sparse feature masks that allow for direct prediction interpretability from the input features[22].

In our framework, 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. The decoder then takes these encoded features and passes them through a fully-connected layer with batch-normalization and a generalized linear unit activation[24]. 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 (Figure 1). For accurate training, the user must input an annotated matrix of gene expression in each cell (Figure 1). After training and production of training statistics, the user can input a new unlabeled dataset (Figure 1). Of note, if the training data was normalized ahead of training, the user must normalize the unlabeled data in a similar manner. 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 designed to require minimal input from the users. To train the model, the user has to only input the data file of the training dataset, a file with the labels, and define the class label (Figure 1C). 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 cell labels and define the class label. The user will then input the new unlabeled dataset (Figure 1D).

2.2 Benchmarking SIMS with existing cell classifiers of single cell RNA data: Peripheral blood mononuclear cells (PBMC68K)

We performed the main benchmark test in the PBMC68K (also known as Zheng68K) dataset[25] derived from human peripheral blood mononuclear cells. The PBMC68K is a complex dataset, containing unbalanced cell clusters and cells of similar molecular identity. Due to these properties, it is a great dataset for benchmarking cell type annotation methods and has been widely used for this task.[26].

We benchmarked the performance of SIMS against other methods built based on different theoretical approaches[26], building on top of the benchmark performed in[21]. We applied a fivefold cross-validation strategy to not let randomness be a deciding factor for our conclusion. SIMS surpassed most label transfer methods in both accuracy (Figure 2; Supplemental Table 1) and Macro F1 score (Supplemental Figure 1; Supplemental Table 2).

2.3 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 a standard dataset, we then asked whether it could perform accurately in a highly complex tissue, such as the brain. We focused in adult mouse cortical and hippocampal data generated by the Allen Brain Institute[27–29].

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[30]. The cerebral cortex is organized in 6 layers, and several cortical areas, each with different composition and proportions of excitatory projection neurons, inhibitory interneurons, glial cells and other non-neuronal cell types[30]. The hippocampus, on the other hand, is part of the archicortex (also known as allocortex)[31]. It is further subdivided into cornu ammonis (CA), dentate gyrus, subiculum, and entorhinal area[31]. While the hippocampus also has a layered structure, made of 3 layers, the cell type composition and numbers vary greatly from those in cerebral cortex[31]. 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 excitatory projection neurons, inhibitory interneurons, endothelial and glia 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 (Figure 3 and Supplemental Figure 2).

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 (Supplemental Figure 3). The Macro F1 after training in 7% of the data is 0.90 (Supplemental Figure 3). Given the low amount of training data needed to obtain a high accuracy in label transfer, we conclude that SIMS is a data efficient machine learning model.

2.4 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 in tissues where obtaining single cells is difficult. For example, adult neurons are difficult to obtain, 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[32]. Importantly, single nuclei sequencing is compatible with cryopreserved banked tissue[33]. 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[34]. Moreover, single nuclei datasets are more prone to ambient RNA contamination from the lysed cells[35]. In the case of the brain, it has been observed that neuronal ambient RNA has masked the transcriptomic signature of glia cells, leading to incorrect classification of glia subclasses in existing atlases[35].

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[27,28]. We trained on 80% of the data and tested the model in the remaining 20%. Overall, we obtained an accuracy: 98.0% and a Macro F1-score of 0.974 (Figure 4; Supplemental Figure 4; Table 1).

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 (Supplemental Figure 5). Unlike classical methods that are based on the differential expression of marker genes, such as Seurat and other clustering-based pipelines[4,36], SIMS assigns probabilities based on a nonlinear combination of genes present and absent in the cell clusters, as well as the variance of gene expression within the cluster (Materials and methods). By measuring the average sequential attention weights across each cell type, we see that SIMS considers a sparse number of input genes to determine cell identity. For example, averaged over all cells for the PVALB+ Interneuron cell type, SIMS defines this cluster by expression LOC105376571, LDB2, NWD2 and TMX2 (Supplemental Figure 6). Microglia, on the other hand, are defined by the high expression of DIAPH2, TRIP12 and ZCCHC11, and the simultaneous absence of NWD2 and MAGI2 (Supplemental Figure 7).

We then asked how SIMS performs in trans-sample predictions. This dataset is made of 3 different postmortem samples. Namely: H200.1023, a 43 years old Iranian-descent woman; H200.1025, a 50 years old Caucasian male; and H200.1030, a 57 years old Caucasian male. We trained the model on one sample and tested it on the other 2 samples. We performed this

Training Sample	Testing Data	Accuracy	Macro F1-score
80% of Data	20% of Data	98.0%	0.974
H200.1023	H200.1025	94.0%	0.84
H200.1023	H200.1030	94.4%	0.865
H200.1025	H200.1023	93.1%	0.769
H200.1025	H200.1030	93.1%	0.779
H200.1030	H200.1023	95.8%	0.862
H200.1030	H200.1025	94.8%	0.87

Table 1: Trans-sample accuracies and Macro F1-scores for human adult cerebral cortex dataset

experiment in each possible combination, obtaining accuracies ranging from 93.1 to 95.8% (Figure 4; Supplemental Figure 8; Table 1; Supplemental Tables 3-8).

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 (Supplemental Tables 3-8). This is consistent with recent work showing that previously annotated single nuclei atlases of the brain often mask non-neuronal cell types[35]. In addition, we observed that Layer 4 Intratelencephalic neurons often get classified as generic Intratelencephalic neurons (Supplemental Tables 3-8). This is in agreement with the fact that Layer 4 Intratelencephalic neurons are a subset of Intratelencephalic neurons[37]. 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.

2.5 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 predicting cells of different ages. Classifying cells during development is challenging, as several spatiotemporal dynamics can mask the biological cell identities[38]. During cortical development, gene networks of competing neuronal identities first colocalize within the same cells and are further segregate postmitotically[30,39,40], likely through activity-dependent mechanisms[41,42].

To test the accuracy of SIMS at classifying developing tissue, we focused on mouse cortical development due to its short timeline[43]. In the mouse cortex, neurogenesis starts at embryonic day (E) 11.5, and it is mostly completed by E15.5[43]. Common C57BL/6 laboratory mice are born at E18.5[44]. Neonatal mice are timed based on the postnatal day (P)[44]. We took advantage of a cell atlas of mouse cortical development that contains 2 samples of E18 mouse embryos and 2 samples of P1 mice[39]. These timed samples, which are 1 day apart from each other represent timepoints at which all mouse neurogenesis is completed[43]. At these timepoints, neurons may still be undergoing fate refinement[45], and consequently retain fate plasticity, albeit limited[46–48].

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 (Supplemental Figure 9 A-B). Across 17 cell types, we find that the model predicts the labels with an accuracy of 84.2% and a Macro F1-score of 0.791 (Figure 5A; Supplemental Table 9).

We then tested SIMS by training on two P1 samples and testing the label transfer in two E18 samples (Supplemental Figure 9 C-D). We find that in this experiment, the label transfer accuracy drops to 73.6% and the Macro F1-score to 0.674 (Figure 5B; Supplemental Table 10). 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 glia cells that retain the ability to divide throughout life[49]. Indeed the major source of astrocytes in the cerebral cortex is other dividing astrocytes[49]. Consequently, the "Cycling Glia Cells" cluster is often predicted as astrocytes (Figure 5B).

In the neuronal lineage, we find that SIMS can accurately predict cell types. Interestingly, the second choice for classification follows closely related subtypes (Figure 5C). For example, the second choice for "Deep layer callosal projection neurons" is "Upper layer callosal projection neurons". Upper and deep layer callosal projection neurons share several indistinguishable features of neuronal identity, such as morphology and intrinsic electrophysiological properties[50]. At the molecular level, both identities are governed by the transcription factor SATB2[51,52], and to date, very few genes that distinguish both subtypes of callosal projection neurons have been identified[53,54]. Strikingly, the second choice selection for the majority of Upper layer callosal neurons is not deep layer callosal projection neurons, but "Migrating neurons". This is consistent with the fact that Upper layer callosal neurons are the last neurons to be generated during neurogenesis, and therefore, migrating neurons are likely of an upper layer callosal fate[30,45] (Supplemental Figure 9 E-F).

Then, 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 projection neurons, and rarely as callosal projection neurons of the deep layers (Figure 5D-E). While morphologically distinct, layer 4 neurons share transcriptional homology with callosal neurons[55]. 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[37]. In agreement, Layer 4 neurons that are mislocalized to the upper cortical layers retain an upper layer callosal identity and fail to refine their identity[56]. 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 the proper identification. Alternatively, SIMS may identify subtle differences in fate transitions that cannot be accurately pinpointed by traditional clustering methods in the reference atlases.

We then asked whether SIMS can accurately predict labels early in development. We focused on E12-E14, the peak of neurogenesis in the mouse cerebral cortex[43,45]. During this period, cortical neurons are still undergoing postmitotic fate refinement[39], and therefore their identity remains highly plastic[46,47]. We trained SIMS in E12 and E13 samples and predicted the cell types in an E14 sample. We find that SIMS can predict the labels of postmitotic neurons, including Cajal-Retzus cells, immature neurons and migrating neurons (Figure 5F-H; Supplemental Table 11). It also can accurately differentiate between apical progenitors (radial glia cells) and intermediate progenitors (Supplemental Table 11). Yet, a small subset of intermediate progenitors gets predicted as migrating neurons. Similarly, a subset of immature neurons get predicted as migrating neurons, and some migrating neurons get predicted as immature neurons (Figure 5F-H). These likely represent a continuum on fate specification. In addition, the accuracy of SIMS is affected by non-neuronal cells, in agreement with our previous observations (Figure 4). Specifically, microglia, pericytes and endothelial cells frequently get predicted as apical progenitors (Figure 5F-H; Supplemental Table 11). This is likely due to the low representation of these cell types in the training dataset[39], as there are less than 60 of each of these cell types in the training data.

2.6 SIMS can identify gene cell line-dependent gene trajectories and label previously uncharacterized cells in cortical organoids

Cortical organoids are a powerful tool to study brain development, evolution and disease[13,57,58]. Yet, 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[59]. Moreover, transcriptomic analysis of cortical organoids have revealed strong signatures of cell stress[60,61], which can impair proper cell type specification[62]. *In vitro* conditions also generate cell types of uncharacterized identity, that do not have an *in vivo* counterpart[61,63]. While some have argued that these cells should be removed from further analysis[63], the most common approach is to annotate them as "Unknown" cell clusters[57].

To understand whether SIMS could be used to uncover cell line differences and identify previously unassignable cells, we took advantage of a dataset of 6 months old human cortical organoids derived from 3 different cell lines, each with their own idiosyncrasy[57]. 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 3) two batches of cortical organoids derived from the PGP1 cell line, which contained major batch effects. One of those batches contributed to a significant number of "Unknown" cells (Figure 6A-B; Supplemental Figure 10).

First, we performed reciprocal label transfers between organoids generated from the 11A and the GM8330. Training on GM8330 organoids and predicting in 11A organoids, we find an accuracy of 71.9% and a Macro F1-score of 0.598 (Supplemental Figure 11; Supplemental Table 12). We observe a high degree of accuracy for most cell types tested, including cycling cells, radial glia, outer radial glia/astroglia, immature interneurons, and callosal projection neurons (Supplemental Figure 11; Supplemental Table 12). Cells in the "immature projection neurons" cluster often get classified as "callosal projection neurons" (Supplemental Figure 11). This is consistent with the fact that callosal projection neurons are the last projection neuron subtype to be generated in the cerebral cortex. Therefore, SIMS likely finds cells that are further in their developmental trajectories toward a callosal fate[53]. Notably, while GM8330-derived organoids contains a cluster labeled as "Unknown" cells which is lacking in 11A-derived organoids, only 2.7% of 11A cells get predicted as "Unknown". Performing the opposite experiment (Figure 6C-E; Supplemental Table 13), in which we train SIMS in 11A-derived organoids and predict the identity of GM8330-derived organoids we find that, as expected from the previous experiment, a large percentage (56.3%) of callosal projection neurons get predicted as immature projection neurons (Figure 6C-E). Strikingly, we find that a large percentage (68.7%) of GM8330-derived radial glia cells get classified as immature projection neurons (Figure 6C-E). Given that, radial glia cells are the stem cells of projection neurons[30], this observation likely represents a difference in molecular trajectories during fate transitions between both cell lines. Lastly, because the 11A-derived organoids lack an "Unknown" cell cluster, we asked whether training on GM8330 could predict the identity of those unknown cells. We find that SIMS classifies 73.71% of unknown cells as immature projection neurons. This is consistent with recent observations that organoid-generated cells without an *in vivo* counterpart are of a projection neuron fate[61]. In addition, we find that 11.50% of unknown cells get classified as outer radial glia/astroglia, and 10.82% as callosal projection neurons. UMAP projection allows us to visualize that proximity of the unknown identities with other cell clusters 6C-D). Altogether, we conclude that SIMS is an effective tool to classify single cell data and perform label transfer in *in vitro*-generated neurons. Moreover, we find that it SIMS can uncover previously

Predicted organoid	Trained on PGP1 Batch:	Accuracy	Macro F1
GM8330	Batch 1+3	78.4	0.648
GM8330	Batch 1	31.1	0.209
GM8330	Batch 3	36.7	0.264
11A	Batch 1+3	78.7	0.681
11A	Batch 1	39.4	0.263
11A	Batch 3	49.3	0.323

Table 2: Effect of cell number and batch effect in accuracy metrics

unidentified differences in gene trajectories that contribute to differentiation variation between cell lines.

We then asked whether a similar approach could be used to classify the large number of unknown cells in PGP1-derived organoids. We trained SIMS with 3 different datasets: 11A-derived organoids, GM8330-derived organoids, and GM8330-derived organoids without the unknown cell cluster (Figure 7A-D; Supplemental Tables 14-15). We then predicted the cell labels in PGP1-derived organoids. First, we checked the accuracy of the label transfer by comparing the agreements between predictions. We focused on the non-unknown cell clusters. There was a 69.97% agreement between PGP1 labels predicted using 11A-trained and GM8330-trained models. Similarly, there was a 73.52% agreement between 11A-trained and GM8330-trained without the unknown cell cluster. We find that in both cases, the majority of disagreements between predictions are in cells that GM8330-trained models call callosal projection neurons, while 11A-trained models predicts them as immature projection neurons, suggesting once again a difference in developmental trajectories between the 11A and GM8330 cell lines. In agreement with this hypothesis, the second and third most common disagreements are between cycling cells and ventral precursors, which are a subset of cycling cells of a different neuronal lineage[30].

Finally, we asked whether cells labeled as unknown in PGP1-derived cortical organoids could be assigned to an identity using SIMS trained on each of the other cell lines. We find that 10.9% of unknown cells are agreed to be of an outer radial glia/astroglia identity by both models (Figure 7G). However, the majority (50.7%) of cells are predicted to still be of unknown identity by GM8330. On the other hand, 11A, which lacks an "Unknown" cell cluster in the training, classifies them as immature projection neurons (41.6%) and outer radial glia/astroglia (28.5%) (Figure 7G). We therefore decided to classify in GM8330 without the "Unknown" cell cluster and found that 16.7% of unknown PGP1 cells get classified as outer radial glia/astroglia cells (Figure 7H). The other major classifications are predicted by GM8330 as immature projection neurons, while 11A predicts them as radial glia and outer radial glia/astroglia cells (Figure 7H). Once again, this suggest a difference in the genetic trajectories governing differentiation in GM8330 and 11A cell lines. Nevertheless, the agreement between both models in the outer radial glia/astroglia subset allows us to confidently assign that identity to previously unresolved cells.

2.7 Cell number increases label transfer accuracy regardless of batch effects

Traditional cell clustering and label transfer methods are often limited by batch effects in the reference datasets which hinder their applications[20,64,65]. Newer transfer learning approaches have overcome these limitations taking ideas from the field of Deep Learning applied to computer vision and Natural language processing.[20,64,65]. When looking at the developing mouse tissue (Figure 5), we noticed that SIMS performed suboptimally when trained in clusters with few cells (<60). Yet, the training data was generated in a single sequencing batch. To understand whether an increase in the number of cells and batches would increase the accuracy of SIMS, we took advantage of PGP1-derived cortical organoids. As discussed above, two batches of PGP1-derived organoids (named by the authors as batch 1 and 3) have strong batch effects[57] (Supplemental Figure 10). Moreover, batch 3 has a large percentage of cells of unknown identity (37.97%). We trained SIMS on each of the two batches, as well as the two batches combined. We then predicted the label of GM8330-derived organoids. We find that training on batch 1, SIMS has an accuracy of 31.1% and a Macro F1-score of 0.209, while training on batch 3 the accuracy is 36.7% and the Macro F1-score is 0.264 (Table 2; Supplemental Tables 16-17)). Strikingly, when training on both batches together, the accuracy increases dramatically to 78.4% and a Macro F1-score of 0.681 (Table 2; Supplemental Table 18).

Finally, we predicted the labels of 11A-derived organoids. We find that training on PGP1 batch 1 alone, SIMS has an accuracy of 39.4% and a Macro F1-score of 0.263, while training on batch 3 alone the accuracy is 49.3% and the Macro F1-score is 0.323 (Table 2; Supplemental Tables 19-20). Yet, training on both batches together gives an accuracy of 78.7% and a Macro F1-score of 0.681 (Table 2; Supplemental Table 20). We therefore conclude that a larger number of cells and batches is preferred over a smaller more curated dataset and that the variability that goes along with the batch effect is useful for SIMS classification as it helps SIMS pinpoint what genes are more important for the classification in a given cell type.

3 Discussion

Currently, over 1.5M cells per month are sequenced and archived through the different cell atlas projects[66]. With the lowering trends in sequencing costs the number of cells sequenced is increasing exponentially[3,66]. Yet, cell annotation remains as 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[21]. However, these approaches require large number of cells for pretraining their algorithms and advanced computational knowledge and resources to further train their models[21]. SIMS does not require pretraining, therefore avoiding large data files and increasing its versatility. An added advantage to SIMS is the speed in which the training can be performed, which allows for the users to run the program in their local computers. For example, training SIMS on 70K cells in a standard computer is accomplished in under one hour, while in other deep learning models, this tasks can take over 3 hours[21].

When it comes to interpretability SIMS is able to output a sparse selection of the most important genes, that can then be easily plotted in the Python ecosystem of Scanpy, while other tools[21] rely on external cross-platform packages. This can hamper the adoption of new users, including non bioinformaticians[67]. 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[67].

We designed SIMS as a low code tool for both, training and performing the 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, such as Azimuth[68]. This is meant to remove barriers in adoption by new labs, medical practitioners, students and non-experts alike. Unlike other deep learning models[21], SIMS can use genes that are defined by the user, allowing the label transfer in novel genomes, or using annotated genomes without standard nomenclature. Other deep learning approaches, such as scBERT[21], have been shown to work well with datasets of up to 16K genes. SIMS, being based on TabNet, and therefore optimized for tabular data[22], can work well with over 45K features (Figure 3). This property would allow, in principle, SIMS to be trained simultaneously on references of multiple species, species with large genomes such as the axolotl[69], as well as multimodal data including combined single cell gene expression and gene accesibility sequencing datasets[70].

After showing that SIMS performs as good or better than state of the art methods, we 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[30,45]. Several efforts, such as the BRAIN Initiative and others, exist to sequence neurons across different ages, species, and diseases[71]. While the neuroscience community has started efforts to agree on naming conventions across the increasing number of datasets[5], there is still significant naming 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%. Unlike many other tools, that define cells by the strong expression of marker genes[7,68], 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 performed at a lower accuracy. When applied to cortical organoids, these properties allowed us to identify subtle differences in developmental trajectories between organoids generated from different cell lines that are missed by traditional clustering approaches[57]. We therefore believe that SIMS can be of great value at studying developmental disorders, such as Autism, where cortical organoids have already shown cell-type dependent asynchronous developmental trajectories in different neuronal lineages[72]. Hybrid pipelines that integrate pseudotime-focused tools, such as BOMA[7], could become complementary to SIMS and have the potential to provide more comprehensive insights into these questions.

While we have shown that SIMS can accurately predict trans-sample labels, we have limited our work to samples within the same species. This is because neuronal subtypes diverge significantly between species[28]. However, some neuronal subtypes, such as medial ganglionic eminence-derived interneurons are transcriptomically more conserved across evolution than other primary neurons, including cortical projection neurons[13,28]. In the future, these interneuron subtypes could be used as a way to optimize SIMS to perform trans-species predictions.

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.

4 Material and methods

4.1 The SIMS Pipeline

The classifier component of the SIMS framework is TabNet[22], 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. The decoder then takes these encoded features and passes them through a fully-connected layer with batch-normalization and a generalized linear unit activation[24]. 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 for the given cell type to understand the average contribution of each gene for classification.

4.1.1 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 utilized during training to improve convergence and stability.[73]. 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 M_i is used. This mask is learned via the attentive transformer, and sparsemax normalization[74] is used to induce sparsity in the feature mask. These sequential feature masks are then passed to fully-connected layers for the classification head, first normalized via batch normalization with a gated linear unit[24] for the activation. In our case, we use the raw output of the fully connected classification layer, as[23] loss functions handle logits.

4.1.2 Interpretability

In SIMS the input features correspond to the genes used for cell type prediction by the classifier. Unlike other machine learning models in where computational restrictions force reduced input data representation[75,76], 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 utilized 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[21], 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[22], allowing us to categorize a cell type via only a small number of genes.

4.2 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 utilized 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.

All the code and instructions to use SIMS are available in the Braingeneers GitHub repository: https://github.com/braingeneers/SIMS

4.2.1 Training details

For all models benchmarked, the Adam optimizer[77] was used. The learning rate varied but was generally between 0.0003 and 0.001, 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. To define error in the model, average over the categorical cross-entropy loss function, defined as

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

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 training regime to learn the 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, we calculate the macro F1-score. For per-class accuracy, we use per-class recall which reports the total number of correctly identified samples out of the total number of ground truth samples for the given label.

4.2.2 Datasets

Peripheral blood mononuclear cells (PBMC68K) dataset. Also known as Zheng68K is the PBMC dataset described in[25]. 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 transciptomic similarities between cell types makes classification a difficult task. Due to these properties, the PBMC68K dataset is widely used for cell type annotation performance assessment. The dataset can be accessed at https://www.10xgenomics.com/resources/datasets/fresh-68-k-pbm-cs-donor-a-1-standard-1-1-0

Adult mouse cortical and hippocampal dataset This dataset was generated by the Allen Brain Institute and described in[27–29]. 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. The dataset can be accessed at https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-whole-cortex-and-hippocampus-10x

Adult human cortical dataset. This dataset was generated from postmortum samples by the Allen Brain Institute[27,28]. 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. The data was obtained from https://portal.brain-map.org/atlases-and-data/rnaseq/human-multiple-cortical-areas-smart-seq.

Developing mouse cortical dataset. This dataset was described in[39]. 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. The data was downloaded from the Single Cell Portal administered by the Broad Institute. https://singlecell.broadinstitute.org/single_cell/study/SCP1290/molecular-logic-of-cellular-diversification-in-the-mammalian-cerebral-cortex

Human cortical organoids dataset. We used 6-months old organoids described in[57]. 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 1 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". We downloaded the dataset from the Single Cell Portal administered by the Broad Institute. https://singlecell.broadinstitute.org/single_cell/study/SCP282/reproducible-brain-organoids#study-summary

4.3 Benchmarking against cell type classification models

We benchmarked SIMS using the PBMC68K dataset, as previously described[21]. Briefly, 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.[21]

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.[78]

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.[76]

Seurat 4.0.3. A single-cell data analysis pipeline, with different ways to perform cell type classification. In this case the reference based annotation pipeline was selected.[36]

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[79]

cellID 1.0. CellID is a statistical method for cell type annotation that does dimensionality reduction and extracts gene signatures for cells or groups, depending on the strategy we select.[80]

scmap 1.14. scmap is a reference-based annotation method. We can use the scmap_cluster strategy where cells are mapped to cell types in the reference dataset or the scmap_cell where cells are mapped to individual cells in the reference dataset.[81]

References

1. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell rna-sequencing for biomedical research and clinical applications. *Genome medicine* **9**, 1–12 (2017).
2. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. biotechnology* **37**, 547–554 (2019).
3. Angerer, P. *et al.* Single cells make big data: New challenges and opportunities in transcriptomics. *Curr. Opin. Syst. Biol.* **4**, 85–91 (2017).
4. Luecken, M. D. & Theis, F. J. Current best practices in single-cell rna-seq analysis: a tutorial. *Mol. systems biology* **15**, e8746 (2019).
5. Yuste, R. *et al.* A community-based transcriptomics classification and nomenclature of neocortical cell types. *Nat. neuroscience* **23**, 1456–1468 (2020).
6. Grabski, I. N. & Irizarry, R. A. A probabilistic gene expression barcode for annotation of cell types from single-cell rna-seq data. *Biostatistics* **23**, 1150–1164 (2022).
7. He, C. *et al.* Boma, a machine-learning framework for comparative gene expression analysis across brains and organoids. *Cell Reports Methods* (2023).
8. Guo, H. & Li, J. scsorter: assigning cells to known cell types according to marker genes. *Genome biology* **22**, 1–18 (2021).
9. Wang, Q. *et al.* Elevated hapln2 expression contributes to protein aggregation and neurodegeneration in an animal model of parkinson's disease. *Front. Aging Neurosci.* **8**, 197 (2016).
10. Wonders, C. P. & Anderson, S. A. The origin and specification of cortical interneurons. *Nat. Rev. Neurosci.* **7**, 687–696 (2006).
11. de Lecea, L., Soriano, E. *et al.* Developmental expression of parvalbumin mrna in the cerebral cortex and hippocampus of the rat. *Mol. Brain Res.* **32**, 1–13 (1995).
12. Lee, B. *et al.* Signature morpho-electric properties of diverse gabaergic interneurons in the human neocortex. *bioRxiv* (2022).
13. Mostajo-Radji, M. A., Schmitz, M. T., Montoya, S. T. & Pollen, A. A. Reverse engineering human brain evolution using organoid models. *Brain research* **1729**, 146582 (2020).
14. Zeng, H. *et al.* Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures. *Cell* **149**, 483–496 (2012).
15. Pasquini, G., Arias, J. E. R., Schäfer, P. & Busskamp, V. Automated methods for cell type annotation on scrna-seq data. *Comput. Struct. Biotechnol. J.* **19**, 961–969 (2021).
16. Zhang, Y., Aevermann, B., Gala, R. & Scheuermann, R. H. Cell type matching in single-cell rna-sequencing data using fr-match. *Sci. Reports* **12**, 1–14 (2022).
17. Kuo, F. Y. & Sloan, I. H. Lifting the curse of dimensionality. *Notices AMS* **52**, 1320–1328 (2005).

18. Akbani, R., Kwek, S. & Japkowicz, N. Applying support vector machines to imbalanced datasets. In *European conference on machine learning*, 39–50 (Springer, 2004).

19. Wang, T. *et al.* Bermuda: a novel deep transfer learning method for single-cell rna sequencing batch correction reveals hidden high-resolution cellular subtypes. *Genome biology* **20**, 1–15 (2019).

20. Lotfollahi, M. *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nat. biotechnology* **40**, 121–130 (2022).

21. Yang, F. *et al.* 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 (2022).

22. Arık, S. O. & Pfister, T. Tabnet: Attentive interpretable tabular learning. In *AAAI*, vol. 35, 6679–6687 (2021).

23. Falcon, W. Pytorchlightning/pytorch-lightning. *Pytorch Light.* (2020).

24. Shazeer, N. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202* (2020).

25. Zheng, G. X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. communications* **8**, 14049 (2017).

26. Abdelaal, T. *et al.* A comparison of automatic cell identification methods for single-cell rna sequencing data. *Genome biology* **20**, 1–19 (2019).

27. Tasic, B. *et al.* Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72–78 (2018).

28. Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).

29. Yao, Z. *et al.* A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell* **184**, 3222–3241 (2021).

30. Cadwell, C. R., Bhaduri, A., Mostajo-Radji, M. A., Keefe, M. G. & Nowakowski, T. J. Development and arealization of the cerebral cortex. *Neuron* **103**, 980–1004 (2019).

31. Anand, K. S. & Dhikav, V. Hippocampus in health and disease: An overview. *Annals Indian Acad. Neurol.* **15**, 239 (2012).

32. Joseph, D. J. *et al.* Protocol for isolating young adult parvalbumin interneurons from the mouse brain for extraction of high-quality rna. *STAR protocols* **2**, 100714 (2021).

33. Larson, A. & Chin, M. T. 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**, 1–8 (2021).

34. Thrupp, N. *et al.* Single-nucleus rna-seq is not suitable for detection of microglial activation genes in humans. *Cell reports* **32**, 108189 (2020).

35. Caglayan, E., Liu, Y. & Konopka, G. Neuronal ambient rna contamination causes misinterpreted and masked cell types in brain single-nuclei datasets. *Neuron* (2022).

36. Stuart, T. *et al.* Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).

37. De León Reyes, N. *et al.* Transient callosal projections of l4 neurons are eliminated for the acquisition of local connectivity. *Nat. Commun.* **10**, 4549 (2019).

38. Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of single-cell rna-seq data. *Nat. Rev. Genet.* **20**, 273–282 (2019).

39. Di Bella, D. J. *et al.* Molecular logic of cellular diversification in the mouse cerebral cortex. *Nature* **595**, 554–559 (2021).

40. Nowakowski, T. J. *et al.* Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* **358**, 1318–1323 (2017).

41. Ozair, M. Z. *et al.* hpsc modeling reveals that fate selection of cortical deep projection neurons occurs in the subplate. *Cell stem cell* **23**, 60–73 (2018).

42. Mostajo-Radji, M. A. & Pollen, A. A. Postmitotic fate refinement in the subplate. *Cell Stem Cell* **23**, 7–9 (2018).

43. Greig, L. C., Woodworth, M. B., Galazo, M. J., Padmanabhan, H. & Macklis, J. D. Molecular logic of neocortical projection neuron specification, development and diversity. *Nat. Rev. Neurosci.* **14**, 755–769 (2013).

44. Ciemerych, M. A. & Sicinski, P. Cell cycle in mouse development. *Oncogene* **24**, 2877–2898 (2005).

45. Lodato, S. & Arlotta, P. Generating neuronal diversity in the mammalian cerebral cortex. *Annu. review cell developmental biology* **31**, 699–720 (2015).

46. Rouaux, C. & Arlotta, P. Direct lineage reprogramming of post-mitotic callosal neurons into corticofugal neurons in vivo. *Nat. cell biology* **15**, 214–221 (2013).

47. Ye, Z. *et al.* Instructing perisomatic inhibition by direct lineage reprogramming of neocortical projection neurons. *Neuron* **88**, 475–483 (2015).

48. De la Rossa, A. *et al.* In vivo reprogramming of circuit connectivity in postmitotic neocortical neurons. *Nat. neuroscience* **16**, 193–200 (2013).

49. Ge, W.-P., Miyawaki, A., Gage, F. H., Jan, Y. N. & Jan, L. Y. Local generation of glia is a major astrocyte source in postnatal cortex. *Nature* **484**, 376–380 (2012).

50. Ramos, R. L., Tam, D. M. & Brumberg, J. C. Physiology and morphology of callosal projection neurons in mouse. *Neuroscience* **153**, 654–663 (2008).

51. Alcamo, E. A. *et al.* Satb2 regulates callosal projection neuron identity in the developing cerebral cortex. *Neuron* **57**, 364–377 (2008).

52. Britanova, O. *et al.* Satb2 is a postmitotic determinant for upper-layer neuron specification in the neocortex. *Neuron* **57**, 378–392 (2008).

53. Fame, R. M., MacDonald, J. L. & Macklis, J. D. Development, specification, and diversity of callosal projection neurons. *Trends neurosciences* **34**, 41–50 (2011).

54. Molyneaux, B. J. *et al.* Novel subtype-specific genes identify distinct subpopulations of callosal projection neurons. *J. Neurosci.* **29**, 12343–12354 (2009).

55. Leone, D. P., Srinivasan, K., Chen, B., Alcamo, E. & McConnell, S. K. The determination of projection neuron identity in the developing cerebral cortex. *Curr. opinion neurobiology* **18**, 28–35 (2008).

56. Oishi, K. *et al.* Identity of neocortical layer 4 neurons is specified through correct positioning into the cortex. *Elife* **5**, e10907 (2016).

57. Velasco, S. *et al.* Individual brain organoids reproducibly form cell diversity of the human cerebral cortex. *Nature* **570**, 523–527 (2019).

58. Velasco, S., Paulsen, B. & Arlotta, P. 3d brain organoids: studying brain development and disease outside the embryo. *Annu. review neuroscience* **43**, 375–389 (2020).

59. Hernández, D. *et al.* Culture variabilities of human ipsc-derived cerebral organoids are a major issue for the modelling of phenotypes observed in alzheimer's disease. *Stem Cell Rev. Reports* 1–14 (2021).

60. Pollen, A. A. *et al.* Establishing cerebral organoids as models of human-specific brain evolution. *Cell* **176**, 743–756 (2019).

61. Uzquiano, A. *et al.* Single-cell multiomics atlas of organoid development uncovers longitudinal molecular programs of cellular diversification of the human cerebral cortex. *BioRxiv* 2022–03 (2022).

62. Bhaduri, A. *et al.* Cell stress in cortical organoids impairs molecular subtype specification. *Nature* **578**, 142–148 (2020).

63. Vértesy, Á. *et al.* Gruffi: an algorithm for computational removal of stressed cells from brain organoid transcriptomic datasets. *The EMBO J.* **41**, e111118 (2022).

64. Park, Y., Hauschild, A.-C. & Heider, D. Transfer learning compensates limited data, batch effects and technological heterogeneity in single-cell sequencing. *NAR genomics bioinformatics* **3**, lqab104 (2021).

65. Zhao, Y., Cai, H., Zhang, Z., Tang, J. & Li, Y. Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. *Nat. communications* **12**, 5261 (2021).

66. Svensson, V., da Veiga Beltrame, E. & Pachter, L. A curated database reveals trends in single-cell transcriptomics. *Database* **2020** (2020).

67. Krampis, K. Democratizing bioinformatics through easily accessible software platforms for non-experts in the field (2021).

68. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).

69. Nowoshilow, S. *et al.* The axolotl genome and the evolution of key tissue formation regulators. *Nature* **554**, 50–55 (2018).

70. Jiang, F. *et al.* Simultaneous profiling of spatial gene expression and chromatin accessibility for mouse brain development. *bioRxiv* 2022–03 (2022).

71. Maitra, M., Nagy, C. & Turecki, G. Sequencing the human brain at single-cell resolution. *Curr. Behav. Neurosci. Reports* **6**, 197–208 (2019).

72. Paulsen, B. *et al.* Autism genes converge on asynchronous development of shared neuron classes. *Nature* **602**, 268–273 (2022).

73. Santurkar, S., Tsipras, D., Ilyas, A. & Madry, A. How does batch normalization help optimization? *Adv. neural information processing systems* **31** (2018).

74. Martins, A. & Astudillo, R. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International conference on machine learning*, 1614–1623 (PMLR, 2016).

75. Alquicira-Hernandez, J., Sathe, A., Ji, H. P., Nguyen, Q. & Powell, J. E. scpred: accurate supervised method for cell-type classification from single-cell rna-seq data. *Genome biology* **20**, 1–17 (2019).

76. Li, C. *et al.* Scibet as a portable and fast single cell type identifier. *Nat. communications* **11**, 1818 (2020).

77. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

78. Kimmel, J. C. & Kelley, D. R. Semisupervised adversarial neural networks for single-cell classification. *Genome research* **31**, 1781–1793 (2021).

79. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. immunology* **20**, 163–172 (2019).

80. Akira, C., Loredana, M., Emmanuelle, S. & Antonio, R. Cell-id: gene signature extraction and cell identity recognition at individual cell level. *bioRxiv* 2020–07 (2020).

81. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell rna-seq data across data sets. *Nat. methods* **15**, 359–362 (2018).

5 Declarations

5.1 Author Contribution Statement

D.H., M.T., V.D.J., and M.A.M.-R. conceived the project. J.L. and J.G.-F. performed the experiments. J.L., J.G.-F. and M.A.M-R. wrote the paper with contributions from all authors.

5.2 Funding Statement

This work was supported by Schmidt Futures (SF857) to D.H. and M.T.; National Human Genome Research Institute (1RM1HG011543) to D.H. and M.T.; National Science Foundation (NSF2134955) to D.H. and M.T.

5.3 Data Availability Statement

All data used in this paper comes from previously published datasets.

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

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

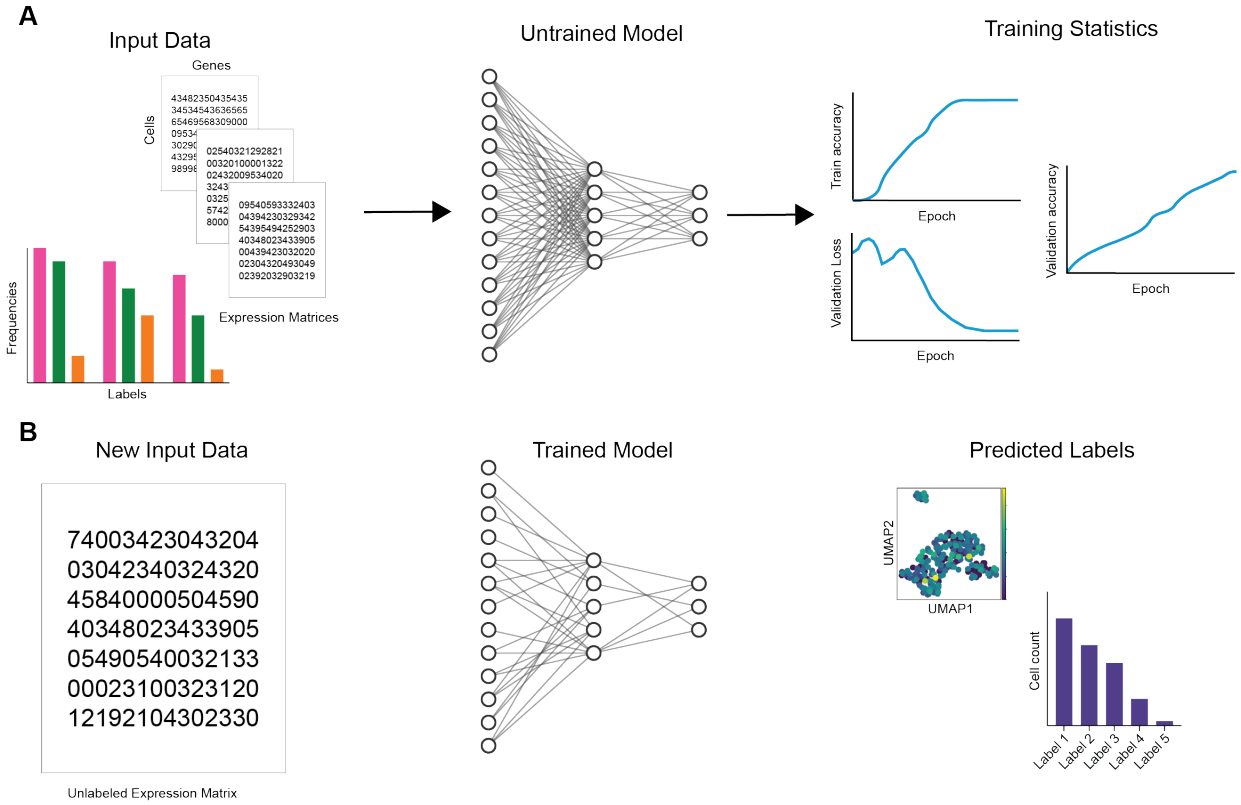
Human cortical organoids: https://singlecell.broadinstitute.org/single_cell/study/SCP282/reproducible-brain-organoids#study-summary

5.4 Declaration of interests Statement

J.L., V.D.J., and M.A.M.-R. have submitted provisional patent applications related to the work in this manuscript. The authors declare no other conflict of interest.

5.5 Acknowledgments

We would like to thank Tomasz Nowakowski, Maximilian Haeussler, Benedict Paten and Hunter Schweiger for their valuable feedback on this manuscript.



C

```python
from scsims import SIMS

# set up the base class with path to data and labels
sims = SIMS(
    datafiles=['datafile.h5ad'],
    labelfiles=['labelfile.csv'],
    class_label='cell_type',
)

sims.train() # Run the training loop
```

D

```python
from scsims import SIMS

# set up SIMS for label transfer with a trained model
sims = SIMS(
    weights="mymodelweights.ckpt"
    labelfiles=['originallabels.csv'],
    class_label='cell_type',
)

predictions = sims.predict('unlabeled_data.h5ad')
```

Figure 1: **Framework of the SIMS Model** A) The SIMS model can be trained with one or several pre annotated datasets. Training includes matrices of gene expression in individual cells, as well as cluster labels. The training model outputs the training statistics. B) A new unlabeled dataset can be inputted into SIMS, which will then apply the trained model to transfer the labels. C) Code snippet on how to train a SIMS model. D) Code snippet on how to use a previously trained model to predict the labels for an unlabeled dataset.

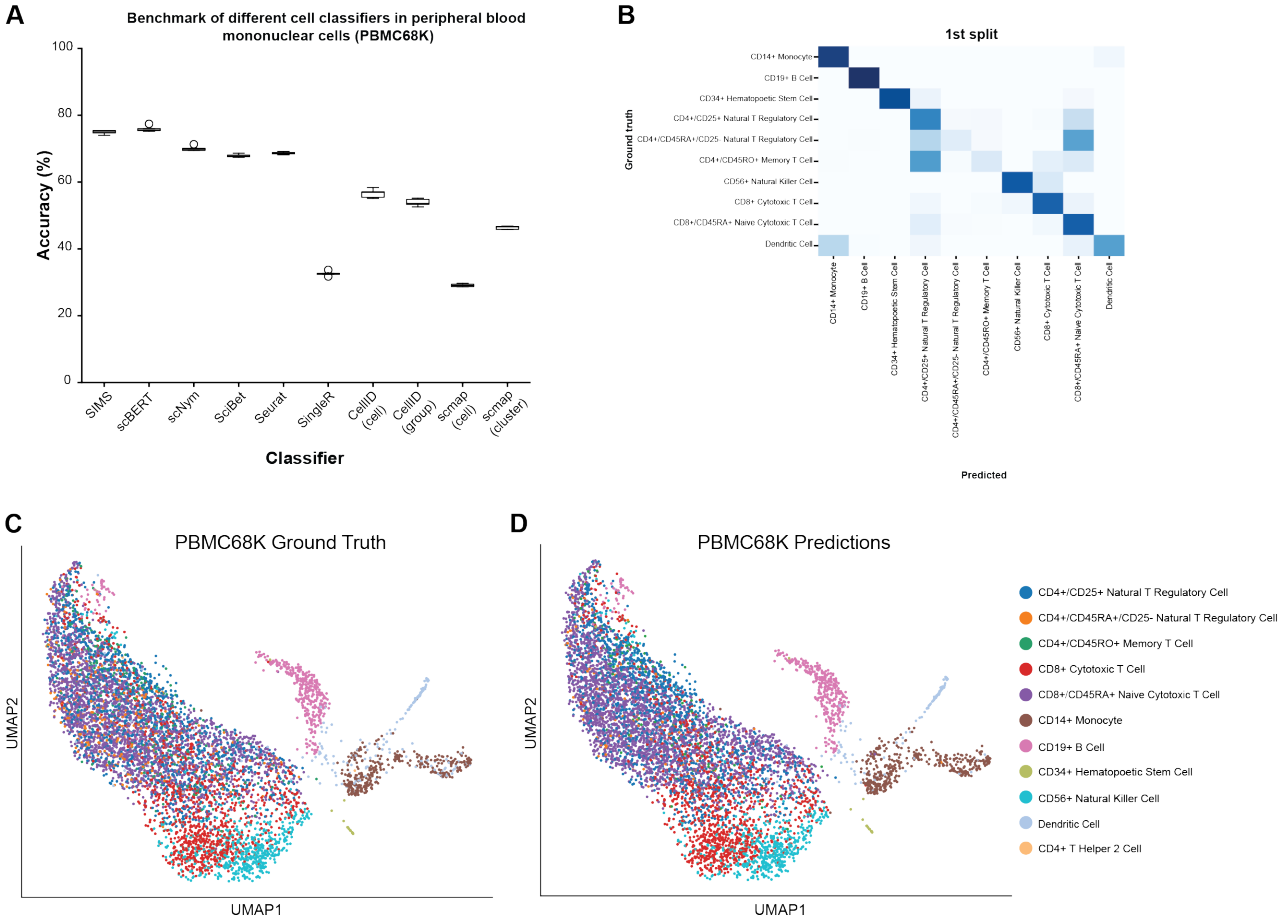


Figure 2: **Benchmarking SIMS against other cell classifiers**. A)Performance of cell type annotation methods measured by accuracy in the PBMC68K dataset using fivefold cross-validation. Box plots show the median (centre lines), interquartile range (hinges) and 1.5-times the interquartile range (whiskers). B) Confusion Matrix for the first split. In the Y axis we have the cell type as it was annotated in PBMC68K, in the X axis we have the prediction made by SIMS. Additional splits are available in Supplemental Figure 1. C) UMAP representation of the PBMC68K cells, colored by ground truth cell type. D) representation of the PBMC68K cells, colored by SIMS predicted cell type.

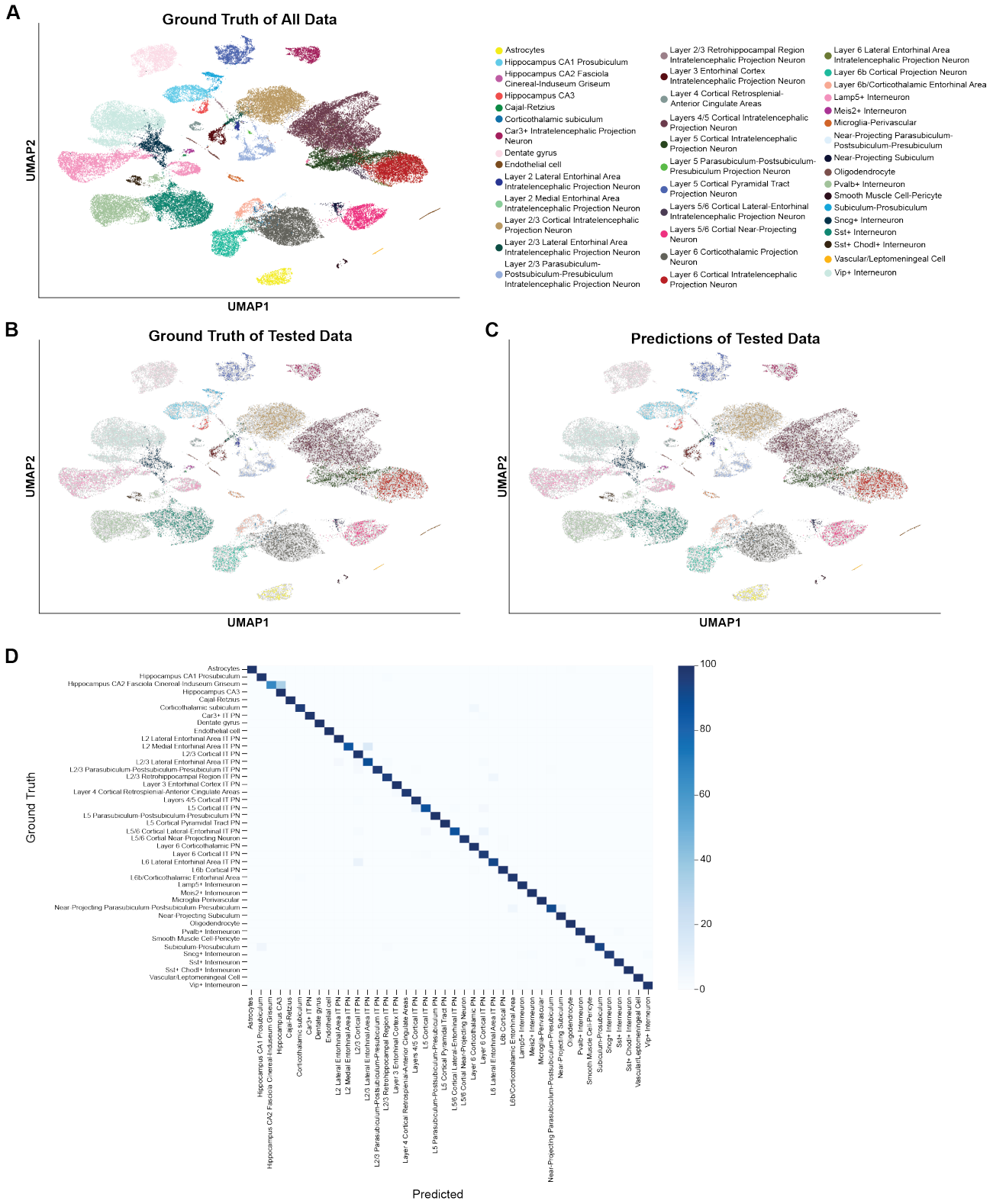


Figure 3: **Application of SIMS to single Cell RNA sequencing: Adult Mouse Cerebral Cortex and Hippocampus** A) Ground truth UMAP representation for the dataset. B) Ground truth UMAP representation for the Subset of Cells used for testing the algorithm in the train-test split. C) Predictions made by SIMS in that subset of data. D) Confusion Matrix for the test-split. L= Layer; IT = Intratelencephalic; PN = Projection Neuron.

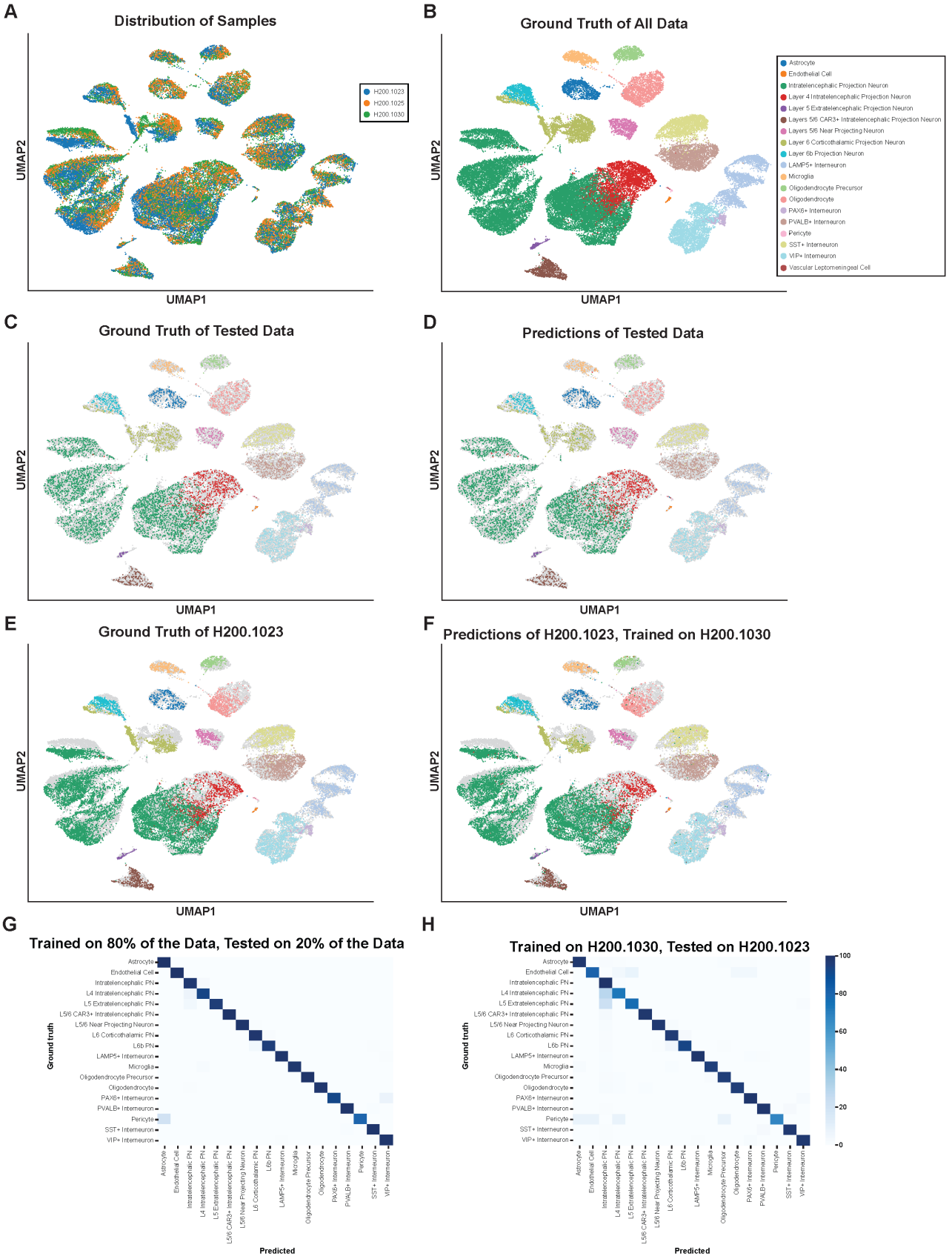


Figure 4: **Application of SIMS to trans-sample predictions of single Nuclei RNA sequencing: Adult human cerebral cortex** A) The adult human cerebral cortex dataset was comprised from three samples, each of which contributed to most clusters. B) Cell types for the whole dataset.

Figure 4: C) Ground truth for the test-split data. D) Predictions for the test-split data. E) Ground truth for the H200.1023 sample. F) Prediction for the H200.1023 sample after training on the H200.1030 sample. G) Confusion matrix for the test Split. H) Confusion matrix for the trans-sample experiment, where SIMS was training on H200.1030 and tested on H200.1023. L = Cortical Layer; PN = Projection Neuron. Additional examples are on Supplemental Figure 8.

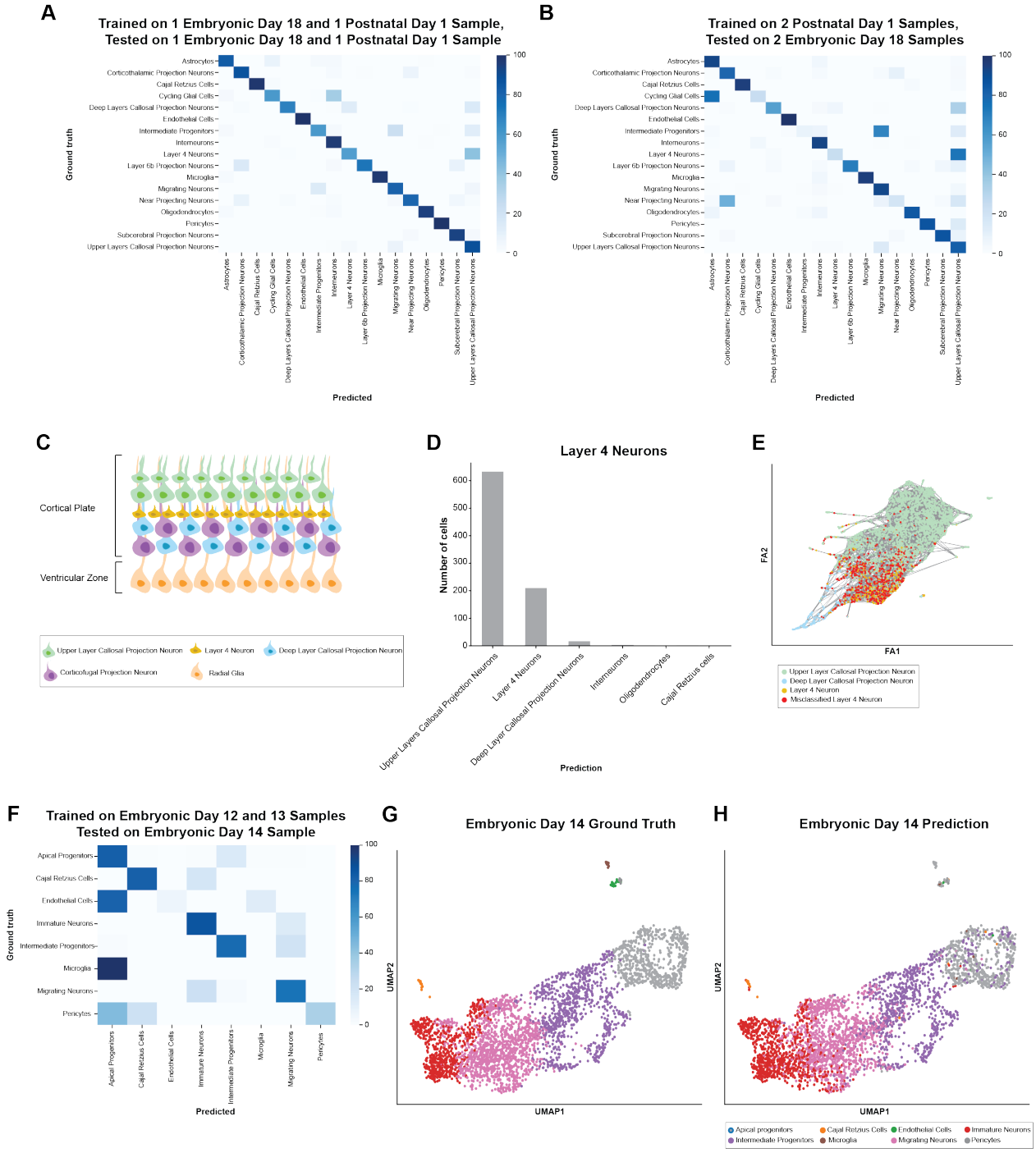


Figure 5: **Application of SIMS to developing tissue: Mouse cerebral cortex** A) Confusion Matrix for E18P1 split, where we trained on Sample 1 E18 and Sample 1 P1 and predicted on Sample 2 E18 and Sample 2 P1 B) Confusion matrix for E18 split where we trained on the P1 samples and predicted on E18 samples C) Diagram of the mouse cerebral cortex after neurogenesis. D) Barplot showing the number of Layer 4 Cells that get predicted as the different cell types. E) Force Atlas representation of Layer 4 Neurons. F) Confusion Matrix for E14 sample prediction, after training SIMS on E12 and E13 samples. G)Ground truth UMAP representation for the E14 sample H)Prediction by SIMS for the E14 sample.

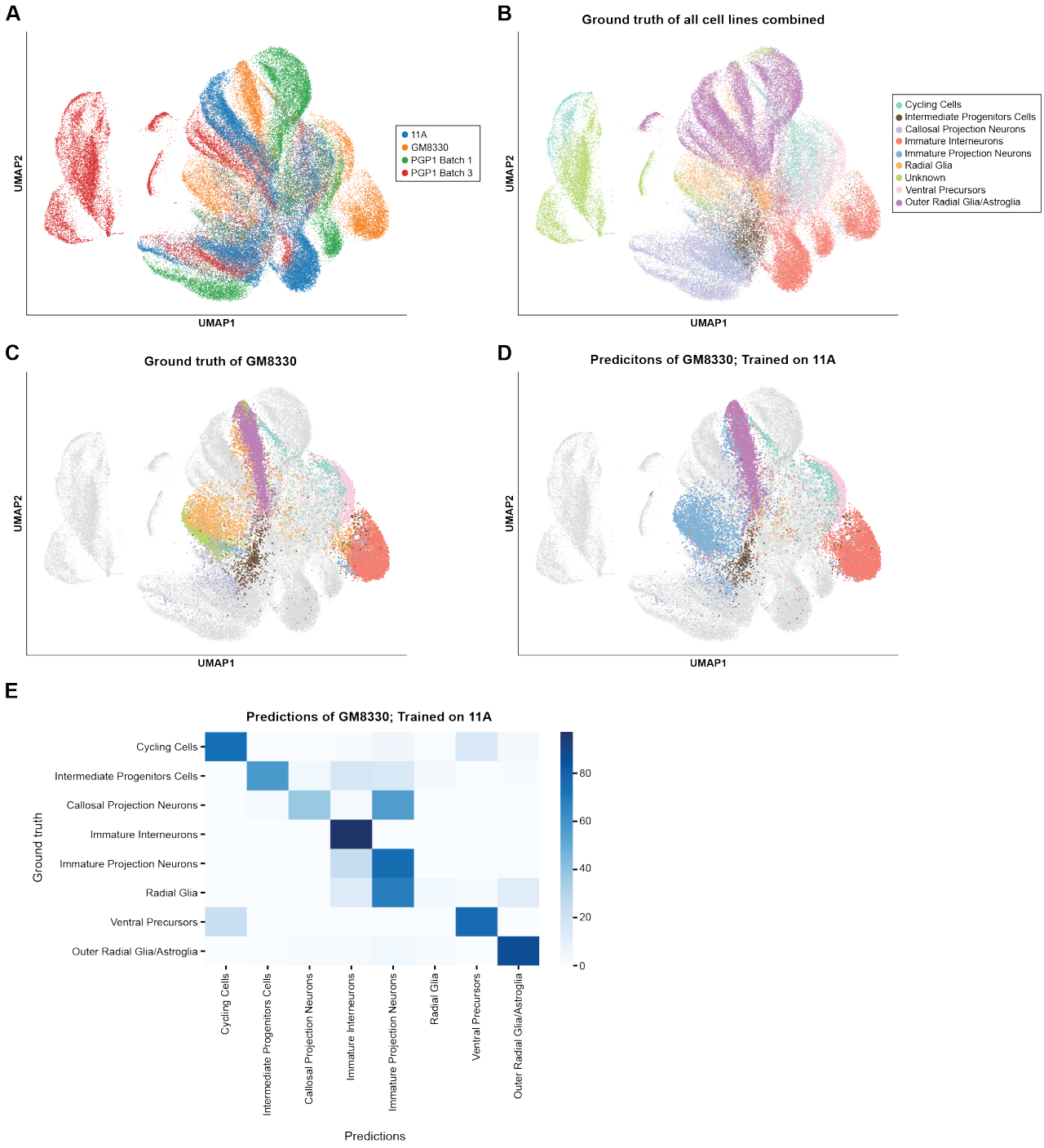


Figure 6: **Application of SIMS to *in vitro* generated models: human cortical organoids** A) UMAP representation of the 6 month organoids colored by batch and cell line. B) UMAP representation of the organoids colored by cell type. C) Ground truth cell type of GM8330-derived organoids. D) Predictions of GM8330 organoids by model trained on 11A organoids E) Confusion Matrix for GM3880-derived organoids, model trained on 11A-derived organoids.

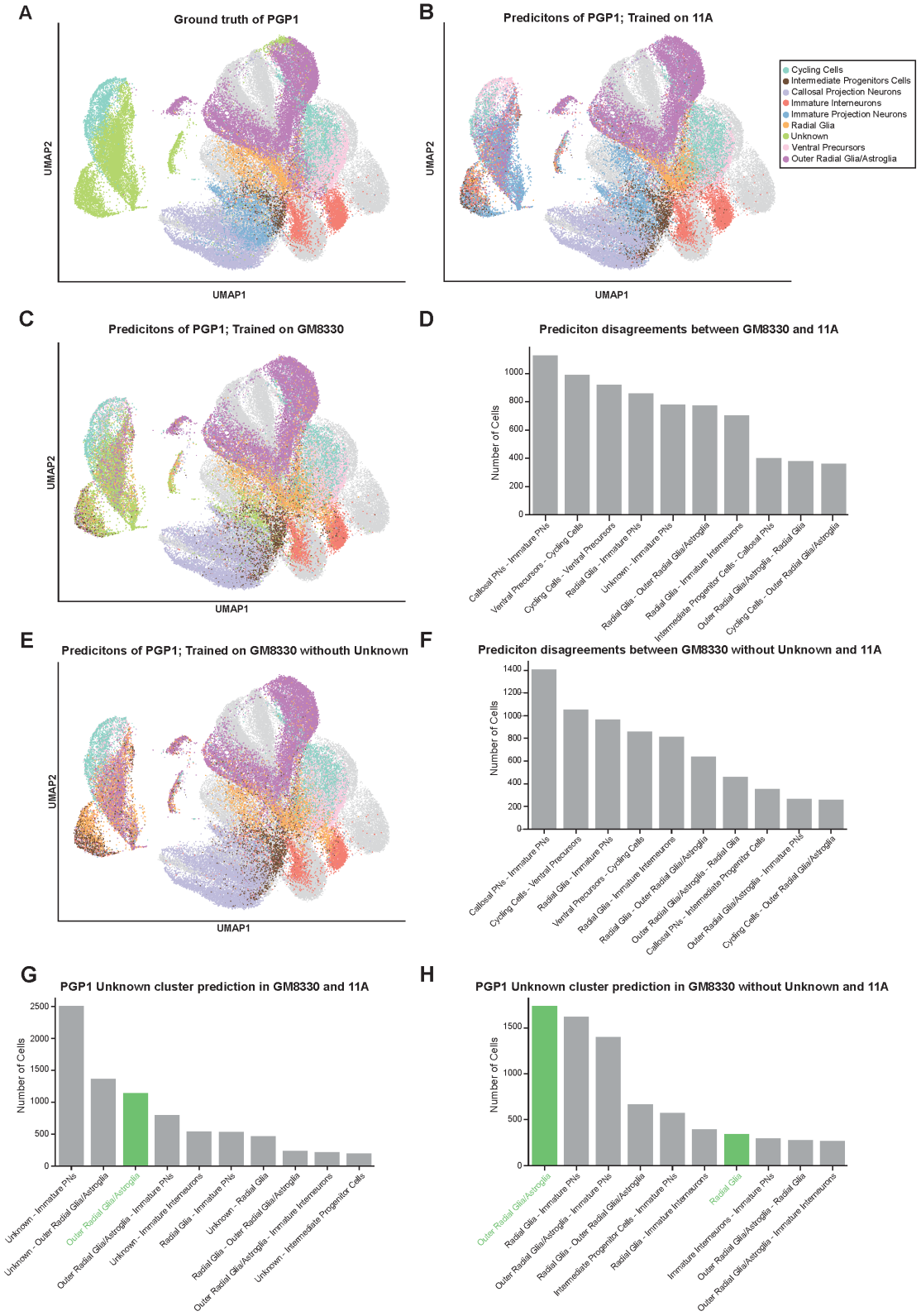


Figure 7: **Application of SIMS to *in vitro* generated models: assigning cell labels to previously unknown cells** A) Ground truth UMAP representation of PGP1-derived cortical organoids. B) UMAP representation of PGP1-derived organoids colored by prediction based on model trained on 11A organoids.

Figure 7: C) UMAP representation of PGP1-derived organoids colored by prediction based on model trained on GM8330-derived organoids. D) Prediction disagreements in cells annotated as not Unknown cells between the models trained on GM8330-derived (first named cell type) and 11A-derived organoids (second named cell type). E) UMAP representation of PGP1-derived organoids colored by prediction based on model trained on GM8330-derived organoids with Unknown annotated cells removed from the training set F) Prediction disagreements in cells annotated as not Unknown cells between the models trained on GM8330 without unknown-derived organoids (first named cell type) and 11-derived (second named cell type) organoids. G) Prediction disagreements and agreements in cells annotated as Unknown cells between the models trained on GM8330-derived (first named cell type) and 11-derived (second named cell type) organoids. H) Prediction disagreements and agreements in cells annotated as Unknown cells between the models trained on GM8330 without unknown-derived (first named cell type) and 11A-derived (second named cell type) organoids