Assessment of Cell type Annotation Tools of Single Cell RNA Sequencing Data

Hussein Lakkis

Department of Human Genetics

Faculty of Medicine and Health Sciences

McGill University, Montreal, Canada

February 2023

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science in Human Genetics

Abstract

Identifying and characterizing the gene expression profiles of cell types in normal and disease tissues is crucial for understanding biological processes and disease mechanisms. Single-cell RNA sequencing (scRNA-seq) provides cellular-resolution gene expression profiles, allowing for the annotation of cell populations based on these profiles. Traditionally, cell type annotation was performed by assessing canonical marker genes. Recently, various machine learning-based annotation tools have been developed to automatically infer cell type identities in new samples from large-scale labeled scRNA-seq datasets. However, the accuracy and limitations of these tools have yet to be thoroughly evaluated, particularly in complex scenarios such as annotation of cell types present during development, annotation of cells from one species using a reference from a different species, and annotation of cancer cells. Potential limitations of machine learning classifiers in these scenarios include difficulty handling complex biological variability and heterogeneity, which can result in poor performance and incorrect cell type annotations. This thesis aims to provide a comprehensive benchmark of cell type annotation tools in complex scRNA-seq scenarios, including developmental datasets, cross-species annotations, and cancer datasets. First, we assessed 17 scRNA-seq datasets from adult and embryonic tissues with complexity metrics to quantify traits of the datasets, including cell type continuity, redundancy, and hierarchical structure. Our comparison of the performance of 18 cell type annotation tools using cross-validation in these datasets showed decreased performance in prenatal developmental datasets. Next, we applied a subset of cell type prediction tools for cross-species prediction on the mouse and human datasets. This analysis demonstrated that individual tools had low cell type annotation accuracy in complex datasets, but a consensus of multiple tools improved accuracy. Lastly, we annotated cell types in brain tumor samples using scRNA-seq references of normal brain tissue and found that a consensus of three tools effectively labeled tumor cells. Overall, machine learning-based tools performed well in normal adult samples. In contrast, a consensus of tools with distinct algorithms improved prediction accuracy for cross-species and cancer datasets. We implemented our consensus annotation workflow into a modular, reproducible pipeline that can easily be extended to new tools. The results of this study offer insights into the performance of cell type annotation tools for complex datasets common in biological and biomedical research and guide the selection of appropriate cell type annotation tools and workflows for these applications.

Résumé

L'identification et la caractérisation des profils d'expression génétique des types de cellules dans les tissus normaux et pathologiques sont cruciales pour comprendre les processus biologiques et les mécanismes pathologiques. Le séquençage de l'ARN à l'échelle d'une cellule (scRNA-seq) fournit des profils d'expression génétique à résolution cellulaire, ce qui permet d'annoter les populations cellulaires en fonction de ces profils. Traditionnellement, l'annotation du type de cellule était effectuée en évaluant l'expression de gènes marqueurs canoniques. Récemment, divers outils d'annotation basés sur l'apprentissage automatique ont été développés pour déduire automatiquement les identités de type cellulaire dans de nouveaux échantillons à partir d'ensembles de données scRNA-seq marquées à grande échelle. Toutefois, la précision et les limites de ces outils doivent encore être évaluées de manière approfondie, en particulier dans des scénarios complexes tels que l'annotation des types de cellules présents au cours du développement, l'annotation de cellules d'une espèce à l'aide d'une référence d'une espèce différente et l'annotation de cellules cancéreuses. Les limites potentielles des classificateurs par apprentissage automatique dans ces scénarios incluent la difficulté à gérer la variabilité et l'hétérogénéité biologiques complexes, ce qui peut entraîner de mauvaises performances et des annotations de types cellulaires incorrectes. Cette thèse a pour but de fournir une analyse comparative complète des outils d'annotation de type cellulaire dans des scénarios scRNAseq complexes, y compris des ensembles de données sur le développement, des annotations inter-espèces et des ensembles de données sur le cancer. Tout d'abord, nous avons évalué 17 ensembles de données scRNA-seq provenant de tissus adultes et embryonnaires à l'aide de mesures de complexité supervisées afin de quantifier les caractéristiques des ensembles de données, notamment la continuité des types cellulaires, la redondance et la structure hiérarchique. Notre comparaison des performances de 17 outils d'annotation de type cellulaire par validation croisée dans ces ensembles de données a montré une diminution des performances dans les ensembles de données sur le développement prénatal. Ensuite, nous avons appliqué un sous-ensemble d'outils de prédiction de type cellulaire pour la prédiction inter-espèces sur les ensembles de données de la souris et de l'homme. Cette analyse a démontré que les outils individuels avaient une faible précision d'annotation des types cellulaires, mais qu'un consensus de plusieurs outils améliorait la précision. Enfin, nous avons annoté les types de cellules dans des échantillons de tumeurs cérébrales en utilisant des références scRNA-seq de tissus cérébraux normaux et nous avons constaté qu'un consensus de trois outils marquait efficacement les cellules tumorales. Dans l'ensemble, les outils basés sur l'apprentissage automatique ont donné de bons résultats seuls dans les échantillons adultes normaux. En

revanche, un consensus d'outils avec des algorithmes distincts a amélioré la précision de la prédiction pour des ensembles de données inter-espèces et cancéreux. Nous avons mis en œuvre notre flux de travail d'annotation par consensus dans un pipeline modulaire et reproductible, qui peut facilement être étendu à de nouveaux outils. Les résultats de cette étude offrent des informations précieuses sur les performances des outils d'annotation des types de cellules pour les ensembles de données complexes courants dans la recherche biologique et biomédicale et guident la sélection d'outils d'annotation des types de cellules et de flux de travail appropriés pour ces applications.

Table of Contents

List of abbreviations

RNA seq; RNA sequencing

NGS: Next-Generation Sequencing

scRNA-seq: Single Cell RNA Sequencing

ML: Machine Learning

DL: Deep learning

KNN: K-Nearest neighbor

SNN: Shared nearest neighbor

PCA: Principal component analysis

PC: Principal component variable

PBMCs: Peripheral Bone Marrow Cells

N1: Fraction of borderline cells

H1: Proportion of cell types belonging to a hierarchical structure

MST: Minimum spanning tree

F2: Volume pf overlapping region

SVM: Support vector machine

HGG: High-grade gliomas

pHGG: Pediatric high-grade glioma

UMAP; Uniform manifold approximation and projection

PFA-EP: Posterior fossa ependymoma

E16: Embryonic day sixteen

P6: postnatal day six

ASTR: Astrocytes

MGL: Microglia

OPC: Oligodendrocyte precursor cell

EPEN: Ependymal cells

OPAS: Proliferating OPC and astrocytic progenitor

NFOL: Newly forming oligodendrocytes

GLIP: Glial progenitors

NEURP: Neuronal progenitors

EXIP: Excitatory neuronal progenitors

INIP: Inhibitory neuronal progenitors

MOL: Mature oligodendrocytes

RGC: Radial glial cell

SPN: Spiny projection neurons

VLMC: Vascular lepotomeningeal cells

SMC: Smooth muscle cells

List of figures

List of Tables

Dedication and acknowledgments

I would like to thank my supervisor, mentor, and the sister figure throughout this journey for all that she has done for me over the last two years. Thank you, Claudia, for giving me this chance to grow as a scientist. I have learned so much from you. I also cannot forget the Kleinman lab members who there for me were whenever I needed help, especially Selin Jessa, Nisha Kabir, and Steven Herbert. Thank you, Selin, for being there whenever I had small questions. All your help got me here!

I would like to dedicate this thesis to my family, my late grandma, my mother, and my father, who have dedicated their lives and efforts for me to have a proper education. Thank you, mom, for keeping me motivated through these rough times. I cannot express enough gratitude and appreciation.

Preface:

Format of thesis

This thesis is organized in the traditional format and comprises six chapters. Chapter I includes an introduction to the subject of scRNA-sequencing-based cell type annotation using automated and ML approaches and ends with the hypotheses and objectives of this thesis. Chapter II details the methodology utilized with a description of the mathematical formulation of the complexity metrics, the datasets used, and the computational pipelines used for this work. Chapter III presents the results of the assessment of the machine learning tools on several datasets from prenatal and postnatal tissues. Chapter IV details the application of cell type annotation tools for predictions across species and normal/disease states. In chapter V, the results are discussed and positioned within the broader context of the field. Finally, chapter VI describes the main conclusions and discusses future directions.

Contribution of authors

The work described in this thesis was performed under the supervision of Dr. Claudia Kleinman.

I led all computational analysis performed in chapters III and IV.1 with the help of Samantha Worme in generating figure 16. Selin Jessa provided help generating figures 24-28.

1. Chapter I: Introduction

Single-cell RNA sequencing (scRNA-seq) is being rapidly adopted in biological research and has produced compelling discoveries of expression changes in disease. This section will briefly introduce theoretical concepts of scRNA-seq and several key features for analyzing scRNA-seq data. Cell type annotation, the primary focus of the work presented herewithin, is a critical step in any single cell study. Previously, annotating cell types in scRNA-seq data has relied on time-consuming, cluster-level labeling through prior knowledge of the expected cell types. However, with the availability of annotated reference datasets, machine learning approaches have become practical methods of reproducible and objective cell type annotation performed at the individual cell level. Furthermore, using reference-based automated approaches can reduce inconsistencies of cell type annotations across single cell studies. This chapter will provide a literature review on current machine learning-based annotation approaches and summarize previous benchmarking studies. Finally, we will discuss the complexities of specific datasets and annotation scenarios that have yet to be addressed.

1.1. Single-cell transcriptomic sequencing (scRNA-seq)

scRNA-seq is a technology capable of detecting mRNA molecules in single cells across populations spanning thousands of cells[1]. It allows for a quantitative measurement of gene expression in each cell, using next-generation sequencing, providing a high-resolution image of the cellular transcriptome. The single-cell resolution makes scRNA-seq an adequate tool for defining cell types, states, and functions within heterogeneous tissues.

The use of scRNA-seq technology has been growing exponentially for the past decade. The number of papers containing scRNA-seq data and the size and quality of published datasets have increased substantially in recent years[2,3]. New technological advances, including physical and biochemical innovations reported in the last decade, have led to the widespread application of scRNA-seq methodology in the field of biology[4]. The concept of single-cell transcriptomic sequencing was first pioneered by Jim Eberwine et al. [5] and Iscove et al. [6], who expanded complementary DNA (cDNA) of single cells by

linear amplification via *in vitro* transcription and PCR, respectively. Jim Eberwine first used this technology to show that morphologically similar hippocampus cells display marked differences in RNA expression[5]. These pioneering technologies were later applied to commercially available DNA microarray chips[5–7].

Subsequent developments in sequencing technologies continued to guide parallel developments in scRNA-seq. The first scRNA-seq study performed using next-generation sequencing (NGS) was published in 2009 by Tang et al.[8], one year after the first bulk RNA-seq experiment[9]. This method was first used to study the murine blastomere. Tang et al. demonstrated that this technology could detect 75% more genes than previous microarray techniques. Moreover, since the blastomere is a fast-developing stage of growth in an embryo, using single-cell to detect which genes are expressed allowed for fate-mapping and other types of analyses that bulk RNA-seq cannot provide[8]. Nevertheless, these first experiments could only sequence a few cells and required manual manipulation of cells. Advances in microfluidics technologies[10,11], the development of droplet-based cell encapsulation methods[12], and laser capture microdissection[13] have made it possible to increase throughput and sequence tens of thousands of cells in one experiment. These novel methods have driven the development of new advanced scRNA-seq protocols, such as CEL-seq2[14], Drop-Seq[15], Smart Seq2[16], and MULTI-seq[17].

The rapid technological advancement described above has resulted in the emergence of international consortia projects to map single-cell gene expression across tissues and ages to establish a baseline reference for comparison against disease conditions. One of these projects is the Human Cell Atlas (HCA), which aims to characterize all cell types across all organs of the human body[18]. This project is expected to ultimately generate data for billions of cells. In addition, the HCA project aims to generate data for disease states at a single-cell resolution[19]. In addition, several projects of narrower scope have also sought to profile organs or specific tissue types in detail at the single-cell level. For instance, LungMAP: The Molecular Atlas of Lung Development Program is a consortium funded by the National Institute of Health (NIH) to integrate single-cell transcriptomics

with other data types to improve our understanding of the lung's cellular ontology [20]. Other examples include the BRAIN initiative[21] and the Human and Mouse Allen Brain Atlases, which aim to characterize neural cells at the single-cell resolution and to understand the regulatory networks behind development, evolution, and neuropsychiatric disorders[22]. These consortia and international efforts have yielded large datasets that can be used to study cell physiology and dynamics of a population of cells, to discover rare cell types in normal healthy tissues[23–31], and to provide a novel view on disease states at high resolution[28–34].

Single-cell expression profiling has also increased our understanding of cellular developmental trajectories through spatiotemporal and pseudo-time analyses[35]. scRNA-seq provides the expression measurements of genes and transcription factors controlling cell states. These measurements can infer how cells differentiate from one state or cell type to another and order them on differentiation trajectories[36]. This *in silico* reconstruction allows for the detection of transient intermediate populations, which can be validated experimentally. This type of reconstruction also allows for the detection of gene changes and biomarkers as cells transition from one state to another. In a recent study, pseudo-time analysis was used to uncover potential biomarkers for the transition of liver cirrhosis to hepatocellular carcinoma, illustrating one application of this method of analysis[37]. Regulatory networks that control cellular activities in normal and disease phenotypes can also be inferred from scRNA-seq data. These analyses leveraging scRNA-seq data have furthered our understanding of tumor biology. Rare cell populations within a tumor, such as tumor stem cells that are associated with aggressive disease and relapse, can be identified through scRNA-seq analyses[38]. Furthermore, scRNA-seq can be used to identify new interactions between cancer cells and immune cells within the tumor microenvironment, which has implications in predicting the success of immunotherapy and other new treatment methodologies[39]. These limited examples demonstrate the vast and diverse information that can be garnered from scRNA-seq experiments.

1.2. Common scRNA-seq analysis workflow

A typical scRNA-seq experiment entails several steps: Single-cell barcoding, scRNA-seq library preparation and sequencing, preprocessing scRNA-seq data, and downstream analysis[40–44]. Regardless of the capture and sequencing protocol, the first two steps usually generate sequencing reads that are first demultiplexed, aligned, and mapped. Data that contains unique molecular identifiers (UMIs) or barcodes then needs to be collapsed and counted, often by means of read processing software (such as CellRanger), to generate a gene expression matrix (N (cells) x M (genes))[45]. Finally, the raw expression matrix must undergo quality control (QC) and corrections before further analysis.

QC is an essential step of any scRNA-seq analysis workflow[40–42,46]. Batch effects, inefficient dissociation, RNA leakage and degradation, dropouts due to low expression of mRNA, and many other factors can lead to technical artifacts that can influence downstream analyses[42,46,47]. QC metrics such as the number of counts per barcode, fraction of mitochondrial genes per barcode, and cell cycle scores, a score based on the expression of G2/M and S-phase markers, can be used to filter out low-quality cells[48]. Following quality control, scRNA-seq data is normalized to remove unwanted sources of variation in the data related to technical variability. This normalization can help prevent highly expressed genes from affecting analyses[43]. scRNA-seq datasets may also undergo data correction and imputation to adjust for batch effects and dropouts[40–44].

The following steps in a conventional analysis workflow are feature selection, dimensionality reduction, and clustering. scRNA-seq suffers from the curse of dimensionality, where the number of genes detected is significantly higher than the number of cells sequenced[49]. The high number of features increases the computational intensity and noise of the data. Moreover, only a handful of genes are required for a meaningful biological inference, and most information required to analyze the data can be summarized in a few dimensions[50]. The first step of dimensionality reduction is feature selection. Highly variable genes (HVGs) are often selected as they retain most of the information in the dataset[51]. Dimensionality reduction can then be performed using linear

methods such as Principle Component Analysis (PCA)[52] or non-linear techniques such as t-distributed Stochastic Neighbor Embedding (t-SNE)[53], or Uniform Manifold Approximation, and Projection (UMAP)[54]. Following dimensionality reduction, a ubiquitous downstream involves defining discrete cell populations. For this, cells are clustered into groups of similar gene expression profiles. This can be done using graph-based methods such as Leiden or Louvain, hierarchical clustering, K-nearest neighbors (KNN), and others[55]. The resulting clusters require proper annotations to identify the biologically relevant subpopulations in the dataset (Cell type annotation)[56,57]. This step is crucial as the properly annotated clusters serve as input for downstream analyses such as differential gene expression (DGE) analysis, identifying marker genes, and modeling gene dynamics across the cell populations.

1.3. Cell Type Annotation

Cell type annotation or classification is the process by which clusters of cells or individual cells are annotated with a specific label that is associated to a particular biological function, group or property. Cell annotation is an active area of research as it is often a critical step for scRNA-seq experiments across various applications. Consistent cell classification allows to compare similar cells found in different body parts or distinct species[58,59]. Furthermore, several downstream analyses require the proper clustering followed by the identification of cell types to identify marker genes, model cell population dynamics, gene dynamics, and many other analyses. Correct annotation is essential for the biological interpretation of the results from these analyses. The presence of poorly annotated cells can negatively affect a single cell workflow by adding bias and noise to the analysis[60]. Accurate annotation allows the derivation of specific cell type signatures to help identify and characterize rare cell types in tissues that bulk RNA-seq misses[61].

Cell type annotation has previously been performed using canonical marker genes[56,57]. The cell clusters can then be annotated based on their expression of cell type-specific marker genes or canonical markers found in previous studies[56,57,62]. Annotation using marker genes assumes that cell types can be distinguished based on the expression of cell type-specific genes that are lowly expressed or absent in other cell types. An example

of this would be the use of CD3 to distinguish T lymphocytes from other leukocytes[63]. Researchers have used this gene to annotate many leukocyte scRNA-seq datasets published[64–67]. However, this methodology is time-consuming and highly dependent on the available literature. Clustering also plays a significant role, as the same dataset can have a different number of clusters depending on the choice of algorithm, and the parameters used can often lead to under-clustering of data[68]. In addition, this technique may miss rare and new cell types due to a lack of available well-studied markers. Therefore, curation using marker genes makes it difficult to generate reproducible, objective, and scalable annotations for the growing scRNA-seq datasets[69]. This technique can result in different studies having conflicting annotation outputs due to the subjectivity of manual curation.

1.4. Reference-based cell type annotation

An alternate approach to cell type annotation consists of training machine learning models on pre-annotated scRNA-seq references, i.e., supervised learning. To exploit the abundance of published data, much research has focused on developing automated, fast, and scalable tools that can decipher cellular compositions of single-cell experiments and identify individual cell types. Furthermore, these tools do not depend on clustering to annotate cells but instead perform the classification at the individual cell level. Therefore, this approach also offers an objective methodology for cell type annotation, which has the potential to reduce annotation inconsistencies across single cell studies.

There are several classes of methods for automated cell type prediction[70]. Recent publications include supervised learning tools (e.g., ACTINN[71], SciBet[72], scPred[73], LAmbDA[74], CaSTLe[75], singleCellNet[76], SVM[77]). There are also correlation-based tools such as SingleR[78], scmap-cell, and scmap-cluster[79]. Correlation-based tools can annotate cells by labeling cells as the closest cell type in the training reference using similarity scores, such as Pearson correlation. In addition, we also have automated marker-based tools (Garnett[80], SCINA[81], DigitalCellSorter[82]) that retrieve prior knowledge from marker databases such as PangaloDB[57] and then annotate cells or clusters based on the expression of marker genes. Lastly, there are hierarchical tools that predict cell

types given a hierarchical clustering constraint of cell types (scHPL[83], CellO[84], scClassify[85], CHETAH[86]). The list of automated cell type prediction tools is expanding, with many tools published in the past year alone. Currently, there are more than 140 tools reported for cell type classification[87].

The aforementioned types of cell type prediction tools can be further stratified. Supervised learning tools can be further stratified into discriminative learning tools such as scPred[73], which uses support vector machines (SVM), and generative modeling approaches such as SciBet [72] which uses multinomial distribution models. Discriminative modeling uses the conditional probability and the borders between classes in a feature space to classify a cell. In contrast, generative models use Bayesian approaches and multinomial models based on joint probabilities and maximum likelihood to classify cells[88]. Moreover, we can also segregate discriminative learning tools into neural network approaches and non-neural network approaches. For example, NeuCA[89] uses neural networks with multiple hidden layers to learn a model that can then be used to predict cell types. ACTINN is also another neural network approach that uses only one hidden layer.

Several deep-learning approaches have also been presented for cell type annotation. Deep learning approaches utilize artificial networks with representation or feature learning[90]. This learning can be performed in a supervised, semi-supervised, or unsupervised manner. Examples of these tools include KPNN[91], which utilizes graph convolutional networks to extract features from gene interaction networks that can aid in classifying cells. Moreover, there are autoencoder methods and unsupervised learning techniques that leverage neural networks with bottlenecks to learn features and ignore the noise in the data. One tool, scIAE, has been shown to perform well for feature extraction, cell type annotation, and predicting disease status[92].

Finally, it is essential to mention that hierarchical cell type annotation tools can use any mathematical approach. Still, the constraints imposed on training and prediction distinguish them and merit them their own class.

1.5. Limitations of ML based cell type annotation

While ML-based cell type annotation has the potential to increase labeling objectivity and consistency, limitations to these methods exist and need to be considered prior to their incorporation into an analysis workflow. Several factors can reduce the performance of cell type classifiers on unseen datasets[93]. ML issues that may arise can be classified into two main categories: data-related limitations and algorithm-related limitations.

One main limitation of ML-based cell type annotation is the dependence on reference quality and balance. Single-cell experiments profile heterogeneous cell populations that are not equally represented in the datasets. Due to this, some cell populations, such as rare ones, are often under-represented in the datasets. Class imbalance is problematic as many ML algorithms assume that the data is evenly distributed among the classes. Class imbalance can result in a strong bias towards the classes with the highest prevalence. Another problem affecting ML-based cell type annotation is reference quality and how well it matches the query datasets[94,95]. Single cell experiments often suffer from dropouts and technical noise (batch effects) that can result in poor performance.

In addition, most ML-based classifiers have assumptions potentially violated with scRNA-seq datasets. For example, cell types in these datasets are often highly similar and dependent, which can affect model performance[96]. This is particularly relevant for single-cell studies profiling overlapping cell types, such as the ones tracking developmental lineages. As a result, many classifiers may underperform in these scenarios.

Moreover, most classifiers might underperform in scRNA-seq datasets due to the curse of dimensionality, where the number of features is larger than the number of data points[97,98]. The curse of dimensionality often leads to overfitting the training data, resulting in models that cannot be generalized to testing scenarios. Lastly, neural network approaches are not interpretable (Blackbox approaches) and require large amounts of data, in the tens of thousands to millions of data points, to outperform traditional ML approaches such as Naïve Bayes and tree-based approaches[99]. All these factors can potentially make ML-based cell type annotation less effective in specific scenarios.

1.6. Cell type classifier benchmarking

More than 140 tools have been developed for cell type annotation. To date, five studies have been reported to benchmark the performance of these tools on different single-cell datasets and testing scenarios[100–104]. Abdelaal et al.[100] benchmarked 32 different automated cell type prediction tools on 11 datasets of varying size, complexity, and annotation level using cross-validation for intra-dataset testing. They employed an inter-dataset assessment where the classifiers were trained on a reference dataset and then used to annotate external datasets. This assessment allowed the authors to evaluate prediction performance across different technologies. The authors also assessed datasets based on inter-correlation between cell types as a complexity measure. Their principal findings indicate that performance varies significantly between classifiers depending on the number of genes, number of cells, annotation level, and dataset complexity. This data highlights the importance of choosing the suitable classifier for a given scenario. Furthermore, their findings suggest that a linear SVM with a rejection option for cells classified with low confidence outperforms other tools in most scenarios. Another study that benchmarked nine tools showed that cell type similarity, dataset unbalance, and size play a significant role in performance[104]. In this study, Seurat, SingleR, and CASTLE performed the best. It was also stated that using a simple majority vote (consensus) between tools outperformed individual predictions by the classifiers.

A more recent benchmarking study[101] found that combining multiple references improves performance by reducing overfitting. The authors recommend using multilayered perceptron (MLP) as the classifier and F-test as a feature selection protocol for selecting genes to include in training[101]. Equally important, a study in 2021 benchmarked 10 R-based cell type annotation tools on six publicly available datasets, including peripheral blood mononuclear cells (PBMCs) and Tabula Muris, an atlas of mouse cell types across many organs, in addition to simulated datasets with varying differential expression scales using Splatter[102]. They reported that tools perform well in intra-dataset cell type annotation but struggle to generalize for inter-dataset annotation. They also found that SingleR and Seurat performed the best across the different scenarios. Moreover, they

reported that tool performance suffers in cases of high cell type similarity. Lastly, Xie et al. assessed thirty-two tools classified into three categories (eager or discriminative, lazy or generative, and marker-based) on four primary datasets[103]. Their findings suggest that top performers from each category perform similarly across the different datasets. However, eager and lazy methods (reference-based approaches) are recommended due to speed and accuracy when pre-annotated references are available. Moreover, it was noted that the performance of marker-based approaches depends on the quality of prior information and the number of cell types, which is consistent with previous conclusions from Abdelal et al. Perhaps, the most important takeaway was that embryonic and tumor datasets pose a challenge for classification tools. Embryonic and developmental datasets may include granular and highly similar cell types, which can decrease classification performance. We should note that the recent benchmarking studies were mainly performed on standard datasets and datasets with a small number of very distinct cell types, such as PBMCs[45] (datasets profiling well-defined peripheral blood mononuclear cells) and datasets profiling a low number of discrete cell types (Baron[105] and Muraro[106] pancreatic datasets). Notably, the datasets used in these studies differ significantly from many scRNA-seq datasets that span tens to hundreds of cell types.

1.7. Specific datasets and scenarios that may pose challenges of ML-based cell type annotation

While previous benchmarking studies of cell type annotation tools offered valuable insights into the performance of cell type annotation tools, most of these studies, used gold-standard benchmarking datasets such as PBMCs and pancreatic datasets and focused on biologically simple scenarios (using a reference that ideally matches the query) where annotation of most cells is successful. These studies did not thoroughly assess the classifiers with samples collected throughout different developmental time points, from different species, or from tumor samples. Here, we will highlight some key characteristics of these applications and how we suspect this might affect the performance of classifiers. We will also present the rationale for the work undertaken in this thesis.

A primary limitation in benchmarking studies is the lack of comprehensive datasets representing a high diversity of cell types. These benchmarking experiments utilized datasets published in past years, such as CellBench[107] (5 distinct lung adenocarcinoma cell lines), Baron[105] (discrete cell types in the adult murine and human pancreases), Zheng[45] (PBMCs), and others. However, current references cover more cell types, include more cells, and capture more genes because of improved sequencing techniques[18,20,22]. In particular, an important focus of research relevant to several fields (development, cancer, developmental and neurological disease, among others) relies on prenatal or embryonic datasets that characterize organ and tissue development from the zygote to birth in humans and other species. These datasets may be more difficult to label due to their inherent characteristics arising from the biology of these tissues.

Three main characteristics are present in prenatal datasets that may pose challenges for cell annotation methods: gene expression redundancy across cell types, cell type or state continuity, and high similarity between cell types along a common hierarchy or lineage.

First, prenatal datasets have high redundancy of cell types and gene expression across the cell types. Expression redundancy means a specific gene can have similar expression values across multiple cells and cell types[108]. This is a characteristic of embryonic datasets because cells profiled during development often express various programs necessary for differentiation[109]. These programs often overlap between cell types, leading to expression redundancy. In differentiated tissues, most cells are fully committed to one cell type and can be distinguished from others in the same tissue. These differentiated cell types often express distinct signatures with a relatively lower overlap of gene expression profiles. This feature might impact machine learning-based cell type prediction because features with similar values between classes have low predictive power and are often removed during feature selection to enhance performance. Moreover, these features tend to overfit the models, leading to decreased performance in query datasets[110].

Second, many experiments generating prenatal data often use discrete timepoints to track developmental processes that cells undergo. This sampling is problematic because cells continuously evolve between these time points and at different rates. Gene expression profiles are not binary in these cases and often represent a spectrum of continuous states[111–113]. Defining cell types and cell states is not an easy task[114]. For example, progenitor cells often have transient intermediate states in the developmental lineage where cells express highly similar expression profiles forming a spectrum[115]. The spectrum of differentiating cells in these datasets includes many transient or intermediate cell types that are difficult to characterize and annotate adequately due to a lack of discriminating markers. As a result, researchers label these cells with discrete classes when biologically, these cells form a continuous spectrum of states rather than discrete states along the trajectory[116,117]. This labeling results in cells often being annotated as the closest cell type to them when they are expressing intermediate profiles between two cell types. We suspect this will cause problems for cell type annotation tools that assume that label classes (here, cell types) are discrete and independent. Furthermore, due to the high degree of expression profile similarity between cell types, cell type prediction tools show high rates of unlabeled cells[100,102]. For these cases of continuous cell states, pseudo-time analysis has proven to be an adequate model for the transition between states or types[30,131].

Third, prenatal datasets often show a hierarchical structure in which we have a progenitor to differentiated cell type relationships with intermediate states that might not be captured due to the discrete timepoint sampling[118]. For example, in the developing brain, neuroectodermal cells are comprised of two main hierarchical lineages; neuronal and glial[119,120]. Many well-performing tools assume cell types are independent and discrete rather than a continuous spectrum. This assumption is unfounded in datasets from cells undergoing differentiation, which often exhibit gene expression redundancy with other cells in the same lineage. However, some tools consider hierarchical constraints on the labels included in the training. Recent algorithms, including CHETAH[86], CellO[84], scClassify[85], and scHPL[83] consider these hierarchical constraints and classify cells into intermediate states or nodes if the cell cannot be accurately classified into a cell type. In

theory, this hierarchical consideration of cell types should mitigate the inability to label cells in cases where intermediate cell types are present [120]. It is imperative to mention that the three abovementioned characteristics are not mutually exclusive and can be related.

While we have described specific characteristics of prenatal datasets, we also note the lack of assessments of automatic classifiers for cancer cell type projection or prediction. A common practice for cell type annotation in cancer research is to use developmental scRNA-seq references to interrogate cell type programs in tumor cells[33,34,121–123]. This is done by annotating cancer cells as their most similar cell type in a healthy reference. We use here the term projection since this classification is done by projecting these cells to cell types of a specific cell population in an atlas. This information, related to the cell type of origin, can guide experimental studies and therapeutic targeting in these cancers. However, this projection method generates additional challenges as cancer cells have aberrant gene expression patterns that confound annotation[124]. Tumor cells have copy number variations and epigenetic changes, further distinguishing them from the transcriptional states found in a normal developmental gene atlas. A cancer cell may also express genes specific to other cell types to have a survival advantage. One example of this phenomenon is the immune checkpoint ligand PD-L1 expression that is typically expressed on various immune cells[125]. These genetic alterations can complicate tumor cell type projection using ML approaches, as multiple cell type programs may be expressed.

Despite these challenges, ML-based cell type classification approaches have been used in cancer transcriptomics. Early studies used machine learning classifiers to identify tumor cells from normal cells using scRNA-seq data[126,127]. However, this classification used only two classes (normal vs. tumor) without considering tumor cell types. Machine learning was also used to identify signatures that could be used to trace back the tissue of origin of metastatic tumors[128]. Importantly, this study only provided insight into the tissue of origin but did not give any information as to the cell type of origin of the tumor. Lastly, some classifiers have the potential to annotate cancer cell types but require annotated cancer references and have not been tested using normal references[129]. A major

limitation of this approach is the lack of well-annotated cancer references in most tumor types. The studies above illustrate the need for a robust and accurate cancer cell type projection method using normal scRNA-seq references.

In addition, many studies in developmental and cancer biology rely on scRNA-seq data generated from animal models, such as the mouse or zebrafish. Importantly, cell annotation has yet to be extensively evaluated in cross-species predictions. Human and mouse single-cell data differ due to practical, ethical, and technical considerations. Mouse data is abundant, and sampling is more accessible, especially for developmental data. In contrast, ethical restraints make the same not feasible for human tissues, especially during fetal development. Ideally, high-resolution mouse reference datasets could be generated and used to annotate human datasets. These datasets would allow researchers to overcome some of the constraints encountered when accessing human tissues. Currently, some tools perform cross-species cell type annotation, such as SingleCellNet and SciBet[72,76]. However, these tools have not been adequately benchmarked against other tools not designed for this task.

The challenging characteristics of developmental (prenatal) datasets and the limitations to annotating cancer and cross-species datasets highlight the need for careful evaluation of cell type annotation tools used in these scenarios. In this thesis, we undertake a systematic benchmarking assessment of cell type identification tools that consider the abovementioned characteristics, including control of dataset complexity and quantification of these characteristics. We will then assess tool classes in different annotation scenarios that vary in complexity. Our hypotheses and specific aims are discussed in the following section.

1.8. Hypotheses and Aims:

We hypothesize that intrinsic features of particular datasets, such as prenatal or cancer datasets, may variably impact the performance of different machine-learning approaches for cell type annotation. These features may be quantified with complexity metrics regarding continuity, redundancy, and hierarchical structure.

In this thesis, we aim first to compare the performance of cell type annotation tools in a prenatal vs. postnatal context using cross-validation approaches. Then, using dataset complexity metrics, we will relate the performance of different tools to a set of defined dataset characteristics. The findings are presented in Chapter III.

Second, we aim to assess the performance of cell type annotation tools for two specific tasks: cross-species predictions and identification of predominant cell-lineage programs in cancer datasets. We present our results in Chapter IV.

2. Chapter II: Methodology:

2.1. Cell type classification methods

18 cell type annotation tools were included in this study (Table 1). We installed all R-based tools with R version 4.0.5 except for NeuCA, which was used with R version 4.1.2 due to its unavailability in previous R versions. Python-based tools were installed with Python version 3.6.5. The classifiers were run with their default settings, including hyperparameters and parameters. In the cases where default parameters were missing, we utilized the values provided in the vignettes or the accompanying studies (Table 2). We provided each tool with gene counts per cell as input. We also performed log normalization for specific tools (SVM, SVM Rejection, and SciBet) when it was required to run them. Eight classifiers included an option to reject cells with low classification probability (Rejection). In these cases, we used the default rejection thresholds set by the original authors (Table 2).

Table 1. Automated cell type classifiers used in this thesis. Neural Net: Neural Network approaches; DGE: differential gene expression; KNN: k- Nearest Neighbors.

Name	Type	Language	Underlying Method	Year	Version	Rejection option of low probability predictions
Support Vector Machine (SVM)[130]	Discriminative	python	Linear SVM classifier using the scikit-learn package	1992	0.23.2	no
SVM Rejection[130]	Discriminative	python	Linear SVM classifier with rejection option set at 0.7	1992	0.23.2	yes
SingleCellNet[76]	Discriminative	R	Random forest classifier on genes selected with DGE analysis	2019	0.1.0	no
scPred[73]	Discriminative	R	Radial kernel SVM with dimensionality reduction projection of query to reference	2019	1.9.2	yes
scLearn[131]	Generative	R	Discriminative component analysis to learn models based on similarity then predict query cells	2020	1.0	yes

SciBet[72]	Generative	R	Multinomial generative models with maximum likelihood estimations	2020	1.0	no
ACTINN[71]	Discriminative/ Neural Net	python	Three layered neural net with 1 hidden layer	2020	GitHub version: 563bcc1	no
NeuCA Large[89]	Discriminative/ Neural Net	R	Five layered neural net with 3 hidden layers	2022	1.0.0	no
NeuCA Medium [89]	Discriminative/ Neural Net	R	Four layered neural net with 2 hidden layers	2022	1.00	no
NeuCA Small[89]	Discriminative/ Neural Net	R	Three layered neural net with 1 hidden layer	2022	1.0.0	no
sclAE[92]	Discriminative/ Neural Net	R	Integrative autoencoders ensemble methods to extract features in presence of noise	2022	Github Commit: Fdac0fa	no
scClassify[132]	Hierarchical	R	Hierarchical ordered partitioning and collapsing hybrid (HOPACH), train weighted KNN classifiers at each node in the tree	2020	1.2.0	yes
scHPL[83]	Hierarchical	python	Classification tree using one class SVM at each node except root on the most informative pc variables	2021	0.0.2	yes
CHETAH[86]	Hierarchical	R	Correlation similarity under the constraint of respecting tree topology	2019	1.6.0	yes
SingleR[78]	Correlation	R	Spearman correlation in the variable gene expression feature space to annotate query cells	2019	1.4.1	no
Spearman Correlation	Correlation	R	Maximum spearman correlation scores between a query cell and a training reference	____	No version	no
Scmap cell[79]	Correlation/ Discriminative	R	KNN classifier that uses Spearman-cosine- and Pearson similarity measures to annotate cells	2018	1.12.0	yes
Scmap cluster[79]	Correlation/ Discriminative	R	Nearest median classifier that uses Spearman-cosine- and Pearson similarity measures to annotate cells	2018	1.12.0	yes

Table 2. Parameters used for the selected Classifiers

Name	Mode of usage and parameters selected
Support Vector Machine (SVM)	Run with a linear kernel using the scikit-learn package.
SVM Rejection	Run with a linear kernel using scikit-learn. Cells with 0.7 or less classification probability were rejected,
SingleCellNet	Run with default parameters. A random forest of 500 trees was trained using top 25 cell type-specific genes identified by the classifier.
scPred	Run with default parameters using scPred package; No parameters were provided to the functions. Default rejection threshold of low probability classifications was set at 0.6.
scLearn	Run using default parameters: Training was performed with 10 bootstrap repeats and default rejection threshold set at 0.6
SciBet	Run with default parameters: number of training genes selected by the classifiers was 1000.
ACTINN	Run using default parameters: number of training epochs was 50, minibatch size was 128 cells, and learning rate was 0.0001
NeuCA Large	Run with three hidden layers (parameter model.size = 'large') using NeuCA package.
NeuCA Medium	Run with two hidden layers (parameter model.size = 'medium) using NeuCA package.
NeuCA Small	Run with one hidden layers (parameter model.size = 'small) using NeuCA package.
scIAE	Run using scIAE package with no parameters provided to the function
scClassify	Run with default parameters: algorithm detailed in table 1. Cells assigned by the classifier as 'node' were rejected and labeled as 'unsure'.
scHPL	Run with default parameters. SVM was used as the underlying classifier and cells with a node label were rejected and labeled as 'unsure'.
CHETAH	Run using default parameters: Spearman correlation was calculated using top 200 genes selected by the classifier.
SingleR	Run with default parameters using SingleR package.
Spearman Correlation	Run using Spearman correlation of gene expression between reference and query.
Scmap cell	Run using default parameters. 500 genes were selected for training and rejection threshold was set 0.7
Scmap cluster	Run using default parameters. 500 genes were selected for training and rejection threshold was set at 0.7.

2.2. Cross validation

Ten-fold cross-validation was performed to assess the performance of annotation tools. For this, each dataset was divided into ten partitions (folds): nine were used for training and one for testing[133]. This partitioning was done in a stratified manner (for each cell type, cells were selected randomly for each fold, maintaining the cell proportions of the original dataset) to ensure that all cell types in each fold were represented as the same proportion

in the original dataset. The training and testing folds were defined once and kept constant for all classifiers.

2.3. Performance evaluation metrics for cross-validation

2.3.1. F1-score

The F1-score was used as a performance metric for cross-validation as in previous benchmarking studies[100,101]. F1 is defined as the harmonic mean of precision and recall. Precision **(Equation 4)** quantifies the proportion of true positive predictions out of all positive predictions:

$$precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (Equation\ 4)$$

In turn, recall quantifies the proportion of true positive predictions made out of all positive examples in a dataset:

$$recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (Equation\ 5)$$

F1 aggregates both metrics into one score:

$$F1 = 2\,.\frac{precision\,.\,recall}{precision + recall} \quad (Equation\ 6)$$

Here, we reported F1 per cell type and we used the median F1 score across all cell types as the measure of performance.

2.3.2. Rejection Rate

Eight cell type classifiers (Table 1) can reject cell labels with low classification probability according to set rejection thresholds (Table 2). Rejected cells were excluded from the F1

calculations and were reported as the rejection rate as a complementary measure of classifier performance **(Equation 7)**.

$$Rejection\ rate = \frac{number\ of\ cells\ left\ unannotated\ by\ the\ classifier}{total\ number\ of\ cells} \quad (Equation\ 7)$$

2.4. Dataset complexity metrics

We calculated three metrics to measure complexity of scRNA-seq datasets: Fraction of borderline cells (N1) [134,135], volume of the overlapping region (F2) [134,135], and proportion of clusters belonging to a hierarchical structure (H1) (see below, sections 2.4.1-3). N1 and F2 were adapted from Valvidia, 2019 on GitHub (https://github.com/jcelias98/Complexity-Measures), with code optimization for increased computational efficiency. The optimization included parallelizing the code to run faster on large datasets.

2.4.1. Fraction of borderline cells (N1 metric)

We used the Fraction of Borderline Points Between Classes[134,135] to quantify cell type separability. This metric is a neighborhood approach that calculates the proportion of cells lying in the boundaries between cell types. For computing N1, a minimum spanning tree (MST) is built using the top 2000 variable genes and the Euclidean distance between each pair of cells in the dataset. N1 is then defined as the percentage of cells connected to cells from other cell types in the MST:

$$N1 = \frac{1}{N}\sum_{i=1}^{n}\left(\left(x_i, x_j\right) \in MST \ \ \Lambda\ y_i \neq y_j\right) \ , (Equation\ 8)$$

Where N is the total number of cells, x_i is the current cell, x_j is the neighbor of x_i in the MST, y_i is the cell type of the current cell, and y_j is the cell type of the neighboring cell.

This metric is bounded between [0,1], with higher values indicating more complex boundaries between cell types, showing the overlap and continuity between the classes and poor separability.

2.4.2. Volume of overlapping region (F2 metric)

F2 computes the overlap in the expression of genes among the cell types present[134,135]. This metric, designed for binary classification problems, can be adapted for multi-class problems using one-vs-one (OVO) decomposition, in which the overlap is calculated between each possible pair of cell types in the dataset. This overlap is then averaged across all possible cell type pairs. Higher values of F2 indicate a higher redundancy and overlapping of gene expression programs between cell types. To compute this metric, we first performed principal component analysis (PCA) on the gene expression matrix and retained the first 50 principal components as features. F2 is defined as:

$$F2 = \prod_1^m \frac{Overlap(f_i)}{Range(f_i)} = \prod_1^m \frac{\max\{0, minmax(f_i) - maxmin(f_i)\}}{maxmax(f_i) - minmin(f_i)}, \textit{ (Equation 9)}$$

Where:

$$minmax(f_i) = \min\left(\max\left(f_i^{c_1}\right), \max\left(f_i^{c_2}\right)\right)$$
$$maxmin(f_i) = \max\left(\min\left(f_i^{c_1}\right), \min\left(f_i^{c_2}\right)\right)$$
$$maxmax(f_i) = \max\left(\max\left(f_i^{c_1}\right), \max\left(f_i^{c_2}\right)\right)$$
$$minmin(f_i) = \min\left(\min\left(f_i^{c_1}\right), \min\left(f_i^{c_2}\right)\right)$$

m is the number of features, and f_i is the current feature. The volume returned by F2 is bounded between 0 and 1.

2.4.3. The proportion of cell types belonging to a hierarchical structure (H1 metric)

To quantify the inherent hierarchy of a dataset, we calculated H1, which quantifies the fraction of cell populations within a dataset that belong to robustly inferred hierarchical groups (i.e. whether a large proportion of the dataset presented hierarchical structure).

For this, we first applied bootstrapped hierarchical clustering to the entire dataset (pvclust)[136] with 100 bootstrap iterations. Next, for every leaf node c, we asked whether it belonged to a subtree with at least three robustly inferred ancestor nodes (bootstrap value > 80%); if so, this leaf is labeled as c_h. Finally, H1 is defined as:

$$H1 = \frac{N_{ch}}{N}\ , \ (Equation\ 10)$$

where N_{ch} is the number of ch clusters, and N is the total number of clusters in the dataset. Higher values of H_1 indicate that a larger fraction of datasets belong to a robustly inferred hierarchical structure.

2.5. SNAKEMAKE benchmarking and annotation pipelines

Snakemake is a workflow engine and management tool that provides a Python interpretable framework for executing pipelines based on user-defined rules[137]. We implemented two Snakemake pipelines: one for the benchmarking experiments and one for annotation of samples using a reference.

2.5.1. Classifier benchmarking pipeline

This is a benchmarking pipeline, adapted from a workflow written by Abdelaal et al., 2019[100], but modified with added scripts to: include more cell type prediction tools not available at the time of the paper publication (NeuCA, correlation, scIAE, and scClassify), include quantitative metrics of dataset complexity metrics **(section 2.4)**, and output UMAP plots. The scripts used to implement this pipeline can be found at https://github.com/HusseinLakkis01/scCoAnnotate/tree/main/Benchmarking

2.5.2. scCoAnnotate: a consensus-based annotation of query datasets

This pipeline trains user-selected cell type classifiers on an annotated training reference, then uses these trained classifiers to annotate new query datasets. This pipeline produces an output of consensus predictions for which individual annotations are used to choose a label using a majority vote (the predicted target label of the ensemble is the

mode of the distribution of individually predicted labels). In cases where two or more predictions are tied, scCoAnnotate outputs a 'no consensus' label. From the 18 classifiers described in Section 2.1, only 12 were selected for this pipeline based on benchmarking results: ACTINN, SingleR, SciBet, SVM, SVM rejection, scPred, Correlation, scmap-cell, scmap-cluster, SingleCellNet, scHPL, CHETAH. scCoAnnotate can be found at https://github.com/HusseinLakkis01/scCoAnnotate.

2.6. Datasets:

We compiled a set of 23 scRNA-seq datasets (**Table 3**) spanning different tissues, developmental ages, species, number of cells, and number of cell type populations to assess ML tools. We downloaded pancreatic datasets from https://hemberg-lab.github.io/scRNA.seq.datasets (Baron Human: GSE84133, Baron Mouse: GSE84133, Muraro: GSE85241). Brain datasets can be found in the Gene Expression Omnibus (GEO) (Jessa 2019: GSE133531, Jessa 2022: GSE188625, Vladoiu: GSE118068, Human Fetal Atlas) or are downloaded from external sources as follows. The Allen Brain datasets were downloaded from the Allen Institute Brain portal https://portal.brain-map.org/atlases-and-data/rnaseq (The primary visual cortex (V1 & ALM): GSE115746, adult human cortex, adult mouse cortex: GSE185862). The Anderson (GSE125290), Dong (GSE137804), and Mizrak (GSE134918) datasets were downloaded from the GEO. For the organism-spanning atlases, we downloaded the Tabula Muris from https://tabula-muris.ds.czbiohub.org (GSE109774) and the Pijuan-Sala from https://github.com/MarioniLab/EmbryoTimecourse2018. For the cancer datasets, we downloaded the CellBench dataset (GSM3618014) and used high-grade glioma tumor data (HGG Dataset) available at the European Genome-phenome Archive (EGA) under accession number EGAS00001005773.

We used the counts matrices of these datasets as they were processed by the original authors. However, we down-sampled the Tabula Muris, the adult human and mouse cortical datasets, Jessa 2022, and the human fetal brain due to computational restrictions (resulting sizes highlighted in table 3). We performed down-sampling in a stratified manner where we kept 50% of each cell type population. The Jessa 2019 mouse brain

atlas contains different levels of labels varying in granularity (from time point level labels such as “E18 pontine astrocytes” to general cell class labels such as “astrocytes”). We used different subsets of this dataset (pons and cortex) as independent datasets. In all cases in this thesis, cell type annotations, as provided by the authors, were used as a ground truth.

Table 3. Summary of the datasets used with the metadata and the chapters it was used in. Dataset type indicates whether the dataset profiles strictly prenatal tissues, postnatal tissues, or both (Mixed). E corresponds to embryonic day and P corresponds to postnatal day. Number of cells highlighted in bold are the final sizes for the down-sampled datasets.

Dataset	Species	Organ	Type	No of cells	No of labels (cell types)	Protocol	Chapter Used
Anderson[138]	Mouse	Brain Striatum	Mixed	14466	40	10X	Chapter III
Jessa 2019 Pons [33]	Mouse	Pons	Mixed	25978	92	10X Chromium	Chapter III
Jessa 2019 [33]	Mouse	Brain	Mixed	58153	14	10X Chromium	Chapter III
Jessa 2022[34]	Mouse	Brain	Mixed	**63936**	290	10X Chromium	Chapter IV
Vladoiu [121]	Mouse	Cerebellum	Mixed	62040	34	10X Chromium	Chapter III
Jessa 2022[34]	Mouse	Cortex E10	Prenatal	3934	14	10X Chromium	Chapter III
Jessa 2019 [33]	Mouse	Cortex E12	Prenatal	8578	14	10X Chromium	Chapter III
Jessa 2019 [33]	Mouse	Cortex E16	Prenatal	7183	21	10X Chromium	Chapter III
Jessa 2019 [33]	Mouse	Cortex P0	Postnatal	4689	20	10X Chromium	Chapter III
Jessa 2019[33]	Mouse	Cortex P6	Postnatal	3886	13	10X Chromium	Chapter III
Allen Mouse Brain 2018[139]	Mouse	Primary Visual Cortex	Postnatal	12832	16	SMART-Seq v4	Chapter III
Baron [105]	Human	Pancreas	Postnatal	8569	14	inDrop	Chapter III/IV
Baron [105]	Mouse	Pancreas	Postnatal	1887	13	inDrop	Chapter IV

Tabula Muris [105]	Mouse	20 organs	Postnatal	**54865**	55	SMART-Seq2	Chapter III
CellBench [107]	Human	Adenocarcinoma	Postnatal	3803	5	10X Chromium	Chapter III
Muraro [106]	Human	Pancreas	Postnatal	2122	8	Cel-seq2	Chapter III
Mizrak [140]	Mouse	Brain	Postnatal	28407	11	SCOPE-seq	Chapter III
Dong [141]	Mouse	Brain	Prenatal	14229	15	GemCode	Chapter III
Pijuan-Sala [142]	Mouse	Whole Organism	Prenatal	18140	37	10X Chromium	Chapter III
Human Fetal Brain [143,144]	Human	Brain Thalamus	Prenatal	**43821**	213	10X Chromium	Chapter IV
HGG[34]	Human	Brain	Postnatal	176934	-	10X Chromium	Chapter IV
Allen Human Brain 2019 [145]	Human	Cortex	Postnatal	**23714**	9	SMART-Seq2	Chapter IV
Allen Mouse Brain 2019[146]	Mouse	Cortex and Hippocampus	Postnatal	**37484**	8	SMART-Seq2	Chapter IV

3. Chapter III: Classifier benchmarking using cross-validation

Single-cell RNA sequencing (scRNA-seq) has been a proven tool for identifying and characterizing cell types in various biological systems. However, the accuracy of cell type classification remains a critical issue, particularly in prenatal and developmental scRNA-seq datasets, where the cell type information is often limited or unknown. In this chapter, we present a benchmarking of cell type classifiers using cross-validation and dataset complexity metrics (figure 1). Our goal is to evaluate the performance of different tools and understand the impact of dataset complexity on their performance.

We use cross-validation, a widely used technique to evaluate and compare the performance of different classifiers, to benchmark the performance of several state-of-the-art cell type classifiers. Additionally, we quantify the complexity of each scRNA-seq dataset using metrics that capture various aspects of dataset complexity, including cell type continuity, gene expression redundancy, and hierarchical cell type relationships. We aim to provide a comprehensive understanding of the relationship between dataset complexity and classifier performance and to identify the classifier that performs best in the presence of high levels of dataset complexity.

This chapter will provide insights into the challenges and limitations of cell type classification in the context of developmental studies, aiming to inform the choice of classifier for future studies. By combining cross-validation and dataset complexity metrics, we objectively and systematically evaluate cell type classifiers. Hopefully, this will aid in advancing our understanding of the underlying biological mechanisms that drive cellular heterogeneity in developmental research.

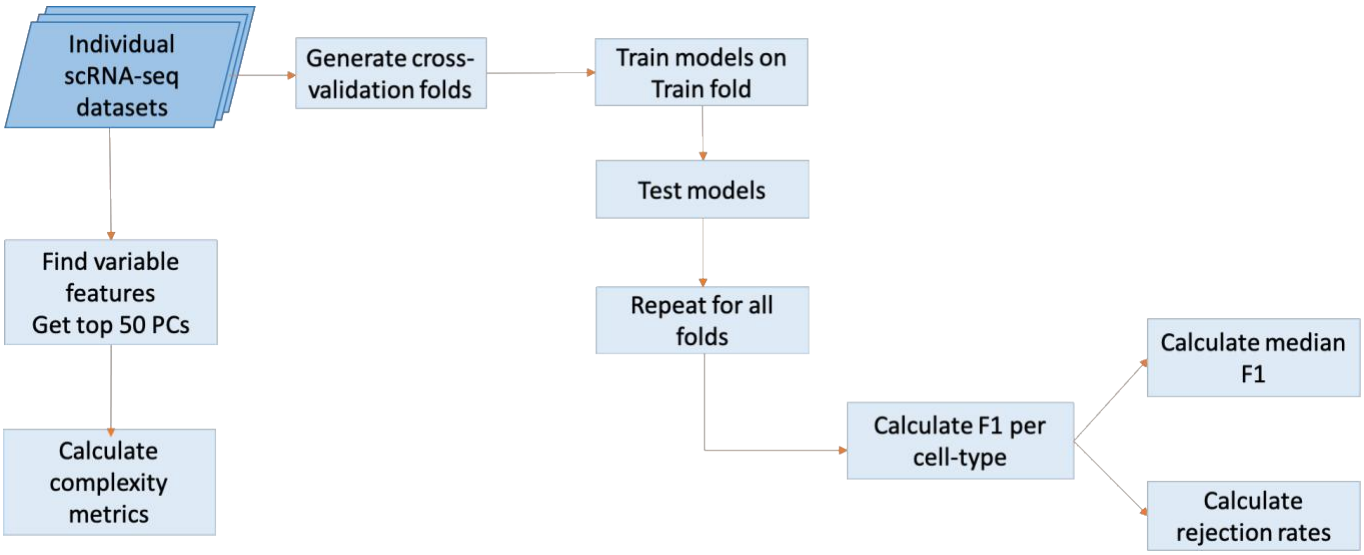


Figure 1 Overview of the benchmarking experiments done to evaluate classifiers in sections 3.3, 3,4, and 3.5.

3.1. Validation of complexity metrics

To evaluate whether prenatal datasets possess specific characteristics that hamper automated cell type annotation tasks, we implemented three complexity metrics to quantify redundancy, continuity, and hierarchical structure (sections II-2.4). Next, we performed a validation study of these metrics in a simulated setting to confirm if they could capture the observed biological complexity.

To begin these analyses, we created four synthetic datasets from the Jessa 2019 mouse brain dataset[33] (Figure 2). Each dataset contains clusters of cells with defined transcriptional overlap that we intend to measure. Dataset *a*, independent cell types, is comprised of five different cell types: astrocytes, microglia, oligodendrocyte precursor cells (OPCs), ependymal cells, and neuro-progenitors. These cell types are transcriptionally distinct and have different biological functions. Dataset *b*, progenitors and differentiating cells, includes 16 different glial and neuronal progenitor cell types. These cell types are related to each other and have some common transcriptional programs. Dataset *c*, astrocytes, contains 16 astrocyte cell types from different brain regions (pons and cortex) collected at three postnatal time points (at birth (P0) and postnatal days 3 (P3) and 6 (P6)). These clusters contain only mature cells. They have highly overlapping

gene programs but express distinct signatures that allow them to be classified as separate cell types. Lastly, dataset *d*, oligodendrocyte lineage, is comprised of 14 cell types from one defined lineage. These cell types range from early OPCs to mature oligodendrocytes. OPCs and oligodendrocytes from both the pons and cortex collected at P0, P3 or P6 were included. These cell types are redundant in their gene expression programs but less so than dataset *c*, which contains only mature cell types. Using these synthetic datasets, we assessed the ability of the metrics N1, F2, and H1 (Section II: Equations 8-10) to detect cell type separability, gene expression redundancy, and hierarchical structures, respectively (Figure 3).

The first measure, N1, calculates the proportion of cells lying in the boundaries between cell types (cell type separability). As the value for N1 increases, the cell types are less separated. N1 was calculated for each of the generated synthetic datasets (Figure 3a). Dataset *a* (independent cell types) showed the lowest N1 value (2.9%), In contrast, dataset *d* (oligodendrocyte lineages) had a much higher N1 value (63%). As expected, the more closely related cell types exhibited a decrease in class separability (N1). This effect is a direct result of the cell types being more transcriptionally similar. Dataset *a* contains disjointed and discrete cell types which do not have complex cell type borders with other cells. Dataset *b* shows a higher N1 value (28%) than dataset *a* due to the inclusion of progenitors from multiple lineages expressing similar gene programs. These progenitors are also related which decreases cell type separability, thus increasing N1. On the other hand, dataset *d* depicts a full differentiation lineage where multiple cell types are differentiating into other mature cell types, forming a continuum of gene expression profiles and increasing the border complexity. Interestingly, dataset *c*, which contains gene expression profiles from mature astrocytes, displays a lower N1 value than dataset *d*. While the mature astrocytes contain many redundant gene expression profiles, the borders between distinct clusters are more defined than cells within one lineage.

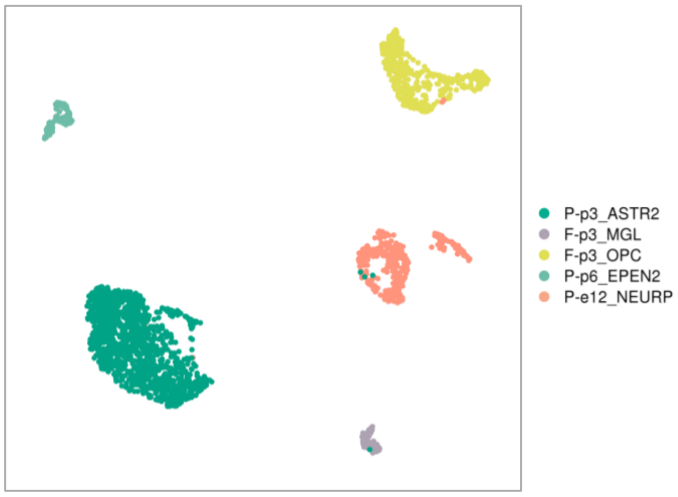
The F2 metric computes the overlap in the expression of genes among the cell types. Similar to N1, the F2 metric shows distinct differences across the four synthetic datasets (Figure 3b). As expected, the dataset with the least similar gene expression programs,

dataset *a*, had the lowest value for the F2 metric (1.3 x 10-10). Dataset *b* showed a higher F2 value (9 x 10-8) than dataset *a* due to the inclusion of different progenitors with similarities in gene expression programs. In contrast, dataset *c*, containing mature astrocytes had a much higher F2 value (7.6 x 10-7). Comparison of these two metrics demonstrates their distinct uses for describing datasets. While dataset *c* (mature astrocytes) contains redundant gene expression programs, the gene expression profiles targeting the cells for distinct clusters allow for the establishment of clear boundaries between different cell types. In contrast dataset *d*, which contains cells at various stages of differentiation, the boundaries between cells are less defined despite a lower redundancy in gene expression.

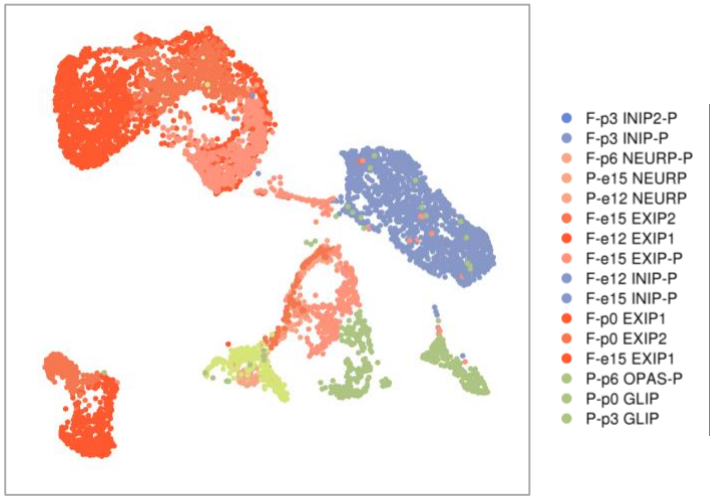
We designed the H1 metric to detect cell type hierarchical structure in scRNA-seq datasets comprised of differentiation lineages. However, we did not detect large differences in the calculated H1 values across our datasets. Dataset *a* and c (independent cell types and mature astrocytes, respectively) should have a low H1 value when compared to lineage datasets *b* and *d*. We observed an H1 value of 0.2 in dataset *a* and 0.1 in dataset *c*. Datasets *b* and *d*, both displayed H1 values of 0.25 (Figure 3c). These similar H1 scores across datasets could be due to the poor clustering of highly similar cell types belonging to the same lineage, which can result in poor hierarchical clustering in datasets *b* and *d*. Nevertheless, H1 detects a larger hierarchical structure in datasets with progenitors (*b* and *d*) compared to other datasets.

Altogether, these results demonstrate that dataset complexity can be measured using quantitative metrics such as N1 and F2. These metrics increase as cell type separability decreases and gene expression redundancy increases, as expected. Furthermore, our data demonstrates the distinct manner that N1 and F2 capture and quantify complexity in datasets. N1 captures the cellular continuity that occurs as differentiating cells take on intermediate phenotypes that reside at the border of a particular cell type. In contrast, F2 captures expression redundancy where cells undergoing development express similar programs and cannot be classified as a distinct class. Lastly, H1 did not strongly

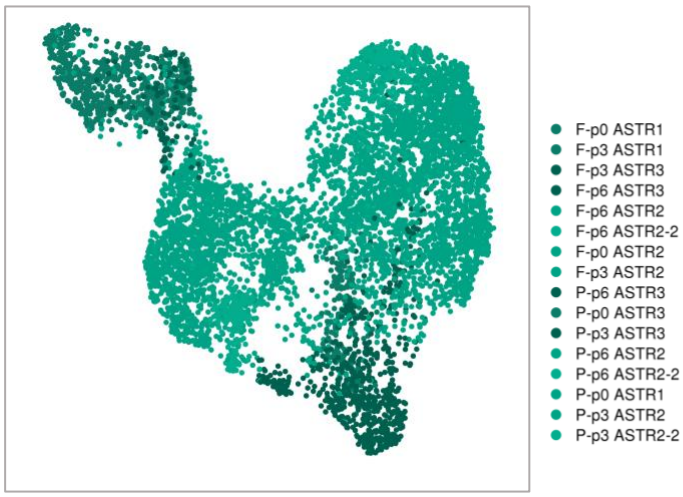
represent cellular hierarchy in our initial analyses. However, it could still detect some hierarchical structure present in the synthetic datasets.



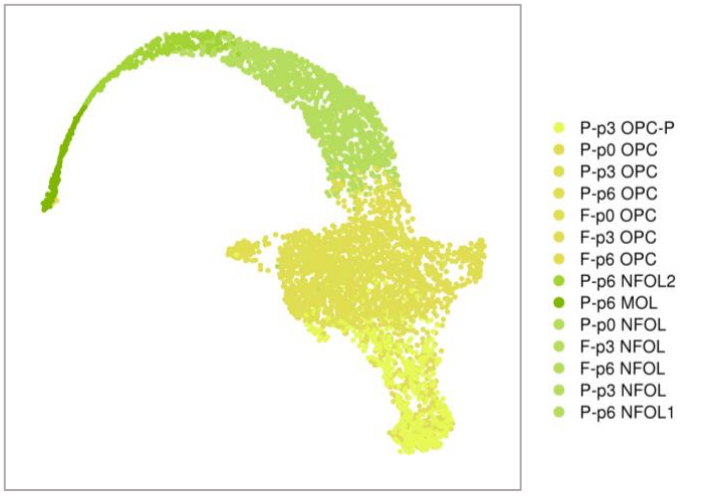
a. Independent cell types



b. Progenitors and related cell types



c. Astrocytes



d. Oligodendrocyte lineage

Figure 2 Four synthetic datasets generated from the Jessa 2019 mouse brain data. Independent cell types (a) progenitors and related cell types (b), mature astrocytes (c), cells of the oligodendrocyte lineage (d).

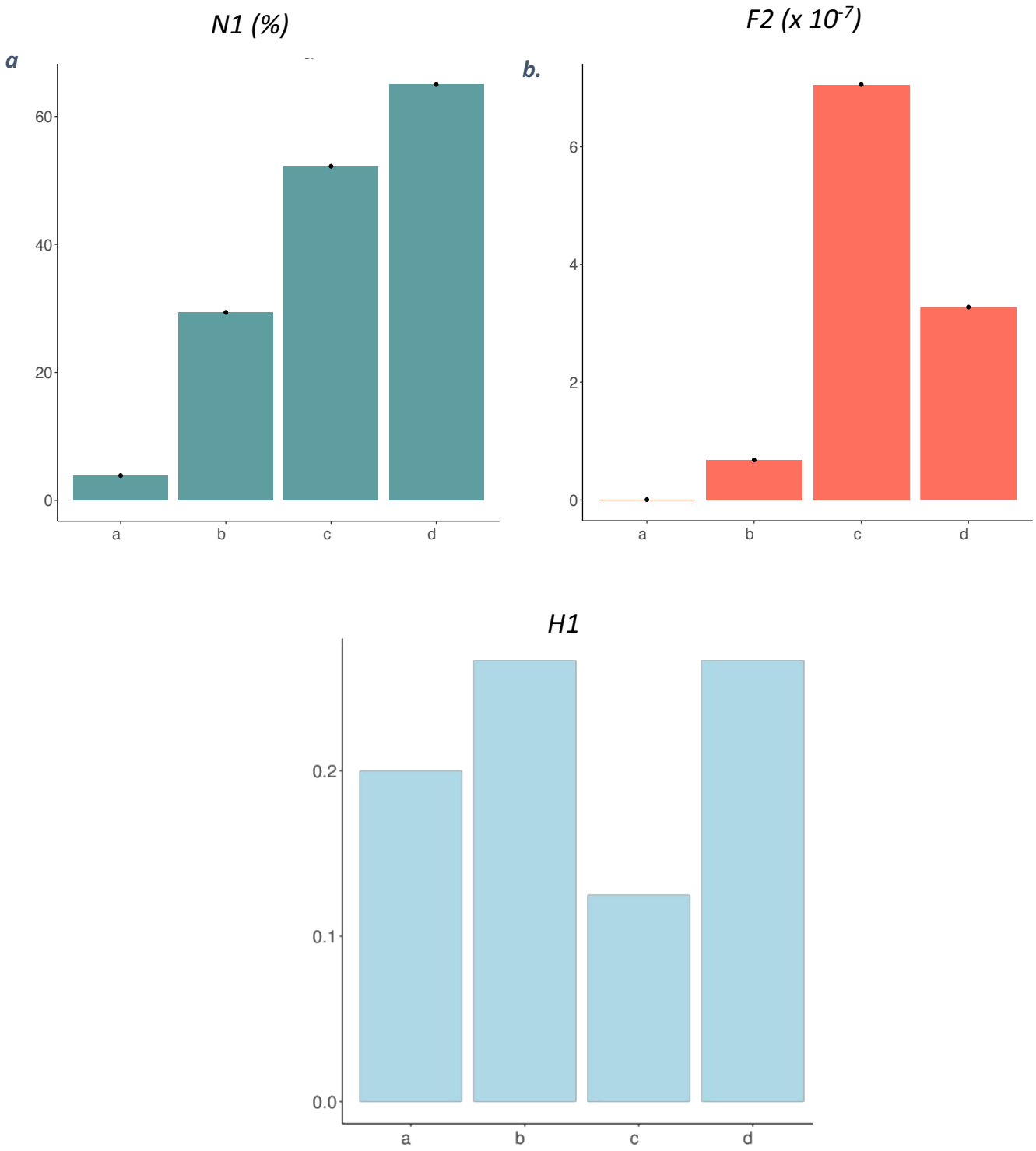


Figure 3 Bar plots showing the N1 (a), F2 (b), and H1 (c) in the four synthetic datasets. Each bar in the plots corresponds to a dataset depicted in Figure 2.

3.2. Quantitative assessment of dataset complexity in different developmental contexts

Following the validation of the three-complexity metrics in synthetic datasets, we expanded our assessment of dataset complexity by calculating N1, F2 and H1 for each dataset described in Table 3 (section II-2.6). These datasets include prenatal datasets containing only cells from embryonic stages, postnatal datasets containing cells from postnatal tissues, and mixed datasets which contain data from both prenatal and postnatal tissues.

We first measured cell separability by calculating N1 for each dataset (Figure 4). We observed an increased proportion of borderline cells (>0.2 or more than 20% of cells in the dataset) in all prenatal and mixed datasets, with the exception of one dataset (Jessa Broad Labels). In contrast to other prenatal and mixed datasets, the Jessa Broad dataset contains diverse clusters from different samples that are grouped together into cell classes. This under-clustering leads to poor labeling which may influence the value of the N1 measure. In contrast, postnatal datasets showed low N1 values (on average less than 10%) than prenatal and mixed datasets. Jessa P0 and Jessa P6 show higher N1 values than other postnatal datasets. Importantly, these datasets were sampled on the day of birth (P0) and six days after birth (P6). At these time points, multiple cell types in the mouse brain are still actively undergoing differentiation. These cells in transient states are less likely to be separated into particular cell types and as such may lead to an increase of N1 for these datasets. All other postnatal datasets were collected after P9. We also observed that the CellBench dataset, which we included as a control, had the lowest N1 value of 0.1%. This dataset contains five disjoint lung cancer cell lines that do not share common cell types and gene expression profiles. In this case, we would expect cells to cluster into distinct values and N1 to be very low. Overall, the N1 metric performed as we expected with a decrease in N1 value as the cells move through development and adopt mature cell fates

Next, we examined the genetic redundancy of our datasets by calculating F2 values. These values were log transformed to facilitate the scaling and visualization of results

(Figure 5). Similar to N1, prenatal and mixed datasets generated higher F2 values (average of 14.1 across datasets) than postnatal datasets (average of 7.25). However, one postnatal dataset, Mizrak, deviated from this average (13.1 vs 7.25 on average). This deviation was expected as this dataset represents postnatal neurogenesis and thus contains various cell types undergoing differentiation. As these cells can be found in intermediate states, they also exhibit redundancy in their gene expression profiles. We would expect these factors to inflate F2 values. We also observed that the CellBench dataset shows null F2 scores, as expected.

Finally, we attempted to evaluate the hierarchal structures among datasets by calculating H1 in real datasets that contain multiple lineages (Figure 6). Unlike N1 and F2, H1 did not show any specific trend across prenatal, postnatal and mixed datasets. H1 scores ranged from 0 to 0.72. We first noticed that the postnatal Jessa cortical dataset displayed the highest proportion of hierarchical lineages (H1=0.72). The Jessa cortical dataset is derived from mouse brain at P0. Three datasets showed no hierarchical structure at all as per H1 calculation (H1=0). While this result may be expected in the unrelated cell types of the CellBench dataset, it was not anticipated in the Mizrak dataset (postnatal neurogenesis). We expected H1 to detect a hierarchical structure in the Mizrak dataset which profiles neuronal lineages in the brain. One explanation for this discrepancy is that the conditions we used to generate H1 scores were stringent conditions (>80 bootstrap p-value). Under these conditions, some hierarchical structures may have been inadvertently excluded.

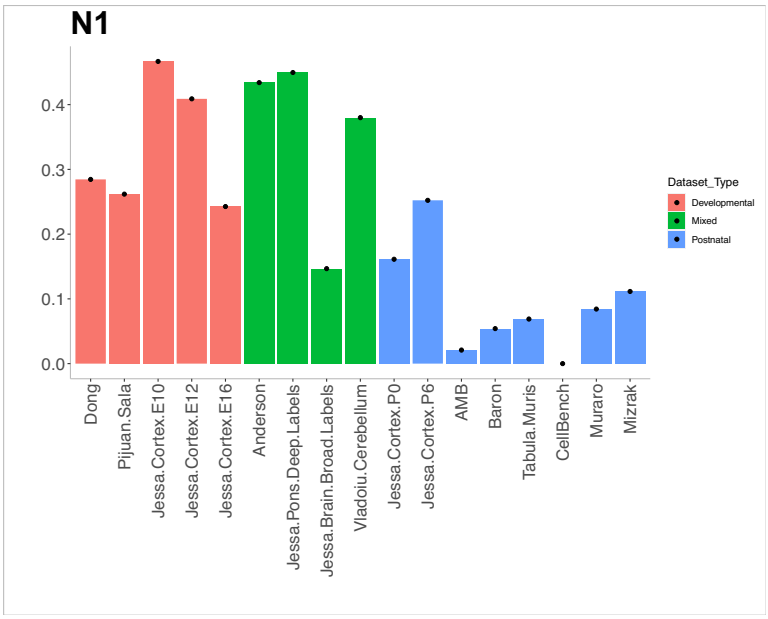


Figure 4 Bar plot showing the fraction of borderline cells (N1) in datasets included in benchmarking

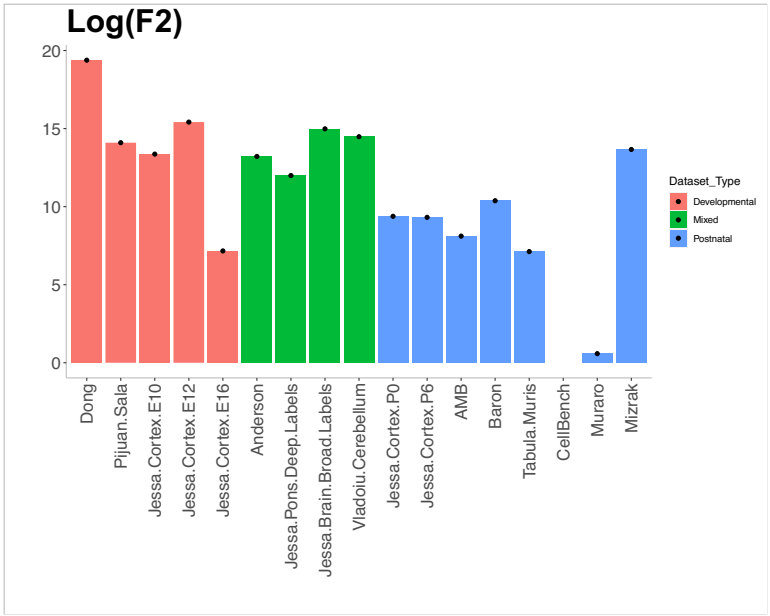


Figure 5 Bar plot showing the volume of overlapping regions (F2) on datasets included in benchmarking

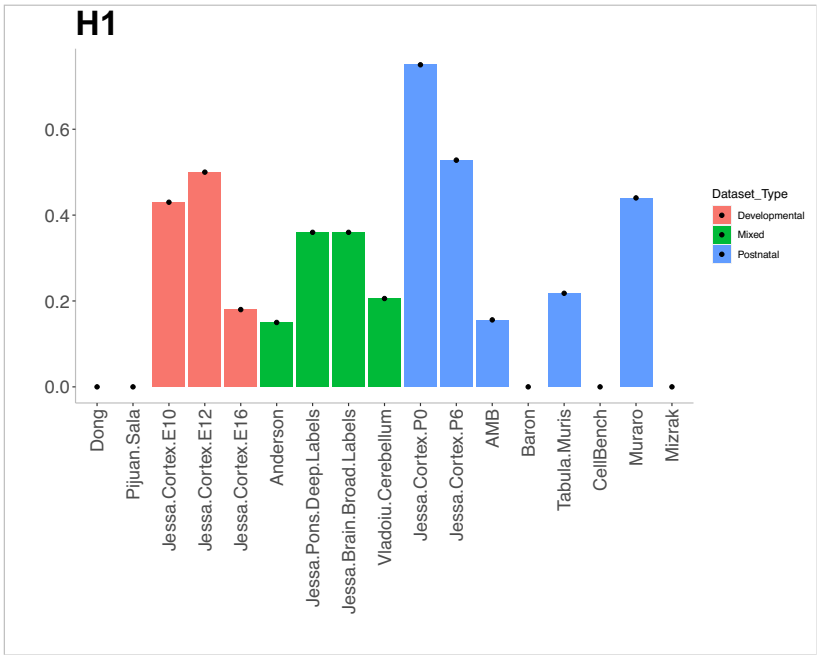


Figure 6 Bar plot showing the proportion of cell types belonging to a hierarchical structure (H1) on datasets selected for benchmarking

3.3. Timepoint benchmarking of classifiers on mouse cortical datasets

After determining the complexity of our datasets, we sought to evaluate the performance of classifiers in a setting where cross-validation performance can be examined as a function of the developmental stage of the data. We evaluated 17 classifiers (Section II – Table 1) using five datasets that profile the murine cortex at three embryonic time points (E10, E12, and E16) and two postnatal time points (P0 and P6), as described in Table 3 (Section II). We reported the performance of each classifier using F1 scores and rejection rates (Section II-2.3). The F1 score is a measure of precision and recall, while the rejection rate reflects the inability of a classifier to annotate a cell.

F1 scores for each classifier across the five datasets are presented in a heatmap and a line graph format (Figures 7 and 8). A high F1 score indicates a well-performing classifier. All classifiers had lower F1 scores in prenatal datasets than the postnatal time points. Specifically, all classifiers except scmap, ACTINN, and scLearn show their lowest F1 scores in the earliest time point (E10). In all cases, each classifier displayed its highest performance score when analyzing the dataset with the latest timepoint at P6 (Figure 7). However, we observed that.NeuCA_small and scmapcell showed near 0 F1 scores in two prenatal timepoints (0.05 for scmapcell at E12 and 0.04 for NeuCA_small at E16). This indicates that these classifiers failed in these datasets. Overall, we saw that discriminative models outperform all other classes across the different timepoints; SVM rejection achieved perfect accuracy (F1 = 1) at P6. On the other hand, we see that NeuCA performed poorly in these experiments. This is perhaps due to the large dataset required for properly optimizing neural network classifiers. Other tools, such as CHETAH and SingleR, were unsuitable for annotating prenatal datasets.

While the F1 metric measures the precision with which a classifier annotates, the rejection rate examines the inability of a classifier to annotate a particular cell in a dataset. A low rejection rate indicates a better performance by the classifier as it signifies that more cells in the dataset are being annotated. Rejection rates for most classifiers were higher in prenatal than postnatal datasets (Figure 9). Intriguingly, as we observed with the F1

score, scmap performed differently from the other classifiers. Although scmap cell shows good F1 scores at E10 and E16, the rejection rate is higher than 40%. This shows us that scmapcell refrains from annotating most prenatal cells, which indicates a lack of confidence in these predictions. This poor performance might be because scmapcell assumes a linear relationship between gene expression values and cell type which might not hold true in complex systems such as prenatal datasets. Additionally, specific tools (SVM_rejection and scHPL) show a linear decrease in rejection rates. These results highlight the higher uncertainty classifiers have for calling a prediction in datasets where cells are highly similar and undifferentiated. This uncertainty decreases as cell types become more separate and distinct. We also note that discriminative and hierarchical classifiers show the lowest rejection rates compared to other classes.

In summary, we measured the performance of 18 classifiers using F1 scores and rejection rates. Our data indicate that all classifiers perform worse when annotating prenatal datasets. Cell populations become more discrete and distinct as they undergo differentiation and maturation. Consequently, cell type classifiers perform better in these cases as it is easier to distinguish one cell type from another. This ability of the classifiers is reflected by the higher F1 scores and lower rejection rates obtained when tested in postnatal tissues. We also conclude that discriminative classifiers perform best in prenatal single-cell data due to their ability to model the differences between classes based on their distinct features. This is particularly useful in prenatal single-cell data where the data is highly variable, and the class structure needs to be better defined. Furthermore, discriminative classifiers can incorporate prior knowledge and utilize regularization techniques to overcome the challenge of limited sample size and high noise levels. These properties make discriminative classifiers well-suited to prenatal single-cell data and explain why they often outperform other classifiers in these applications.

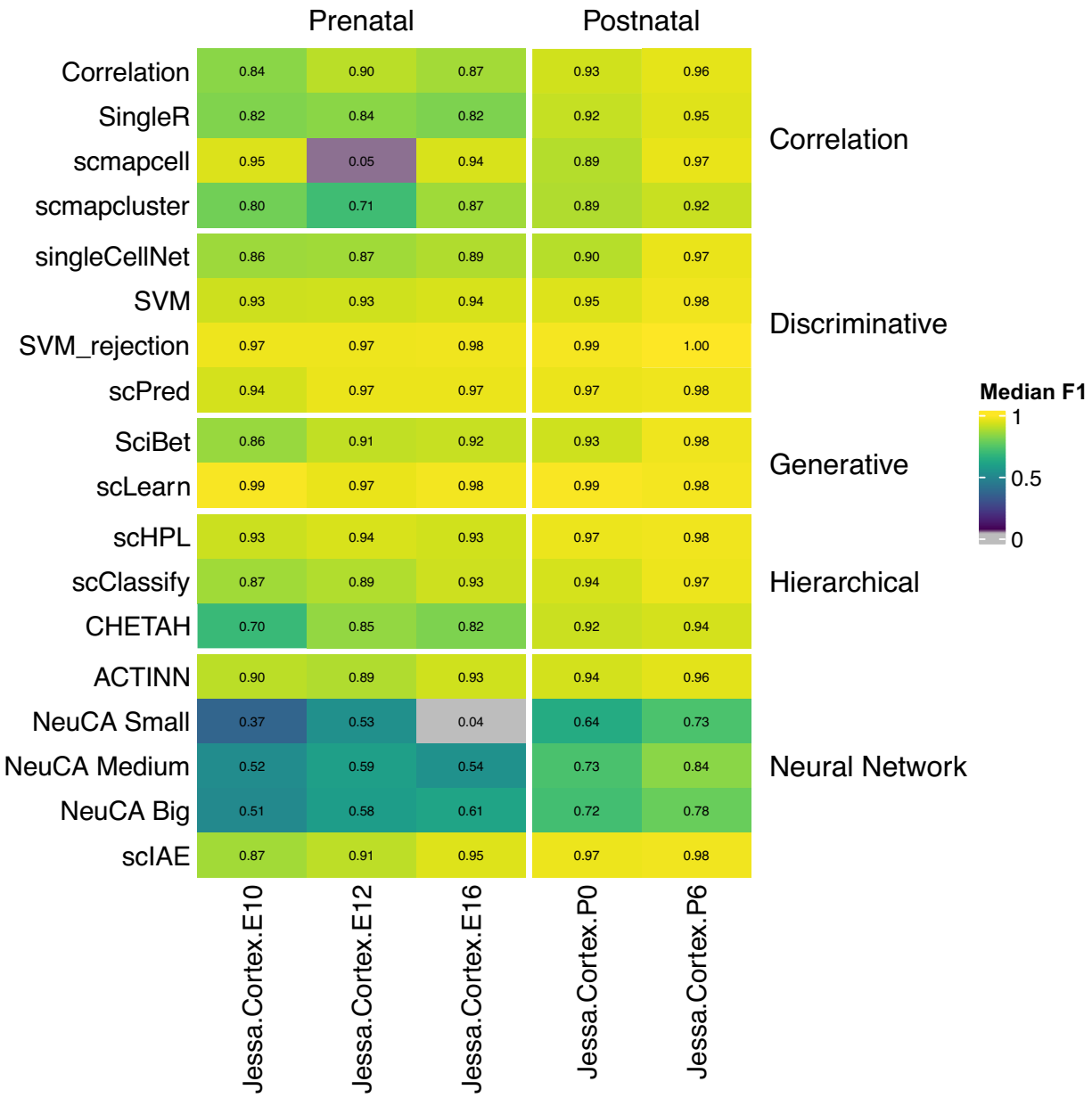


Figure 7 Performance of classifiers on mouse cortex datasets from five sequential timepoints (E10: embryonic day 10 to P6: Postnatal day 6). Heatmap of median F1 across all cell populations per classifier (rows) per dataset (column). Datasets segregated into the two classes (prenatal and postnatal)

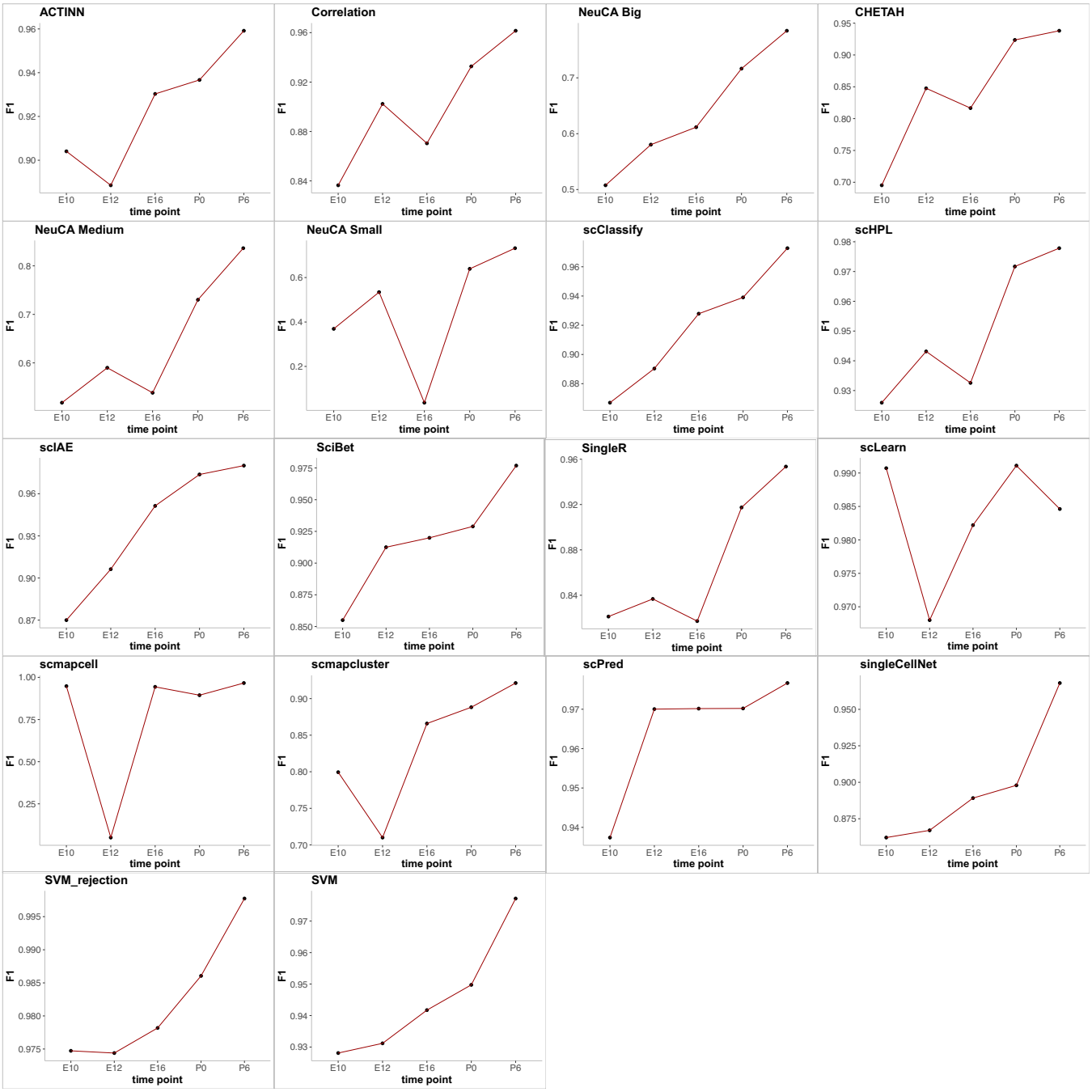


Figure 8 Line plots showing the variation in median F1 score across all cell populations per tool (plot) on the five-time series cortical datasets

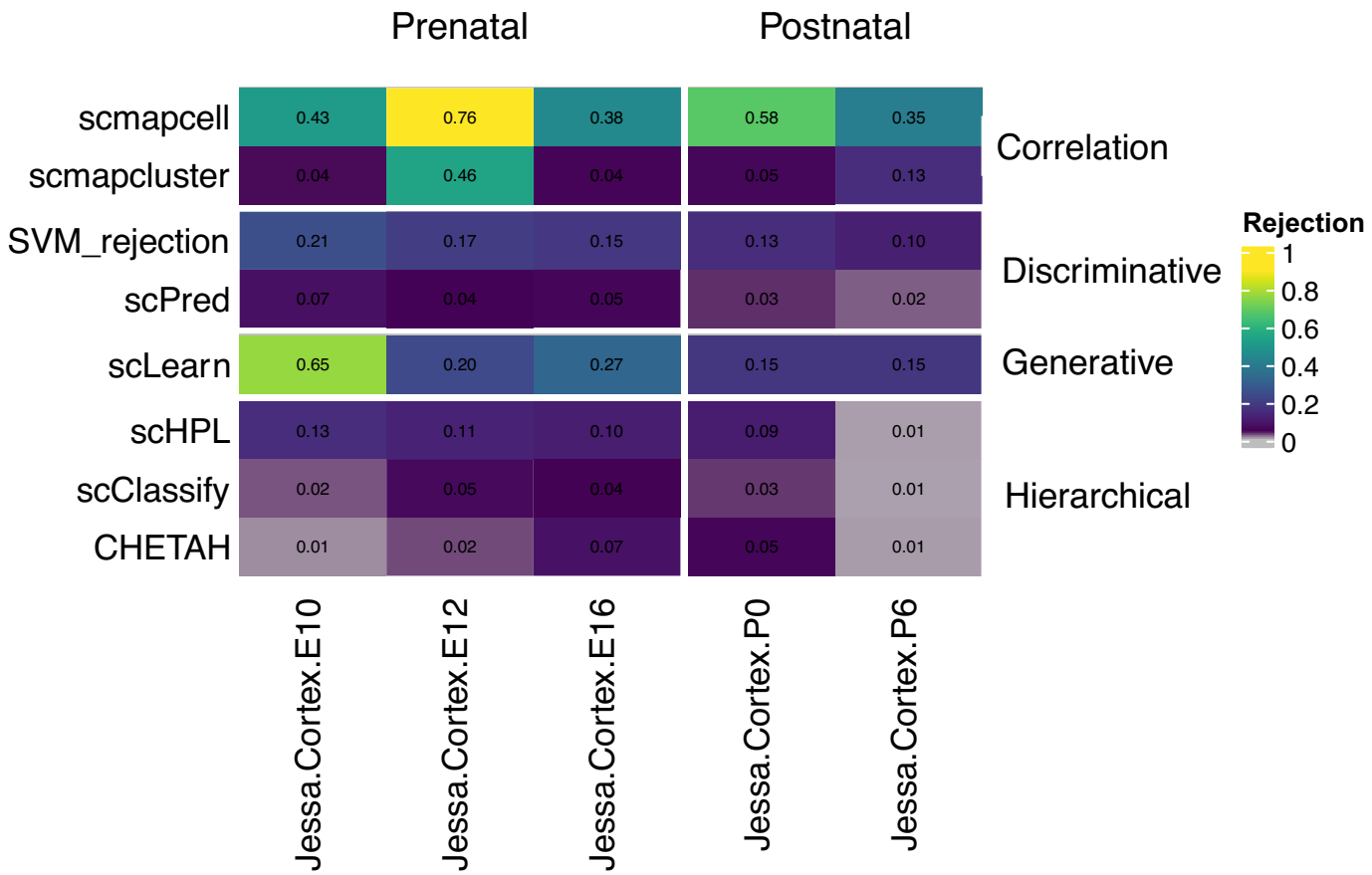


Figure 9 Rejection Rates of classifiers analyzing mouse cortex datasets from five sequential timepoints (E10: embryonic day 10 to P6: Postnatal day 6). Heatmap of proportion of rejected cells across all cell populations per classifier (rows) per dataset (column). Datasets segregated into the two classes (prenatal and postnatal).

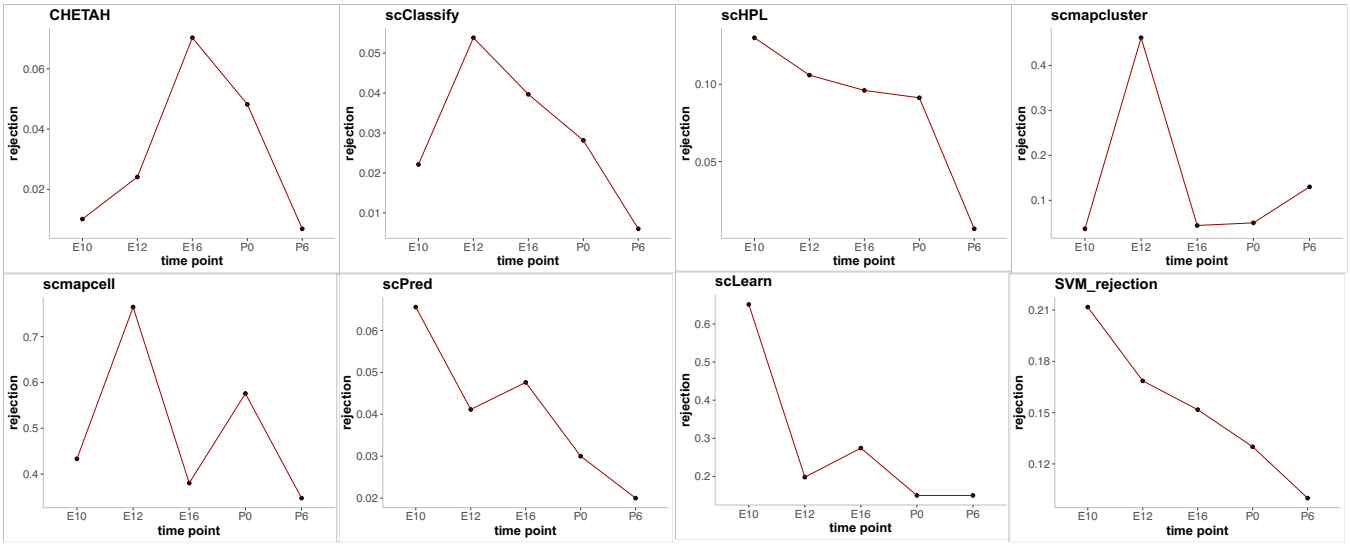


Figure 10 Line plots showing the variation in proportion of rejected cells across all cell populations per tool (plot) on the four-time series cortical datasets

3.4. Performance within a single dataset depends on the cell type and ongoing differentiation

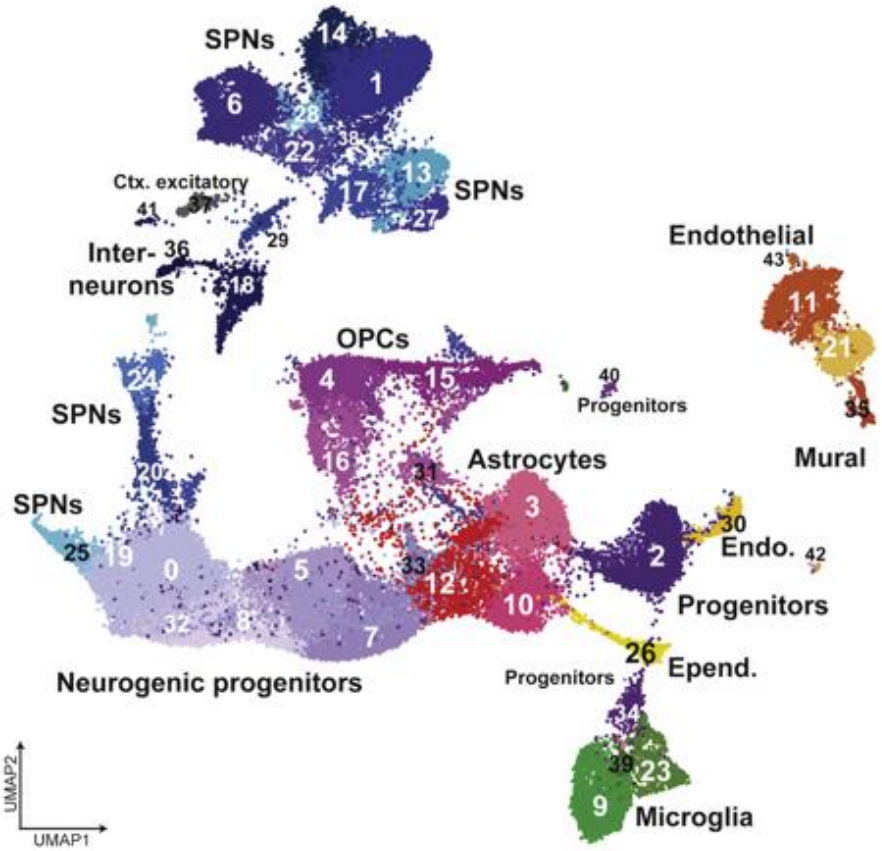
`Our previous data suggests that different cells along the differentiation spectrum have distinct complexity characteristics (Section 3.2). To evaluate whether the performance of classifiers varies according to the inherent cell properties found in progenitor, intermediate and differentiated cell populations, we measured the performance of classifiers within one dataset that contains multiple cell populations. The Anderson dataset is derived from the mouse brain striatum (Section II – Table 3) and is comprised primarily of differentiated cell types. However, this dataset also contains a proportion of progenitor cells and intermediate cells that are undergoing differentiation (Figure 11a).

We calculated the F1 score of for each cell that was annotated by one of 13 classifiers (Section II – Table 1). The F1 scores were plotted as a UMAP to visualize the performance across cell type for each tool (Figure 11b). We observed significantly divergent F1 scores (ranging from 0 to 1) across cell type and across classifiers. Endothelial, interneurons, and committed cell types such as spiny projection neurons (SPNs) show relatively F1

scores (between 0.8 and 1) across most classifiers. In contrast, Neurogenic progenitors, OPCs, and cells with high differentiation potential show lower F1 scores across all classifiers. We also observed that discriminative and generative cell type classifiers such as SVM_rejection, SVM, SciBet, scLearn, and scPred performed well on the various cell types (>0.8 F1), However, hierarchical and correlation-based classifiers performed poorly (<0.1 F1 for some cell types). In both cases, the cells that have higher complexity, such as progenitor cells, are correlated with a lower performance by cell type prediction tools compared to fully differentiated cell types.

Altogether, these results indicate that the cell annotation of mature cells with distinct expression profiles that are delineated, such as endothelial and fully differentiated neurons, are more easily classified by the 13 classifiers we assessed. Within the same dataset, using automated cell type annotation we determined that cells with higher complexity and differentiation potential, such as progenitor cells, cannot be classified as accurately by the automated classifiers. These cell types are often encountered in prenatal datasets, where cells are actively differentiating to form a continuum of cell types and states. We suggest that these features of dataset complexity drive the poor F1 scores and prediction accuracy for these cell types.

A


SPNs
14
1
6
28
38
22
Ctx. excitatory
41
37
17
13
SPNs
27
29
Inter-
neurons
36
18
Endothelial
43
11
21
35
OPCs
4
15
40
Progenitors
16
31
Astrocytes
Mural
SPNs
24
20
SPNs
25
19
0
5
3
30
Endo.
42
33
12
2
32
8
7
10
Progenitors
26
Ependymal
Progenitors
Neurogenic progenitors
34
39
23
9
Microglia
UMAP2
UMAP1

B

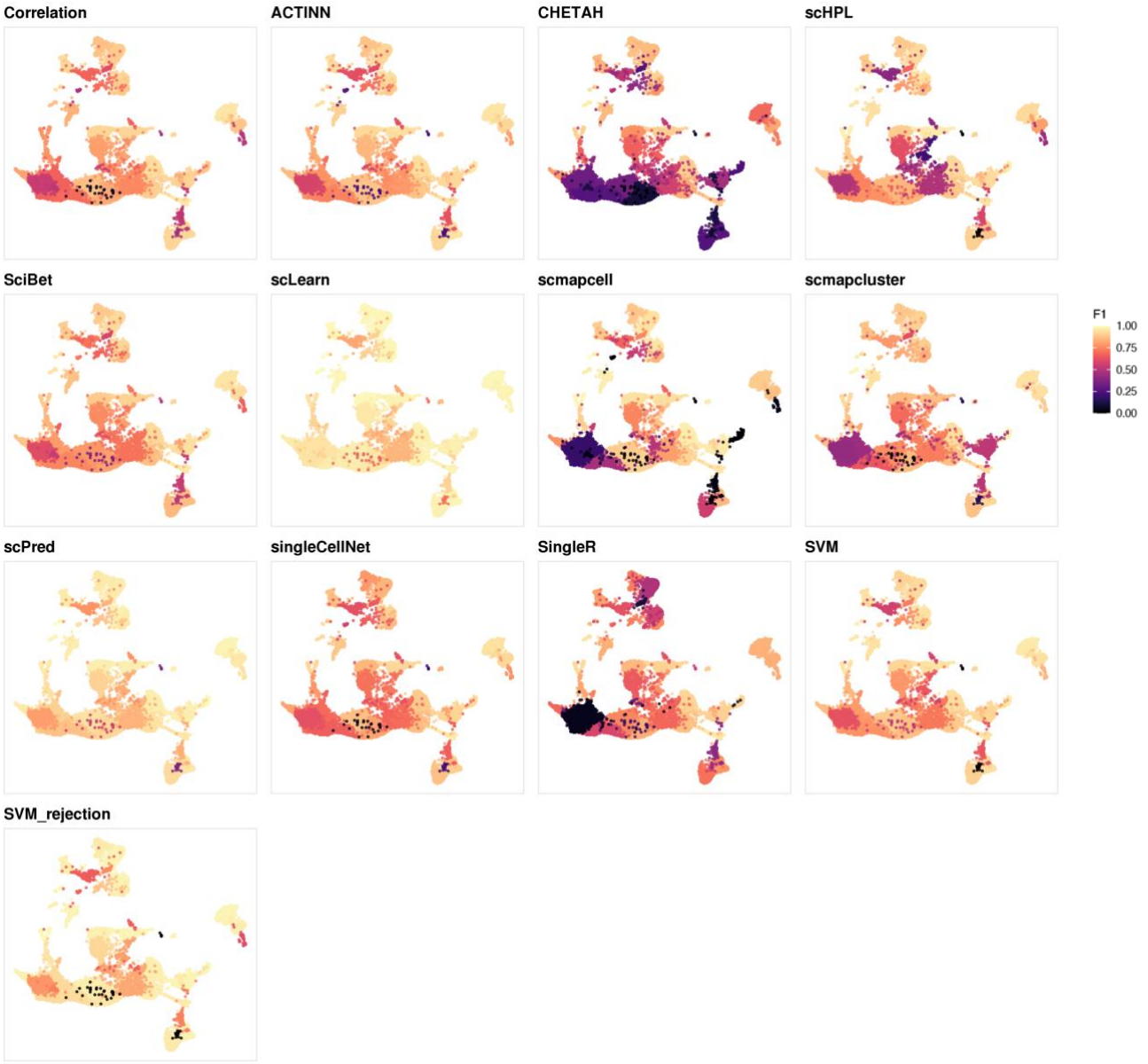


Figure 11 Figure showing classifier performance across cell types in the Anderson dataset. A. UMAP of the Anderson dataset (14,466 cells) colored by cell type. B. UMAPs showing F1 score per cell type per tool. Lighter shades indicate higher performance. Figure generate by Samantha Worme.

3.5. Cell type prediction tool evaluation

3.5.1. Developmental and mixed datasets show decreased performance across classifiers

We next assessed classifier performance on all 17 benchmarking datasets (Section II – Table 3) with cross-validation, as in section III-3.3, reporting F1 scores and rejection rates per dataset and classifier. We also calculated the Pearson correlation of F1 scores and rejection rates with the validated complexity metrics N1 and F2 (section II-2.4).

First, F1 scores showed a variation in performance between classifiers across the dataset types, as shown in the F1 heatmap **(Figure 12).** All classifiers, except for Spearman correlation and NeuCA, generally performed well in all postnatal datasets (>0.9 F1-score). On the other hand, most classifiers showed a significant drop in performance when annotating prenatal and mixed datasets. Correlation-based tools and NeuCA showed the worst performance when tasked with annotating prenatal datasets. CHETAH also showed a decreased performance in these datasets. Scmap performed significantly worse in the prenatal setting than in the postnatal setting. Discriminative tools showed the best performance overall and were the least affected class. SVM rejection showed the most consistent F1 scores (more than 0.96 F1-score) across all datasets regardless of type. Generative tools such as SciBet and scLearn also performed well. However, scLearn needed to be more computationally efficient and could only be run on some datasets.

A further interesting finding was the poor performance of neural network-based classifiers compared to other discriminative tools when tasked with annotating the prenatal and mixed datasets. This low-quality performance could be due to a need for a large number of cells to properly optimize the parameters, given the complexity of prenatal datasets. Our prenatal and mixed datasets included less than 70,000 cells. Given the number of genes detected in these datasets (> 20000), we think we need larger datasets to reach optimal accuracy. Moreover, scPred, scLearn, and scClassify failed to run within a reasonable time due to their time complexity on several occasions, highlighted in the figures with NA. NeuCA also encountered runtime errors due to the model not converging on some datasets. This occurred exclusively with prenatal and

mixed datasets. Lastly, the correlation classifier exhibited inconsistent behavior in which it performed well on specific datasets such as the CellBench dataset, composed of 5 lung adenocarcinoma cell lines, but failed on other datasets such as the Mizrak postnatal neurogenesis dataset. This behavior is likely due to a technical problem where the classifier annotates all cells as one cell type. It is important to note that all classifiers performed well (F1=1.00) with the Cell Bench dataset, as we expected for this control dataset.

Next, we calculated the rejection rate of the 8 classifiers with a rejection option (Figure 13). All classifiers have a significantly lower rejection rate (less than 10%) in postnatal datasets (except for scmap-cell) with values close to zero. On the other hand, rejection rates were significantly higher in prenatal and mixed datasets (up to 91% of cells for scmap). Overall, scmap-cell, scmap-cluster, and scLearn show the highest rejection rate out of all classifiers. SVM rejection and scHPL show significantly higher rejection rates (on average 20%) in developmental and mixed datasets than postnatal datasets (average of 4%), but not to the same extent as the previously mentioned classifiers. Poor cell type separation in prenatal tissues may lead to lower prediction confidence when annotating. In these cases, classifiers reject giving a label instead of outputting a likely false label.

These results confirm the hypothesis of lower classifier performance in prenatal datasets. We observed that classifiers perform well in postnatal datasets (>0.9 F1 score) but suffer from a performance drop in prenatal tissues. Moreover, many classifiers resort to rejecting labels due to the higher complexity of cell types in prenatal tissues. This limitation of the classifiers, in turn, means a large proportion of the dataset needs to be evaluated manually. Similar to what was reported by Abdelaal et al[100], we also observed high performance by SVM classifiers. SVM outperformed neural network classifiers, which may suffer from a lack of sufficient data to properly train. Lastly, we observed a superior

performance of discriminative models, which performed well across prenatal, postnatal and mixed dataset types compared to the other classes.

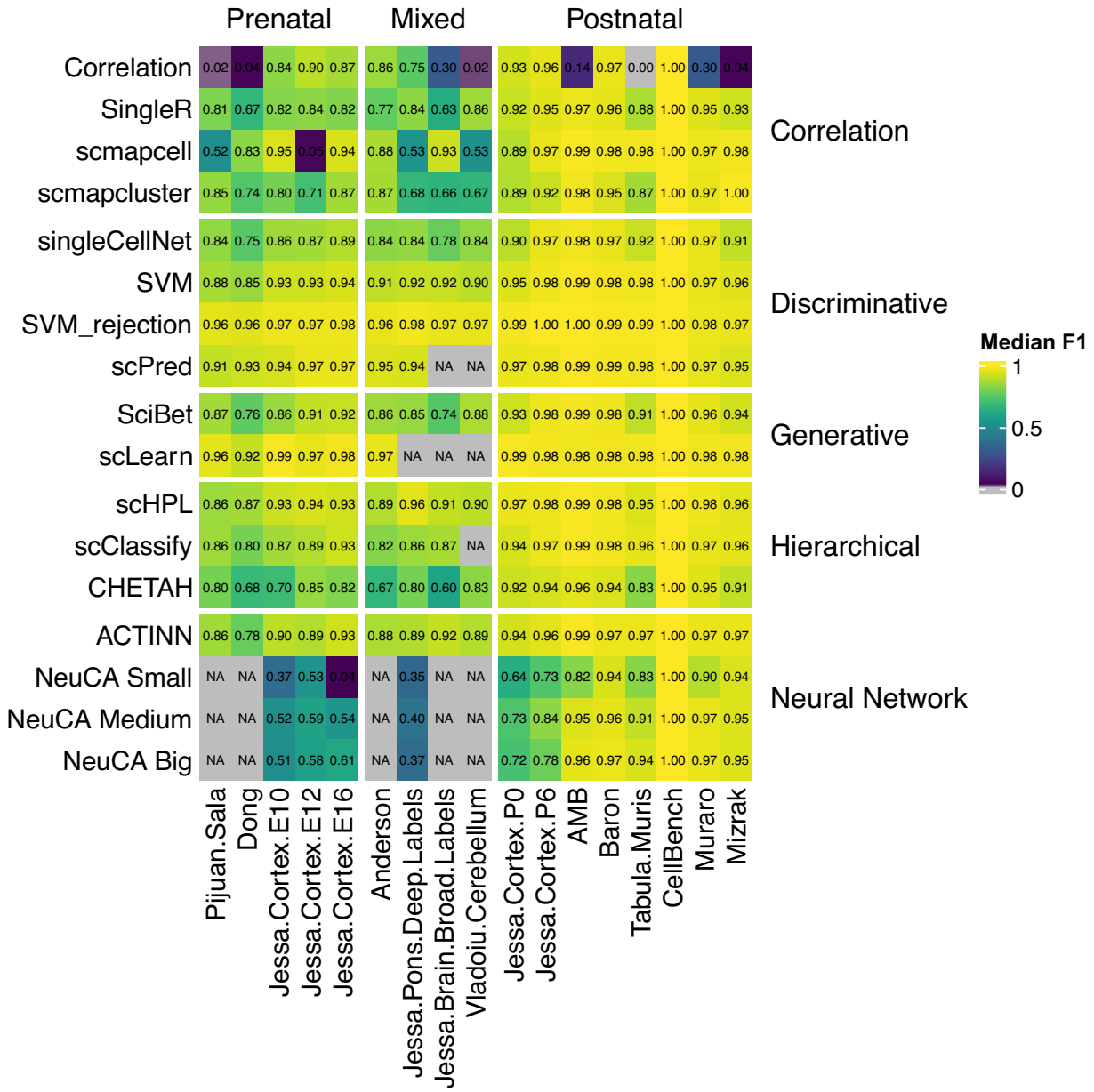


Figure 12 Performance of classifiers on benchmarking datasets. Heatmap of median F1 across all cell populations per classifier (rows) per dataset (column). NA indicates the classifier could not be trained on the corresponding dataset. Datasets segregated into the three classes (prenatal, mixed (prenatal and postnatal), and postnatal.

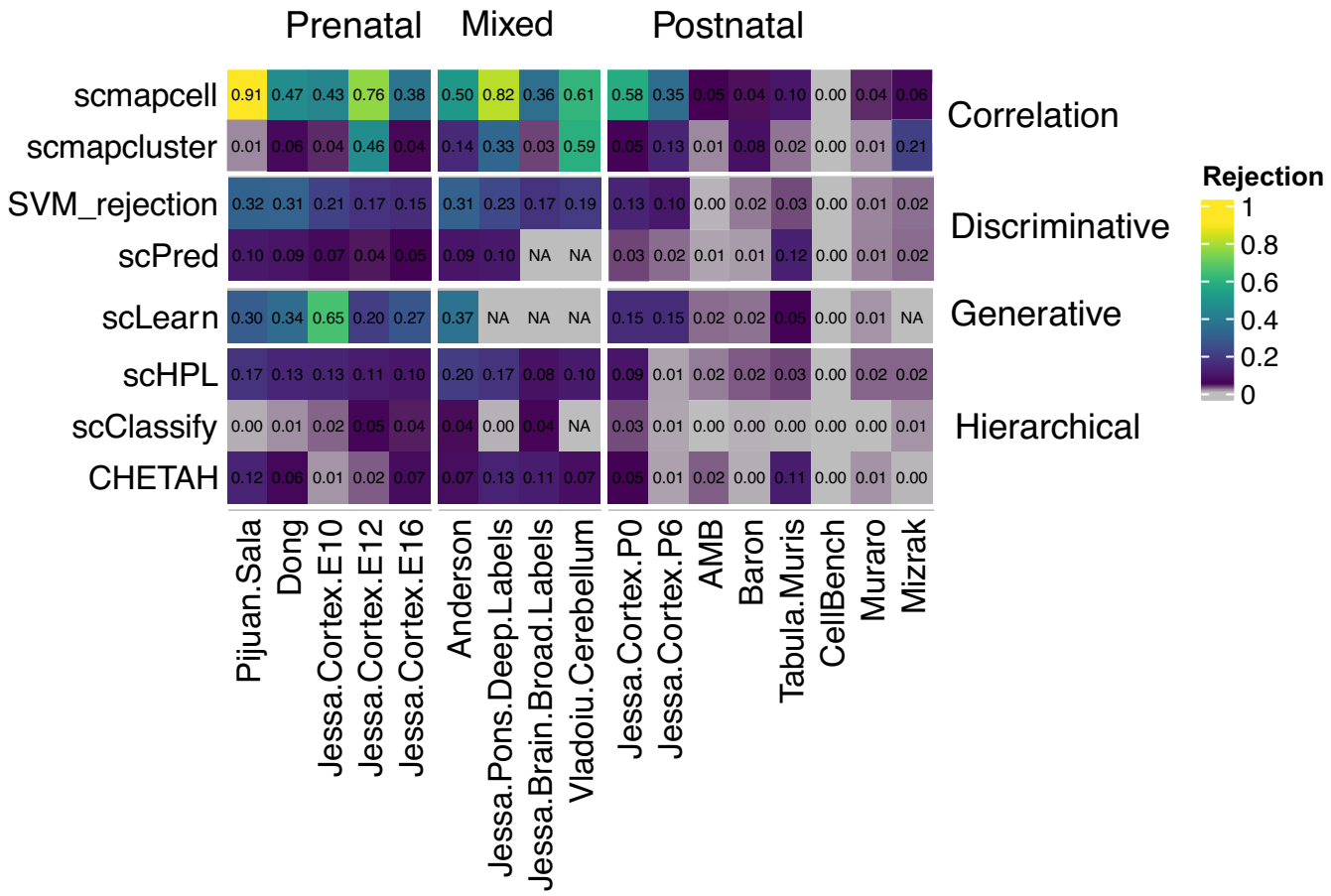


Figure 13 Performance of classifiers on benchmarking datasets. Heatmap of proportion of rejected cells across all cell populations per classifier (rows) per dataset (column). NA indicates the classifier could not be trained on the corresponding dataset. Datasets segregated into the three classes (prenatal, mixed (prenatal and postnatal), and postnatal.

3.5.2. Performance of classifiers is correlated with complexity of datasets (N1 and F2)

To systematically assess whether classifier performance worsens with increasing dataset complexity, we calculated the Pearson correlation (R^2) between classifier performance scores, median F1 score and rejection rate, and complexity metrics, N1 and F2. We also fitted regression lines with 95% confidence intervals. We present these results in the scatterplots, where each dot represents a dataset.

First, we calculated the Pearson correlation between the F1 performance score and the two-complexity metrics (Figure 14 and Figure 15). Several tools such as NeuCA big, NeuCA medium, and scClassify, display a high negative correlation ($R^2 \geq 0.6$) between performance (F1) and cell separability (N1). NeuCA big performance was most strongly correlated to N1 complexity (R^2=0.89). Other tools also show some correlation with N1 but to a less extent. All classifiers displayed lower correlations between their performance and the F2 score for genetic redundancy (Figure 15). Nevertheless, some tools, such as SVM, singleCellNet, scPred, and scHPL, showed a moderate correlation ($R^2 \geq 0.4$). The Correlation classifier seems to be an outlier with no relationship between the F1 performance score and complexity measures, N1 and F2 (R^2=0.04 and R^2=0.018, respectively). This effect is likely due to its inconsistent behavior across the different datasets used.

We also investigated the correlation between rejection rates and N1. Here, we can see a significant correlation ($R^2 \geq 0.64$) between rejection rate and N1 for most tools, such as SVM rejection, scLearn, and scClassify (Figure 16). This data confirms our hypothesis that poorly separated cells, as measured by N1, may lead to higher rejection rates in some classifiers. This result suggests that, in some cases, poor confidence in a cell type label may lead a classifier to reject a cell rather than provide a false label.

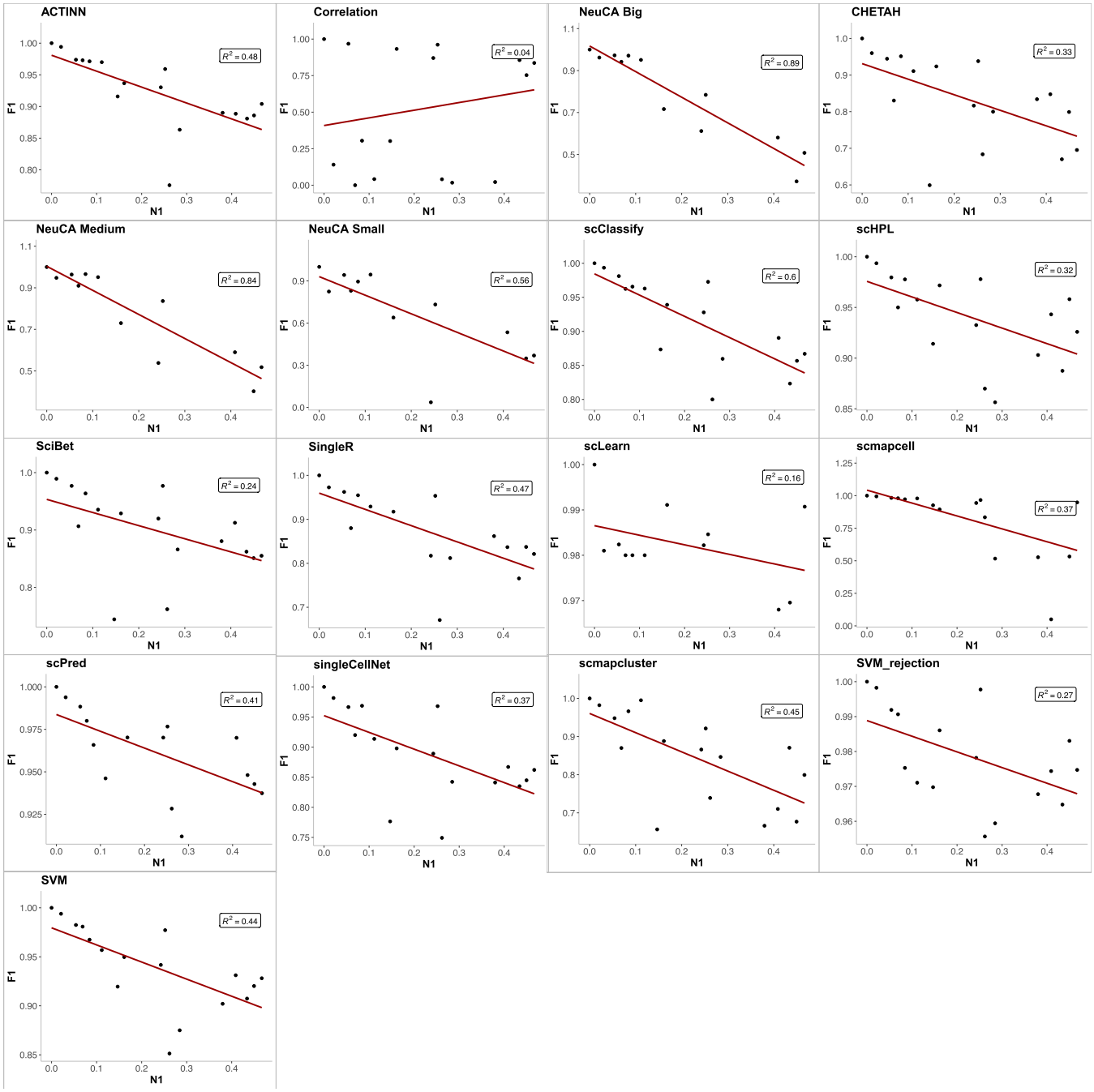


Figure 14 Correlation between the fraction of borderline cells (N1) and median F1 performance score per classifier. Each graph shows a classifier with a regression line fitted and R^2 computed. Individual points represent the benchmarking datasets with their calculated N1 and F1-scores.

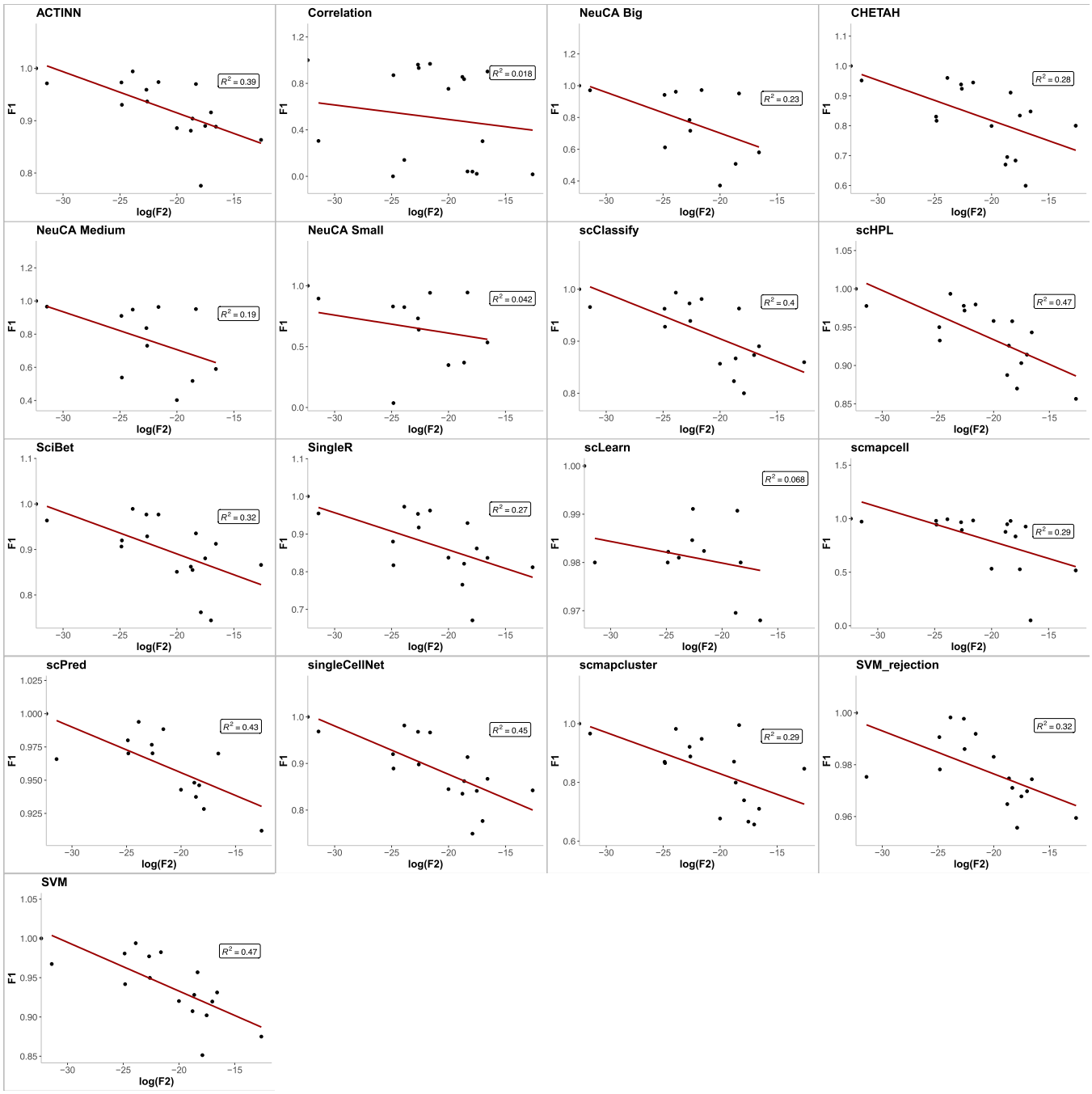


Figure 15 Correlation between volume of overlapping region (F2) and median F1 score per classifier. Each graph shows a classifier with a regression line fitted and R^2 computed. Individual points represent the benchmarking datasets with their calculated log(F2) and F1-scores.

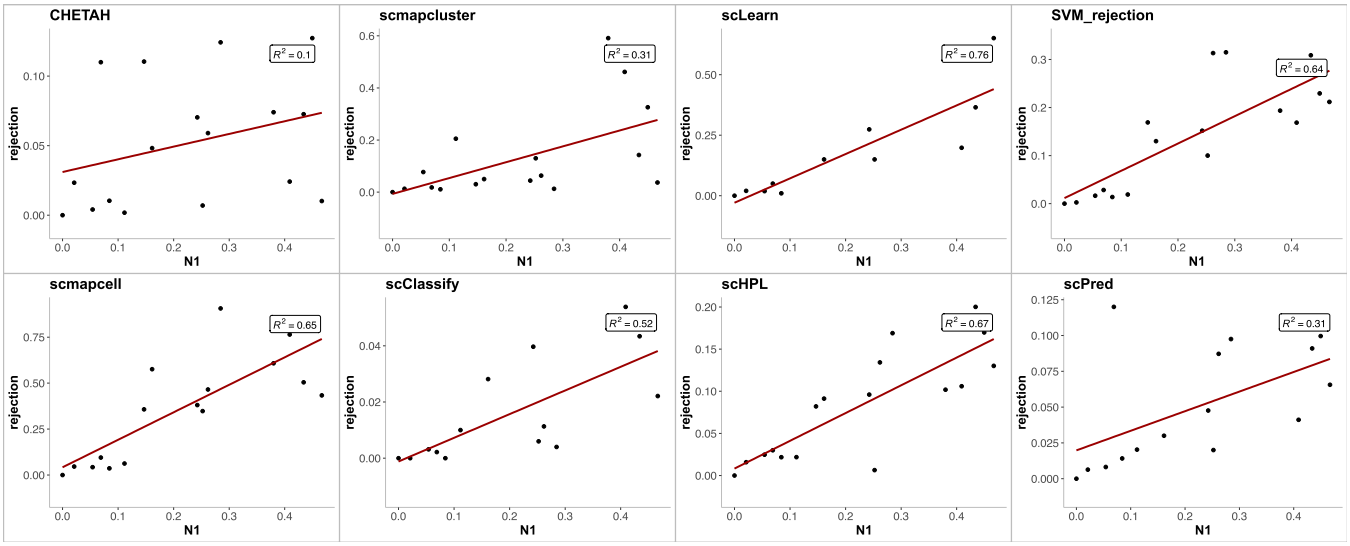


Figure 16 Correlation between fraction of borderline cells (N1) and rejected proportion per classifier for classifiers with a rejection option (N=8). Each graph shows a classifier with a regression line fitted and R^2 computed. Individual points represent the benchmarking datasets with their calculated N1 and proportion of rejected cells (rejection proportion).

4. Chapter IV: Applications in cross-species predictions and cancer

This chapter will explore cell type annotation tools in the context of two applications of single-cell RNA-seq. First, we will assess their performance for cross-species prediction, where cells of one species are annotated using references from a different species. Given the widespread use of animal models in biomedical research, cross-species annotation has become a common task. There now exists an abundance of well-annotated murine datasets. Furthermore, there are fewer challenges in generating murine datasets for human datasets, given ethical constraints for research with human specimens. Exploiting the abundance of data from one species to annotate data from other species would simplify annotation tasks. The second application of cell type annotation that we will examine involves using cell type classifiers to identify cell types of origin implicated in cancer development. Identifying cell types of origin can strengthen our understanding of the biology of tumors and guide the medical community in both the diagnosis and treatment of certain tumors subtypes. Annotation of tumor datasets is often called "projection", since we are projecting cancer cells onto specific normal cell types from normal tissue reference datasets. For these two applications, we use the computationally efficient cell type classifiers implemented in scCoAnnotate (section II-2.5.2). This method involves training and testing classifiers using both mouse and human datasets (Figure 17).

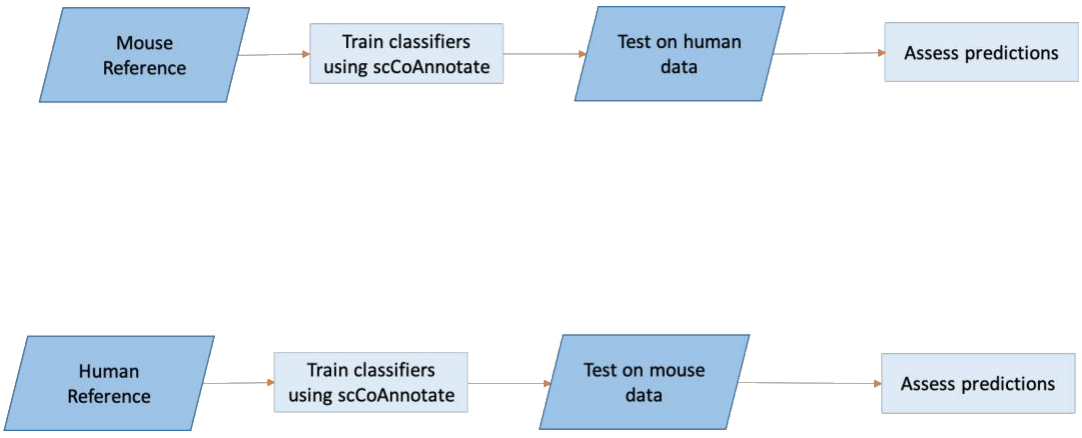


Figure 17 Workflow of the cross-species cell type prediction experiments. For each human-mouse pair, two experiments were performed.

4.1. Assessment of cell type annotation tools for cross-species predictions

We assessed 12 cell type annotation tools and their consensus, implemented in scCoAnnotate, on two pairs of mouse and human datasets (section II – Table 3). The tools selected were ACTINN, CHETAH, correlation, SciBet, scHPL, scmap-cell, scmap-cluster, scPred, SingleR, SingleCellNet, SVM, and SVM rejection. Some tools, such as scLearn and scClassify, were excluded for computational reasons (failure of the tool, high runtime requirement). Moreover, NeuCA failed to converge in these experiments. We first assessed these tools on the Baron pancreatic datasets[105] and then applied the same tools to the Allen Brain datasets[145,146]. These datasets are well suited for this task as they profile the same tissues in mice and humans. In addition, these datasets profile similar cell types across the two species with few inconsistencies in labeling, such that most cell types present in the mouse dataset can also be found in the human dataset. For our analysis, we maintained the inconsistent labels when encountered to assess if classifiers can annotate these cells as similar cell types present in the training reference.

4.1.1. Cross-species prediction assessment using Baron pancreatic datasets

We first trained the classifiers on the human Baron pancreatic reference, which contains 14 distinct cell types. With only 13 cell types, the Baron mouse pancreatic dataset contains fewer cell types than the human reference. Moreover, not all cell types are the same as the human reference. For example, the mouse dataset contains B cells and immune_other labels, which are absent from the human reference. On the other hand, the human reference contains mast cells, epsilon cells, and acinar cells, which are not present in the mouse dataset. The inconsistent labels were maintained in the training and query datasets to assess the performance of classifiers when they encountered a cell type missing from the reference. We used the trained classifiers to annotate the mouse pancreatic dataset. We then visualized the predictions using a confusion matrix (Figure

18), which shows the proportion of each mouse cell type (rows in the heatmap denoted as ground truth) projected to human cell types in the reference (columns).

As expected, the performance of cell type annotation tools displayed high variance across the different cell types (Figure 18). SciBet and SVM were the best performers, performing well on the cell types common between the reference and query datasets. Moreover, for the mouse cell types that were not present in the reference, such as B-cells and immune_other, SciBet predicted the corresponding cells as immune cells present in the reference and not as random cell types. On the other hand, SVM had the most correct predictions in the cell types common between the reference and query but had worse performance than SciBet in the cell types absent from the reference. For example, B-cells were projected as Beta cells.

Conversely, most other tools performed poorly in this experiment. For example, tools such as scPred, CHETAH, scmap-cell, and scmap-cluster could not project most mouse cells into a cell type present in the human reference and therefore did not assign any labels. Other tools, such as SingleCellNet and correlation, performed well on certain cell types, such as endothelial cells and macrophages. Still, they could accurately project other cell types, such as Alpha and Delta cells. Not only did we observe a high variance in performance between tools, but the performance of a particular tool in predicting different cell types was also highly variable. This data suggests that some cell types are easier to classify due to their distinct gene expression profiles.

Next, we repeated the same experiment with the reference and query dataset reversed. In this experiment, we trained classifiers on the mouse dataset and then annotated the human dataset (Figure 19). Here, we observed similar performance trends as the previous experiment, although the performance of classifiers was lower than when we used a human reference to annotate mouse data. The proportions of certain cell types were higher in the mouse than in the human datasets, which may play a role in the poorer performance observed in this experiment. Specifically, the mouse dataset contained mostly alpha and beta cells, with fewer ductal cells than its human counterpart.

Similar to previous results, SVM and SciBet demonstrated the best overall performance across all cell types. Correlation also performed relatively well in this experiment. On the other hand, all other tools perform poorly, particularly the ones with a rejection option. scHPL, scmap-cell, scmap-cluster, and scPred did not annotate most of the query dataset. These results demonstrate that, in many occasions, classifiers could assign a high-confidence label when annotating

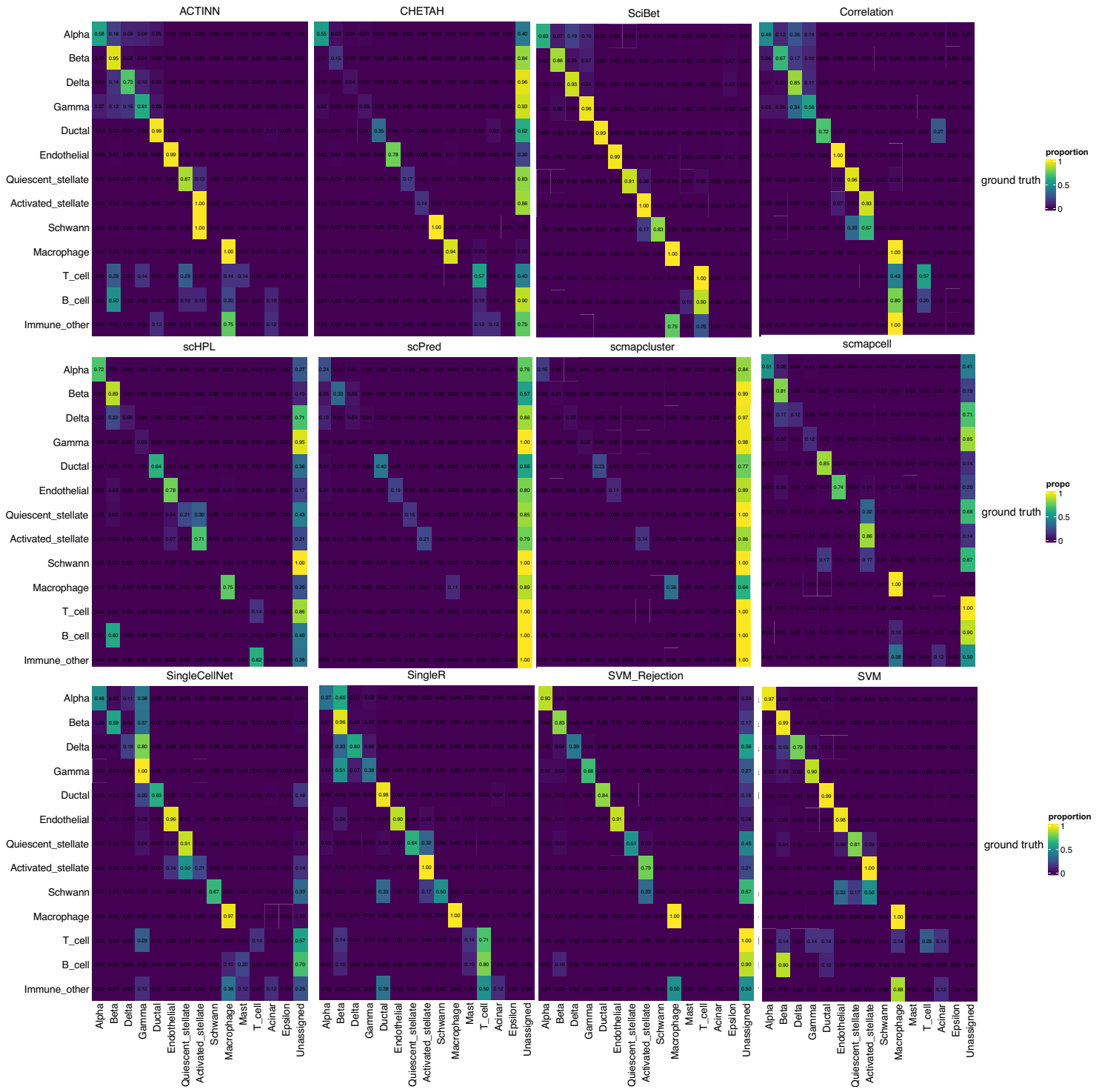


Figure 18 Confusion matrices comparing projected cell types in a mouse pancreatic dataset based on human pancreatic reference. For each heatmap, proportions were computed row-wise and represent the fraction of cells from each mouse pancreatic cell type which were assigned to each human pancreatic label. Each heatmap represents the predictions from one single-cell annotation method.

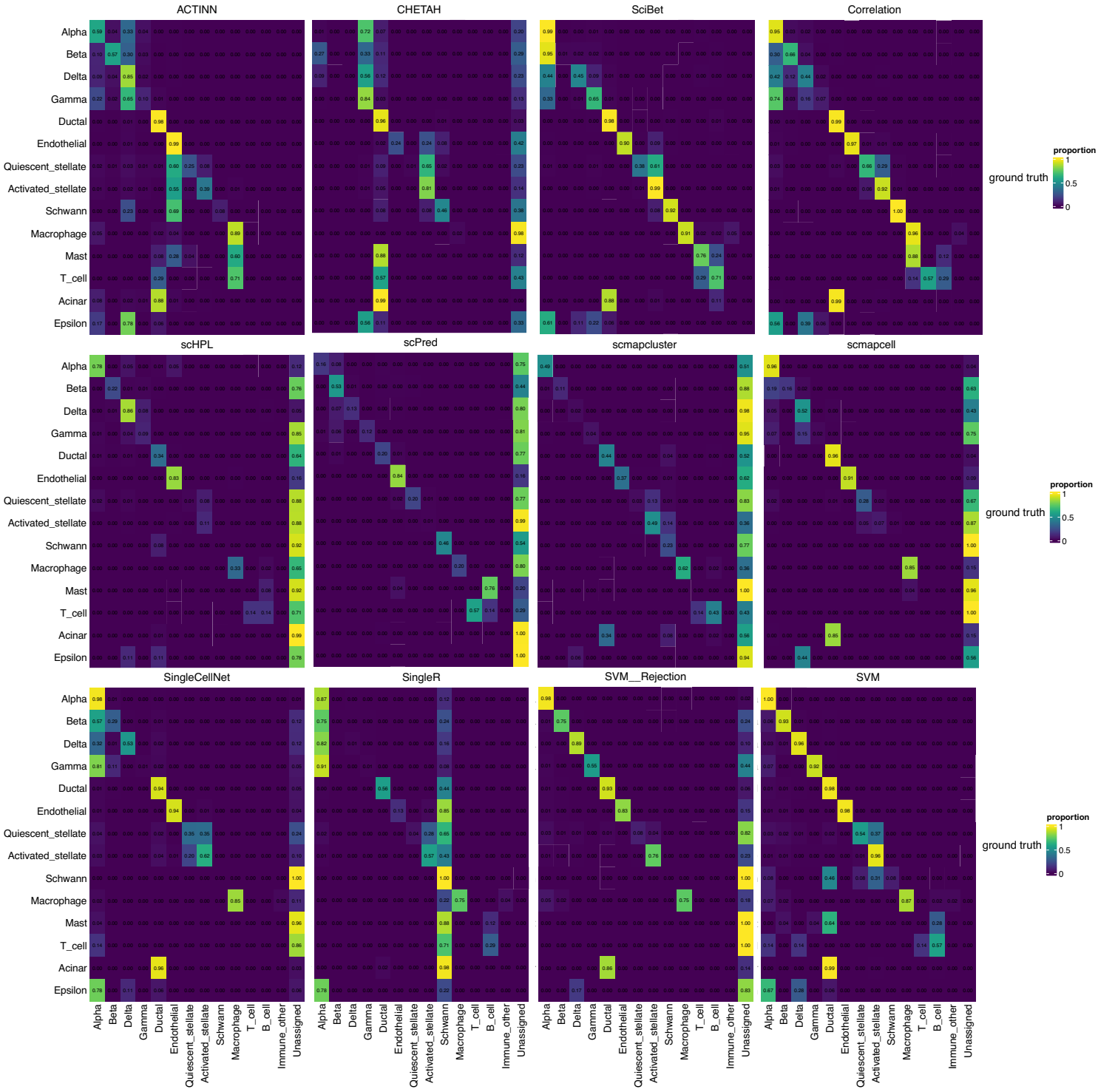


Figure 19 Confusion matrices comparing projected cell types in a human pancreatic dataset based on mouse pancreatic reference. For each heatmap, proportions were computed row-wise and represent the fraction of cells from each human pancreatic cell type which were assigned to each mouse pancreatic label. Each heatmap represents the predictions from one single-cell annotation method.

4.1.2. Assessment using the Allen brain datasets

To ensure the reproducibility and generalizability of results, we repeated the previous experiments using the Allen brain datasets originating from the adult mouse and adult human cortices (Section I-Table 3). First, we trained the classifiers on human brain data and used the trained models to annotate mouse brain data. Most cell types were common to both datasets. However, specific cell types were present in one dataset and absent in the other. For example, macrophages were present in the human dataset, whereas the mouse dataset contained microglia but no macrophages. Other distinct cell types include oligodendrocytes progenitor cells (OPCs) and pericytes, found exclusively in the mouse dataset, and smooth muscle cells (SMC), found only in the human dataset.

First, we noted that SciBet had the best performance overall across all the cell types (Figure 20). However, we observed poor performance of scHPL, CHETAH, scPred, scmap-cell, and scmap-cluster. These tools refrained from annotating all cells and left them unassigned. Most of the classifiers successfully identified astrocytes, glutamatergic, and GABAergic neurons (but showed decreased performance in other cell types). Moreover, we also noticed that most tools classified some of the oligodendrocytes as OPCs. OPCs are progenitors of the oligodendrocyte cell population and likely have transcriptional similarities. As such, it is expected that this projection is not random or incorrect. We also noted that most classifiers projected microglia as macrophages, the closest cell type found in the reference.

Next, we trained the classifiers on the mouse reference and annotated the human reference (Figure 21*).* As expected, the same tools with a rejection option failed to annotate cells. Unexpectedly, SVM rejection also performed poorly in this experiment. However, we observed that some tools successfully annotated human cells, such as correlation and SingleR. SciBet and ACTINN also performed well, but SciBet had poorer performance on glutamatergic neurons. As expected, tools that performed well overall labeled OPCs (not present in the reference) as oligodendrocytes and microglia as macrophages. This prediction is biologically relevant because of their respective transcriptional similarities. Moreover, classifiers mostly predicted pericytes as vascular,

endothelial, or smooth muscle cells. Together, these cell types form the brain vasculature. Lastly, we observed a decreased performance of classifiers in brain datasets compared to pancreatic datasets. This may be due to the higher cell type similarity in brain cell tissues to the pancreas, which could cause incorrect annotations.

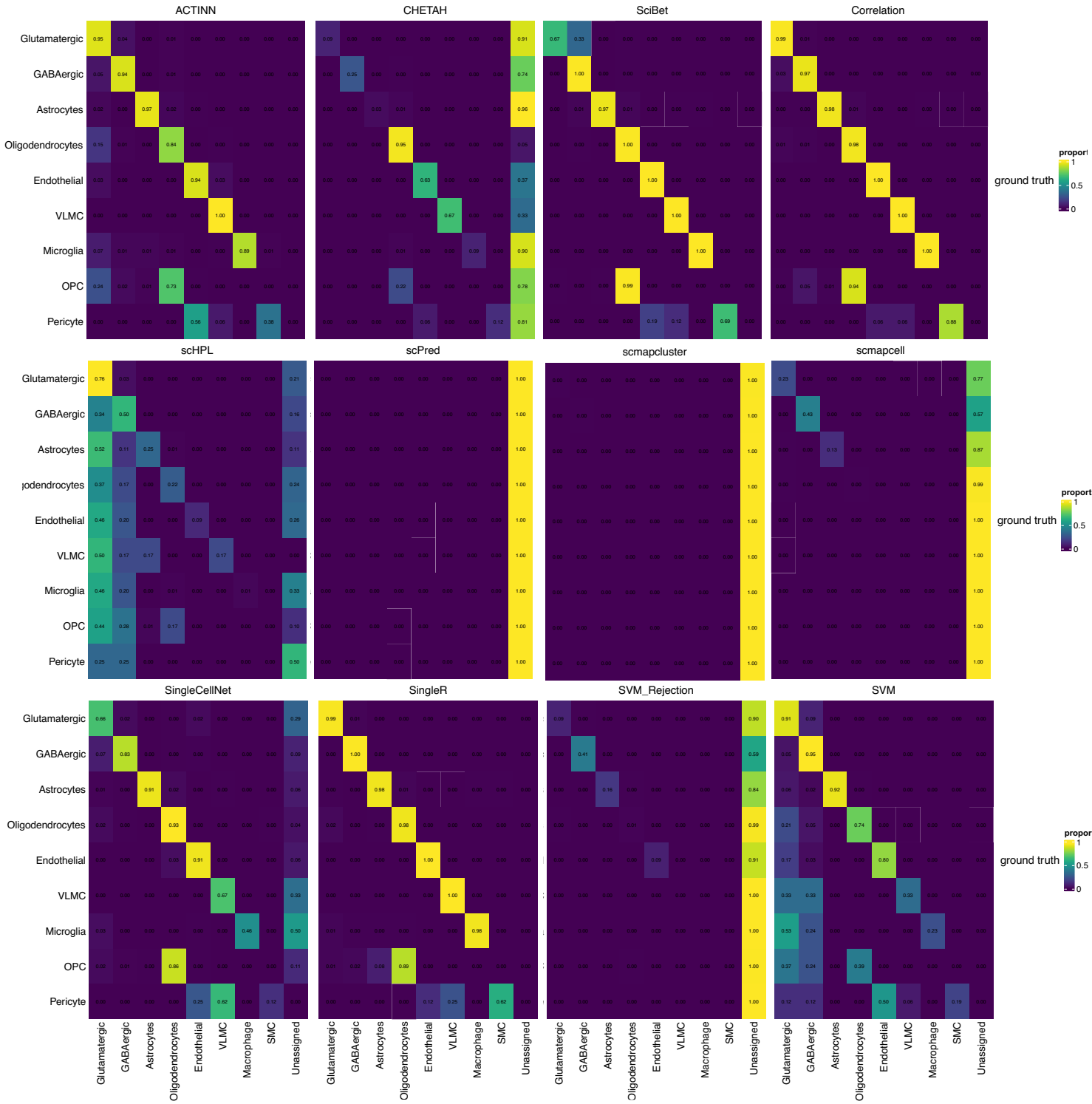


Figure 20 Confusion matrices comparing projected cell types in a mouse brain dataset based on human brain reference. For each heatmap, proportions were computed row-wise and represent the fraction of cells from each mouse brain cell type which were assigned to each human brain label. Each heatmap represents the predictions from one single-cell annotation method.

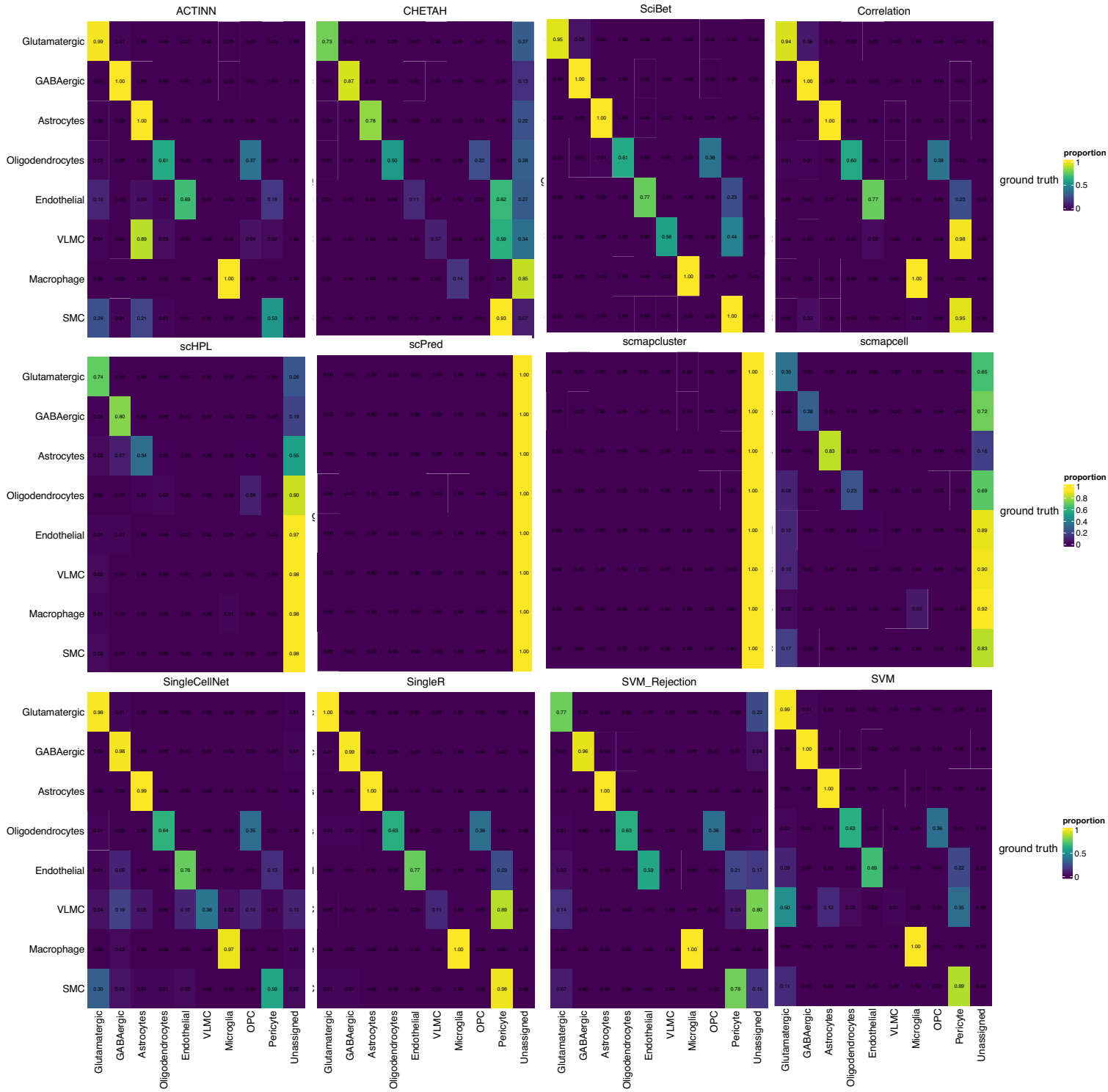


Figure 21 Confusion matrices comparing projected cell types in a human brain dataset based on mouse brain reference. For each heatmap, proportions were computed row-wise and represent the fraction of cells from each human brain cell type which were assigned to each mouse brain label. Each heatmap represents the predictions from one single-cell annotation method.

4.1.3 The consensus approach outperforms any classifier and yields relevant predictions

We next sought to investigate the applicability of an ensemble consensus prediction in the cross-species setting. The use of consensus prediction in machine learning (ML) classifiers for cell type annotation has several benefits. First, combining the predictions of multiple classifiers can reduce the variance and improve the overall accuracy of the cell type annotation. Second, it can also address the limitations of individual classifiers, such as overfitting or limited performance on specific subpopulations. Third, consensus prediction can also provide a more robust and reliable classification, as it accounts for the inherent noise and variability in the data. By combining multiple classifiers' strengths, a consensus approach can improve performance and confidence in cell type annotation in scRNA-seq datasets. Here, we utilized the consensus approach implemented by our pipeline scCoAnnotate. The consensus approach uses a simple majority vote of 12 classifiers (Section II – 2.5.2) to call a prediction. In cases where the vote on a label is tied, the result would be an "unsure" label. We tested our pipeline using the same datasets from the previous two sections.

First, using the Baron pancreatic datasets, the consensus prediction of mouse cells when using a human pancreatic reference correctly predicted most cell types (Figure 22a). Since the human reference did not contain B cells, so mouse B cells were annotated as T cells. In addition, Immune_other cells were predicted as macrophages. This is valuable because the absent cells were classified as the next biologically similar cell type. However, some predictions, such as a proportion of B and T cells, were annotated as beta and gamma cells. Overall, the consensus predictions for this dataset yielded high concordance with the actual labels compared to any individual tool (Figure 18).

On the other hand, the consensus approach yielded less accurate predictions when we used the mouse pancreatic reference to annotate the human reference, particularly for Beta and Gamma cells. From Figure 22b, we observe that the consensus prediction in this dataset still outperformed any single tool (Figure 19), with most inconsistencies occurring between highly similar cell types, which we expected. For example, human

quiescent stellate cells were predicted as a mix of activated and quiescent mouse cells. Moreover, T cells were predicted to be a mix of T and B cells. All human acinar cells were annotated as ductal cells (the most similar cell type), and most epsilon cells were annotated as alpha and delta cells which are also endocrine cells.

The consensus approach yielded high prediction concordance when annotating the mouse Allen brain dataset using the human cortex reference (**Figure 22c**). We can observe that most cells were annotated as relevant cell types found in the reference. For example, mouse Oligodendrocytes were partially annotated as OPC, and mouse macrophages were annotated as human microglia. However, we noticed that vascular and leptomeningeal cells (VLMCs) were annotated as human pericytes instead of human VLMCs. These cell types, in addition to SMCs, are from the same lineage.

Lastly, we used the consensus approach to annotate the human cortex cells using a mouse cortex reference. This experiment yielded the most accurate predictions (**Figure 22d**). Almost all human cell types were assigned correctly to their murine analog. However, only human glutamatergic neurons had a low concordance (58%), with the remaining cells in that cluster having no consensus. The lower concordance for these neurons may be due to the higher complexity of human neuronal cells. The added complexity could make assigning these cells to simpler murine neuronal cell types more challenging.

In conclusion, cross-species cell type classification in scRNA-seq datasets can pose significant challenges due to reference-query inconsistencies. Moreover, differences in gene expression patterns and cellular characteristics between species can pose unique challenges for cell annotation tools. In such cases, it is common for some machine learning tools to fail while others provide accurate results. A consensus approach could help address these challenges by combining the predictions of multiple tools and leveraging their strengths to produce a more robust and precise classification. However, the choice of consensus approach and the specific methods used must be carefully

considered and modified to account for the specific requirements of the project and the limitations of each tool.

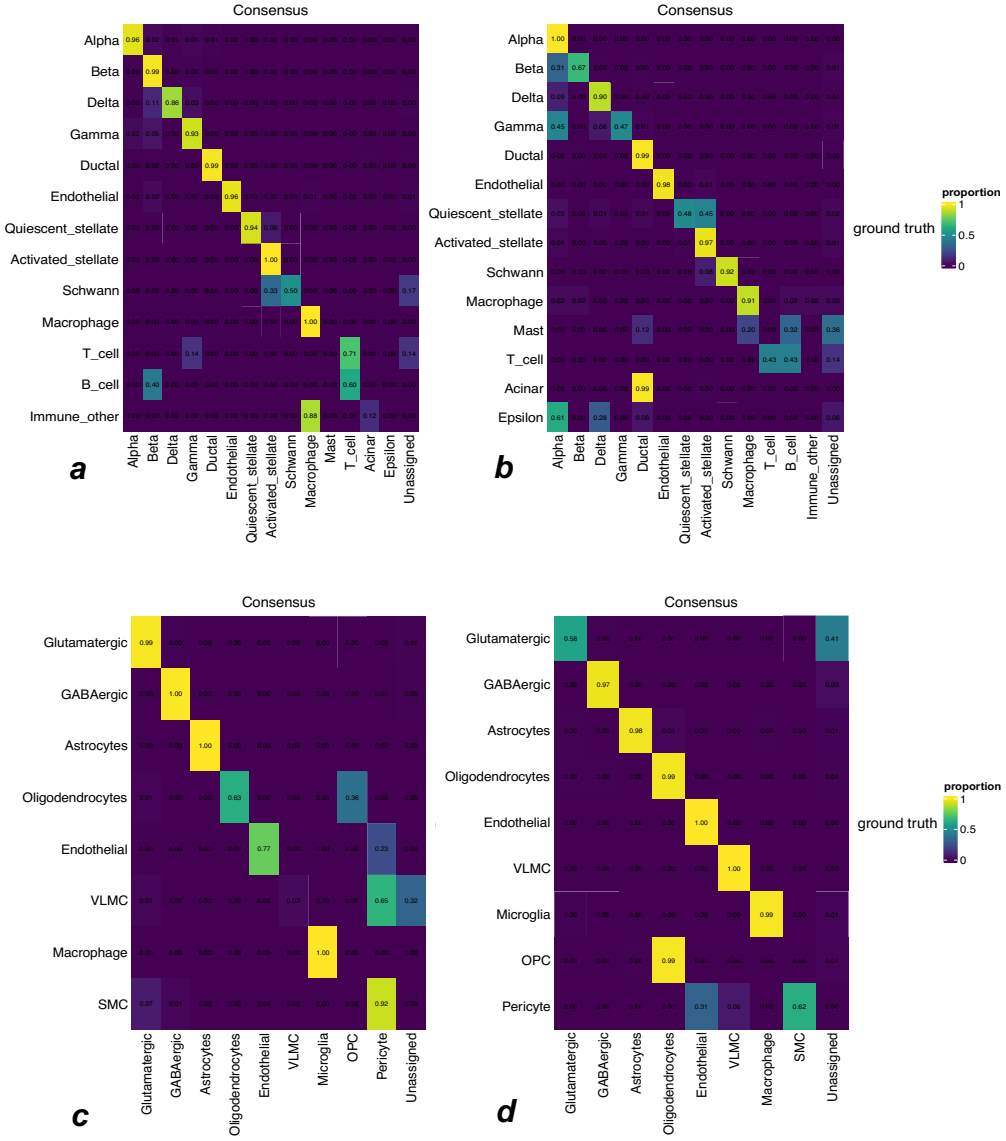


Figure 22 Consensus predictions for a. using a human pancreatic reference to annotate mouse pancreatic cells; b. Using a mouse pancreatic reference to annotate human pancreatic data; c. Using a human cortex reference to annotate mouse cortex cells; d. Using a mouse cortex reference to annotate human cortex cells. Consensus labels are the majority prediction of the 12 classifiers, cells with no majority prediction as left unassigned

4.2. Identification of the closest normal cell type for high grade gliomas using automated approaches

Cell type identification is crucial for understanding the underlying biology and developing effective treatments for many types of cancer, including pediatric brain tumors. Pediatric brain tumors are a heterogeneous group of diseases, and survival varies widely with the tumor subtype[147–149]. High-grade gliomas, a subset of tumors presumed to originate in glial cell types, represent the greatest cause of cancer-related childhood mortality [150,151]. These tumors are characterized by their aggressiveness, resistance to therapy, and fast progression. Histone-mutant gliomas are a subset of these cancers that are caused by mutations in the genes encoding Histone 3 proteins, which play a crucial role in regulating gene expression. For these tumors, the context of the cell type of origin is critical to understanding tumorigenesis. Histone-mutant pediatric gliomas are thought to arise from cells that have stalled differentiation[152]. Moreover, different histone mutations that occur in distinct brain regions and involve different cell types can lead to distinct tumor types. For example, Histone 3.1, 3.2, and 3.3 K27M high-grade mutant gliomas are thought to arise from OPC-like cells, whereas ependymomas arise from ependymal-like cells[33,34,153,154]. Understanding cell-of-origin and cellular hierarchies within tumors is critical for studying tumor biology. An accurate projection of cell types within these tumors can provide valuable insights into their molecular and functional diversity.

This section will discuss the application of cell type annotation tools to identify cell types of origin in pediatric high-grade glioma scRNA-seq datasets of various tumor types. First, we will present the projections obtained using a consensus of classifiers trained on the developing mouse brain using scCoAnnotate (section II-2.5.2). Second, we present the validation of these projections by assessing the agreement between mouse projections and projections obtained using a human thalamic fetal atlas as a reference. We performed this process on a thalamic subset of the samples for which the relevant normal human brain data were available. The process is depicted in the flowchart below (Figure 23).

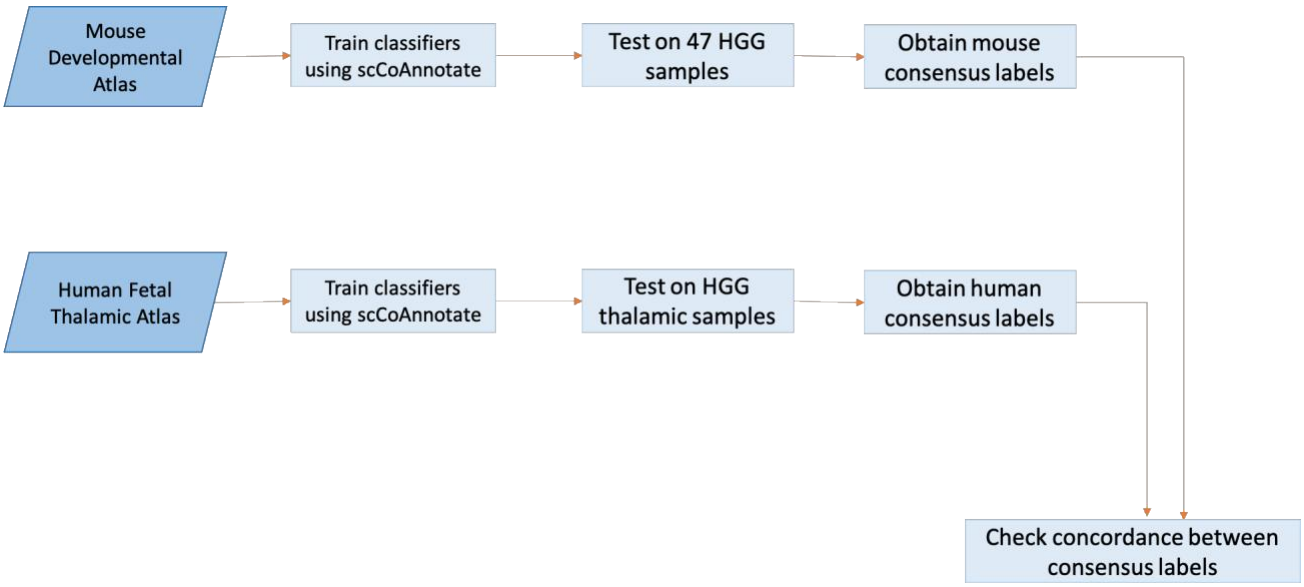


Figure 23 Workflow for identifying cell types in high grade gliomas using the mouse developmental atlas. Verification of predictions was performed on a subset (thalamic samples) using a human fetal thalamic atlas

4.2.1. Cell type projections of high-grade glioma samples using a developmental mouse atlas

We used scCoAnnotate to annotate 47 sc/snRNA-seq datasets obtained from 43 patients using 10X Genomics technologies[34] (Section II - Table 3). Three classifiers (SVM, Spearman Correlation, and SciBet) were trained on the mouse developmental atlas and used to identify cell types in 176,934 cells. These classifiers were chosen because of their computational performance and suitability for tasks where there is large reference-query mismatch (i.e. annotation of cancer cells using normal cells). We annotated cells when the correlation method agreed with at least one of the two other classifiers. In addition, we excluded cells with no consensus label from downstream analyses. The results are summarized below (Figure 24).

As expected, H3.3K27M mutant gliomas consisted of OPC and oligodendrocyte-like cells, consistent with previous studies[33,34,149,150]. In contrast, PFA-EP tumors showed a clear ependymal-like state. In addition, H3.1/2K27M mutant tumors showed a unique profile of cell type predictions (Figure 24a). A subset of samples showed a robust ependymal-like signal later enriched in specific genes conferring the ependymal status (Figure 24b). Moreover, projections of H3.1/2K27M gliomas also showed strong astrocytic signals, confirming the unsupervised analysis of the samples performed by other authors[34].

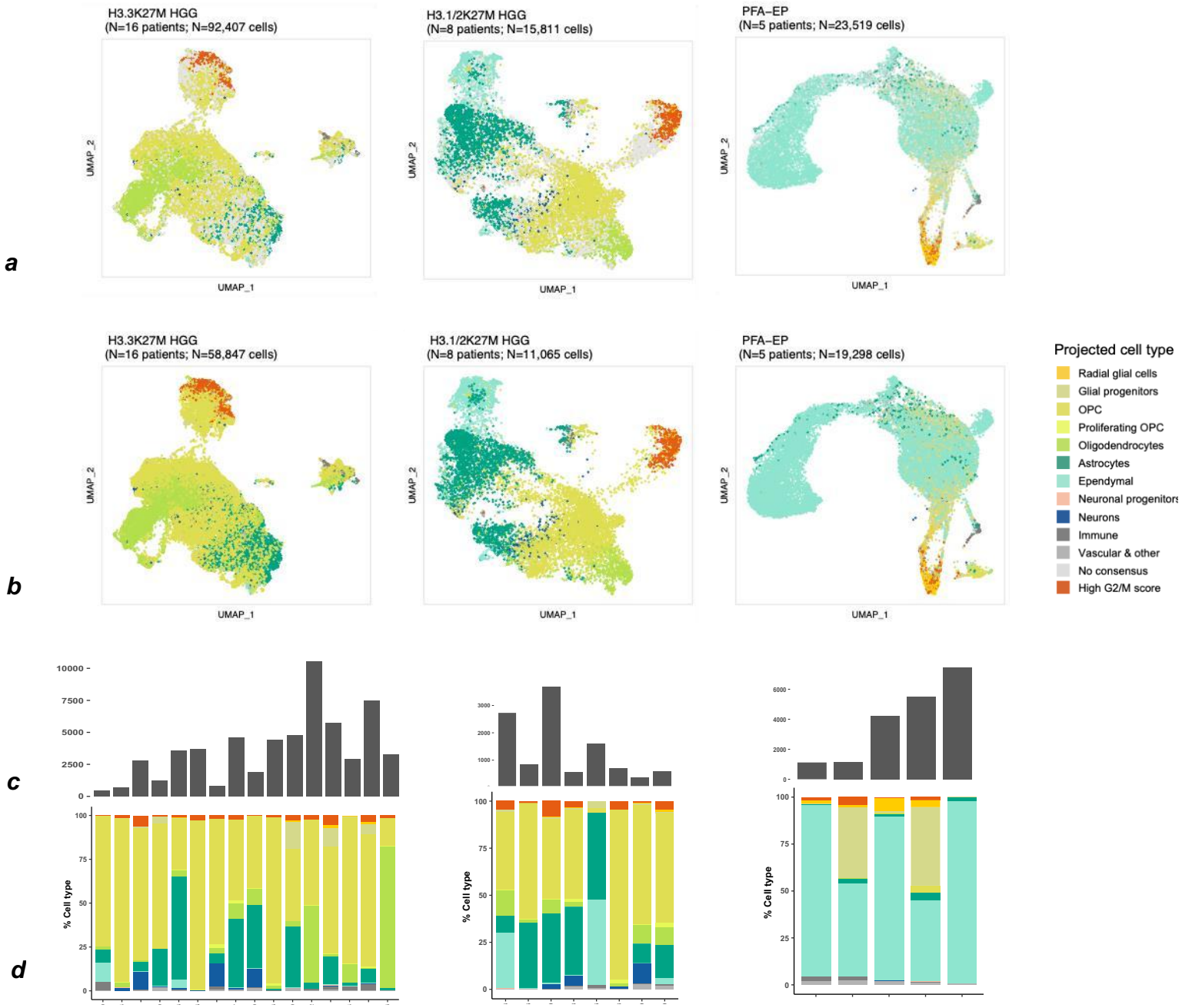


Figure 24 . a. UMAP for tumor malignant cells per tumor type colored by projected cell type obtained using the consensus approach: HGG: high grade glioma, PFA posterior fossa ependymoma. Cells with no consensus and with high G2/M cell cycle scores were colored in orange. Cells with no consensus but low G2/M scores are colored in light gray. b. Same UMAPs as figure 24.a, after removing cells with no cell type consensus. c. Number of malignant cells per sample per tumor class as in a and b. d. Stacked bar plots showing the cell type projection composition of each sample per tumor class.

4.2.2. Verification of mouse projections using a human thalamic atlas

To verify the cross-species projections obtained using a mouse reference, we annotated a subset of thalamic HGG tumor samples using the consensus classifiers trained on either a human prenatal thalamic reference (12 donors) or the Jessa 2022 Mouse Developmental Atlas[34,143,144]. We defined the human consensus label in the same manner as the mouse consensus label, requiring that the Correlation prediction agreed with at least one other method, SVM or SciBet, to produce a consensus label. Otherwise, the cell type was labeled as uncertain. We then generated UMAPs to visualize the projections of malignant cells using the mouse and human atlases (Figures 25 and 26).

UMAPs of mouse and human projections (Figure 25) show similar annotations of clusters with predominant OPC and oligodendrocyte projections. However, we noticed that using the mouse atlas tended to project more cells as astrocytes (Figure 25b). Moreover, mouse projections had significantly more uncertainty, represented by a larger number of unlabeled cells (Figure 26). With both the human and mouse atlas training sets, the most frequent projections are OPCs and oligodendrocytes with some astrocytes and glial progenitors (Figure 26). Mouse-based projections had double the uncertain cells compared to human-based projections. Species-specific gene signatures may make it more difficult to identify cell types. This effect may be particularly striking in malignant cells where aberrant expression profiles with gene copy number variation are more frequent.

Finally, we sought to quantify the concordance between the two consensus projections per cell type. We used the human consensus projections as ground truth and compared them to the mouse consensus projections**.** For the most prevalent cell types (OPCs and oligodendrocytes), the agreement between the projections obtained using the mouse reference and the projections obtained using the human fetal reference is high (>90%, Figure 27). The remaining cells account for a low percentage of total cells. We calculated the F1 scores per cell type to examine the performance of the classifiers (Figure 28). Median F1 was 93% which is significantly higher than what was obtained using only the correlation method (78%). Moreover, for the abundant cell types that are found in this

class of tumors, cell type F1 scores were high (>94% for OPC and oligodendrocytes). The cell types that are the least abundant had lower F1 scores, but this effect was negligible on the overall performance. The lowest F1 scores represent cell known to present biologically challenging situations for cell type annotation, such as differentiating cell types and progenitors. In this case, the radial glial cells (RGCs) are plastic and thus transcriptionally variable, the vascular and other clusters have high heterogeneity, and the proliferating OPCs have a cell cycle signal affecting their gene expression.

In conclusion, we developed and validated a consensus method using multiple ML classifiers for cross-species annotation of high-grade glioma datasets. We demonstrated that the annotation performance of the consensus method was more accurate than using one individual method. Moreover, we identified difficulties encountered by the classifiers when annotating human datasets from a mouse reference. Understanding these obstacles to cross-species annotation is critical, as human tissues are often limited and less accessible due to ethical constraints.

Cancer cell type annotation or projection is critical in analyzing tumor single-cell data. It allows researchers to understand the composition and diversity of cancer populations in cell type dependent tumors such as pediatric brain tumors. Machine learning (ML) algorithms are widely used for this task due to their ability to identify complex patterns in large datasets accurately. However, inconsistencies in cell type annotations arise when comparing individual annotation tools. To address this issue, consensus approaches can be used to integrate the predictions of multiple ML models. This method produces more robust and accurate cell type annotations. The improved accuracy and robustness of cell type predictions demonstrate the feasibility and importance of consensus approaches in cancer cell type annotation. Ultimately, increasing the precision of cell type annotation may lead to a more comprehensive understanding of cancer heterogeneity and its implications for diagnosing and treating cancer.

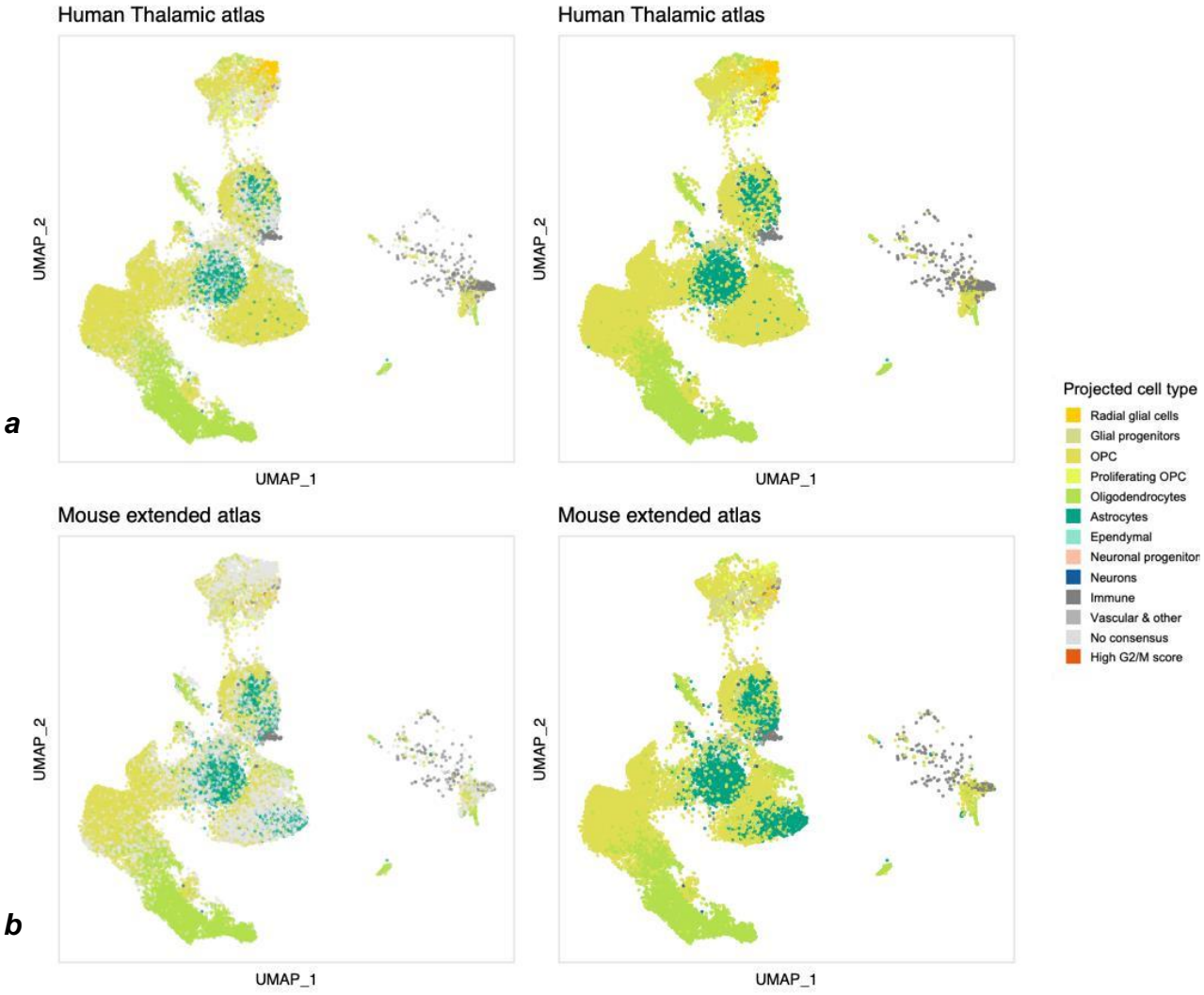


Figure 25 a. left: UMAP of malignant cells colored by consensus mouse projections including cells with no consensus. right: UMAP of malignant cells' mouse projections excluding cells with no consensus. b. left: UMAP of malignant cells colored by consensus human fetal projections including cells with no consensus. right: UMAP of malignant cells' human fetal projections excluding cells with no consensus

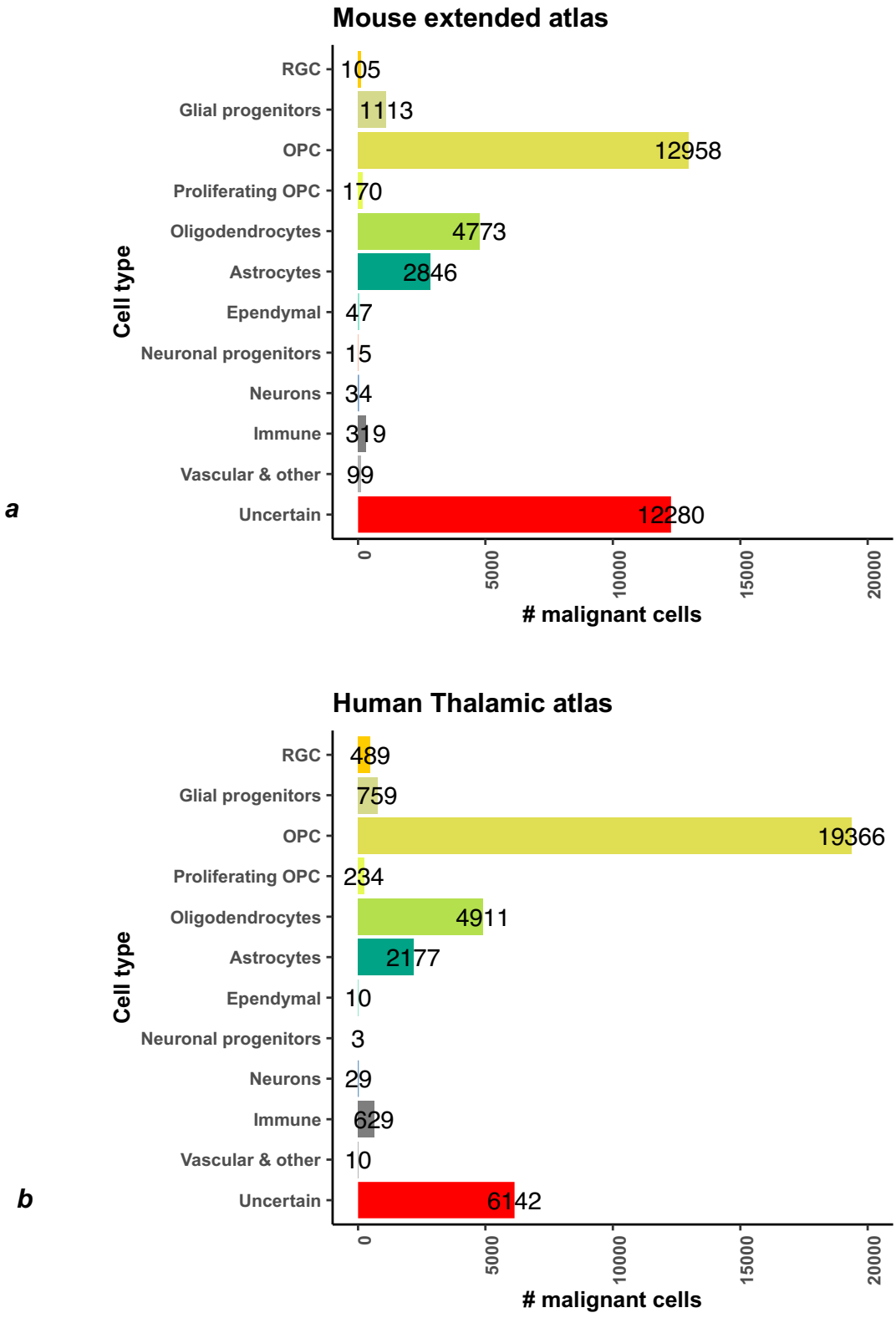


Figure 26 Quantification of cell type projections for H3.3K27M Thalamic malignant cells using a. mouse consensus projections b. using human fetal consensus projections

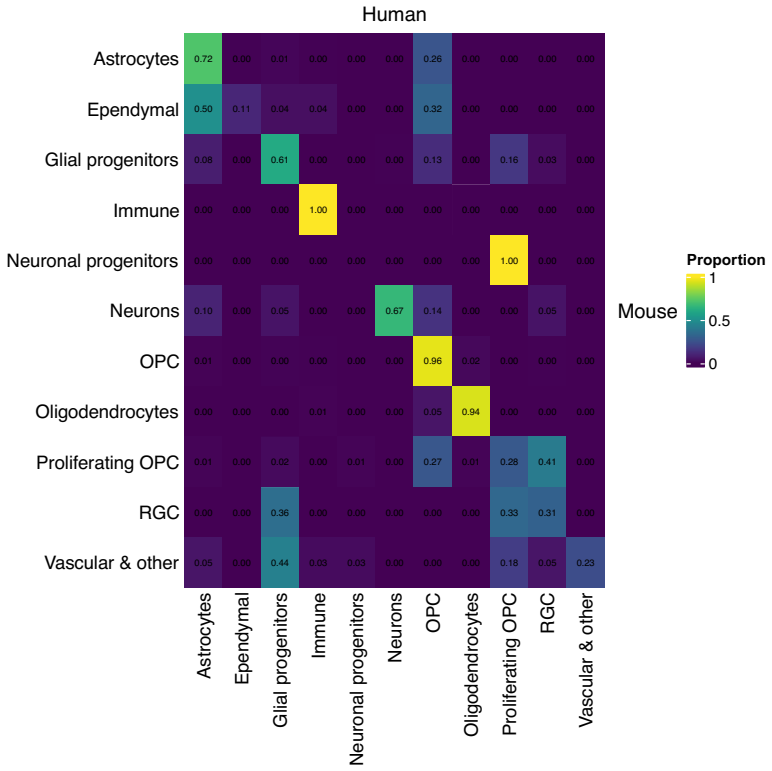


Figure 27 Confusion matrices comparing consensus projection labels obtained using a mouse prenatal reference and a human fetal atlas on H3.3K27M thalamic HGG. Proportions were computed row-wise and represent the fraction of cells from each mouse label which were assigned to each human label.

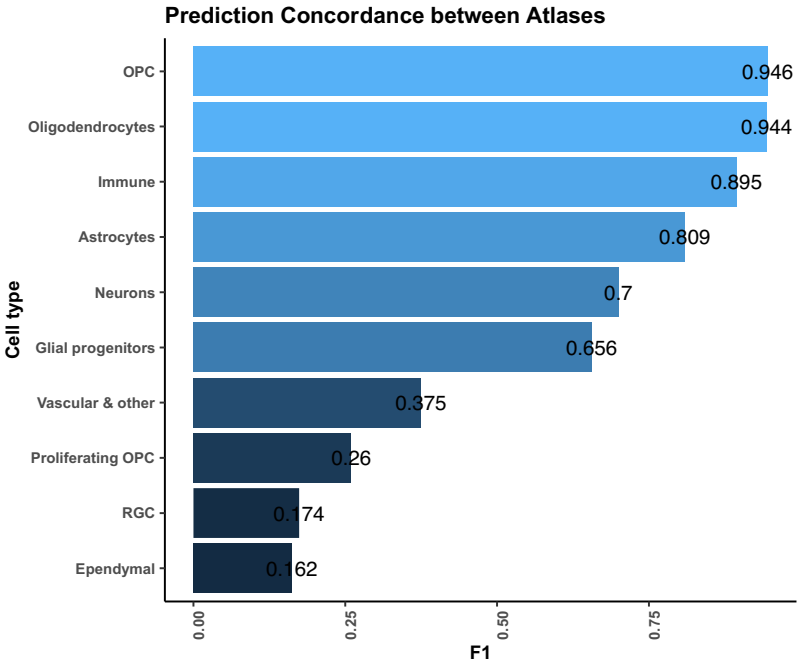


Figure 28 F1 scores per cell type when using human labels as ground truth and mouse labels as predictions.

5. Chapter V: Discussion

Single cell RNA sequencing has led to new discoveries in the field of genomics [25,27,32–34,61,156,157]. With the advent of increasingly large data sets and accessibility of sequencing, novel methods are required to analyze the data. In particular, the task of inferring cell type identity through the transcriptome of thousands of single cells lends itself to machine learning based approaches[70]. Machine learning-based cell type annotation became necessary in single cell genomics as the technology and amount of data continues to grow exponentially. Computational biologists are releasing new packages and algorithms to exploit abundant scRNA-seq data. As of 2023, more than 140 cell type annotation tools have been published or submitted to journals as per scRNA tools[87]. While these tools have shown some potential in individual studies, there are limited studies assessing the these tools on multiple, complex datasets containing tens of thousands of cells[100,103]. No studies in this field have systematically examined these tools using complex datasets, such as prenatal datasets, which pose significant challenges for annotation. Furthermore, most published studies did not explore how ML-classifier tools annotate in when reference and query datasets are derived from different species or tumors. These tasks are essential in the cancer biology field where animal models are frequently used, and human tissues may be limited.

Here, I present my work investigating how several published cell type annotation tools perform with complex datasets and their related challenges. First, I implemented several metrics to measure the complexity of a datasets according to three pre-defined characteristics (Chapter 3). When applying these metrics to prenatal and postnatal datasets, we observed that prenatal and mixed (containing both prenatal and postnatal data) datasets indeed show higher complexity than postnatal datasets. Next, I compared the performance of cell type annotation tools using datasets of varying complexity. We observed that classifiers performed less well when analyzing the more complex prenatal and mixed datasets. To address the limitations of individual classifiers, I designed a consensus-based cell type annotation pipeline (scCoAnnotate). We used this pipeline to efficiently classify cell types and study the performance of cell type annotation tools for cross-species cell type annotation (Chapter 4). Cross-species cell type annotation can

have a tremendous impact on our ability to leverage existing references, especially in fields that are reliant on animal models. Here, we demonstrated that most cell type annotation tools we analyzed performed poorly for cross-species predictions. We conclude that any single annotation method does not perform uniformly well across datasets of high complexity, such as prenatal datasets, or when annotating across species.

However, despite the poor performance of individual tools, our data indicates that using a consensus agreement between different tools can improve cell type annotation performance. To this end, we used scCoAnnotate to identify cell types in patient-derived high-grade glioma (HGG) data bearing H3.1 and H3.3 K27M histone mutations (Chapter 4). Cell type projections using a mouse reference dataset showed distinct oligodendroglial populations in this group of cancers [33,34,153,154]. These predictions were validated against a human fetal atlas. Overall, we find that most cell type annotation methods do not perform uniformly well across biologically distinct datasets. In the following sections I will discuss the challenges of complex prenatal data sets, cross-species annotations and cancer cell type annotation.

5.1. Intrinsic properties of prenatal datasets increase data complexity

Limited work has been done to objectively study dataset complexity in scRNA-seq for the task of annotation. Currently used metrics were not highly discriminant across datasets such as the inter-cluster similarity correlation metric which displayed similar values across datasets of various complexity [100]. Here, we aimed to describe the complexity of prenatal datasets in a supervised manner and according to pre-defined characteristics. We applied two metrics, N1 and F2, to quantify dataset complexity. N1 and F2 have previously been used as measures of dataset complexity in machine learning. For example, previous studies used these complexity metrics to optimize the complexity of synthetic datasets created for efficient classifier evaluation and parameterization[158]. Moreover, researchers also used N1 and F2 to increase the effectiveness of feature selection. In this case, including only features shown to decrease dataset complexity would theoretically lead to better classifier performance[159]. Some studies suggest changing classifiers based on

data complexity measures such as N1. High N1 values indicate complex class boundaries, which might necessitate nonlinear methods[160]. Using these metrics, we demonstrated that datasets derived from cells of prenatal tissues possess complex features, such as increased redundancy of cell types and gene expression programs across cell types, and a continuity of cell type and state. Using synthetic datasets, we saw that complexity metrics adapted to single-cell data capture their complexity and quantify both prenatal, postnatal and mixed datasets containing tissues from both pre- and postnatal tissues (Figure 3). We also examined the metric H1, a measure of lineage hierarchy; however, we not able to validate H1 using our synthetic datasets (Figure 3).

We measured the complexity of 5 prenatal, 4 mixed and 8 postnatal mouse and human datasets using the three metrics (Figures 4-6). Our results indicate that prenatal datasets possess significantly higher complexity than their postnatal counterparts. Cells in developing organs often exhibit plastic intermediate states in which they are continuously transitioning from one cell type to another. The existence of transient intermediate cell states presents a challenge for cell type annotation classifiers. All tools considered in this thesis must assign a discrete label to each cell. This limitation to the classifiers may lead to different labels for intermediate transitioning cells that express gene signatures from two cell types in a lineage. In turn, the discrete labeling of intermediate cells would lead to complex class boundaries and thus higher N1 values. Furthermore, differentiating cells, abundantly present in prenatal tissues, often express numerous genes that are common to multiple cell types. The co-expression of overlapping cell type specific gene signatures, measured by F2, further complicates classifier annotation. In the datasets we examined, the H1 metric, used to assesses the hierarchy, did not capture any significant hierarchical structure in prenatal or postnatal datasets. This result was surprising given it is well established that hierarchical lineages exist throughout brain development [161]. One possibility for the lack of detectable structures might be the stringent conditions we imposed in the H1 metric calculation. Hierarchical structures can be present in any dataset, and it is unclear if their presence is indicative of a true biological lineage.

Importantly, there are limitations to these dataset complexity metrics. Many factors can lead to an overestimation[103] or underestimation of the true dataset complexity. An example of a factor that may lead to overestimation of complexity is the sensitivity of N1 to label noise. In this case, it would be capture weak labeling instead of true biological complexity. Many of the datasets we have used in our benchmarking effort have some annotation limitations. One example is the cerebellum dataset[121] that had the identification and annotation of clusters performed jointly over all samples instead of on a per-sample basis. This method may lead to coarser labeling and hence more label noise, ultimately, affecting the complexity metrics. Nevertheless, these complexity metrics can still serve as a useful diagnostic measure to identify label noise and quality. Future investigation into dataset complexity assessment should tackle unsupervised metrics that are insensitive to external factors such as label quality, a common problem in scRNA-seq datasets that are labelled manually.

5.2. Cell type annotation tools show significant decreased performance when annotating prenatal datasets

Although some studies have assessed the performance of cell type annotation tools on scRNA-seq datasets[100,101,103], these studies did not include systematic benchmarking of cell type prediction tools using prenatal datasets. We showed that prenatal datasets are quantifiably more complex and hypothesized that cell type annotation tools will exhibit decreased performance when annotating these complex datasets. The median F1 scores, a measure of classifier performance, decreased significantly for classifiers such as scmap, correlation, and CHETAH, when annotating prenatal datasets (Figure 12). In contrast, other tools such as SVM and scPred perform relatively well on both pre- and postnatal dataset (Figure 12). Hierarchical tools such as scHPL and scClassify also performed well across the datasets but were slower than SVM. Generally, correlation and marker-based methods had a poor performance in the developmental setting. All classifiers were unable to accurately label certain prenatal datasets that were noted to have inadequate labeling (such as Cerebellum)[121]. The performance discrepancy between prenatal and postnatal datasets was further highlighted when we investigated mouse cortical data from development to maturity (E10 to P6). Neural network methods

failed in some instances could not outperform more straightforward approaches such as SciBet and SVM. This data is supported by a recent study that found that deep learning tools such as neural networks do not outperform classical ML tools for cell type annotation[99]. Overfitting of these neural networks on a relatively small number of training data or 'data greediness', in which neural networks require more data to generate proper models, may be responsible for this effect. In agreement with this hypothesis, we observed that NeuCA failed to converge in several datasets. Neural networks also need to be trained on a more significant number of parameters compared to a few kernel methods such as SVM. This requirement may also contribute to the lower performance of these tools when annotating prenatal datasets. Regardless, neural networks remain a promising tool for the task of cell type prediction. With the generation of larger training references, the training size bottleneck may become less of a problem and these tools will be able to be further exploited.

These findings confirm previous studies[100] where linear SVM with rejection was found to be the most consistent classifier across various scenarios. Although SVM rejection can refrain from labeling a proportion of the dataset (rejection), the proportion of rejected cells was not as high as other classifiers (such as scmap). Rejection, in many cases, can be a beneficial strategy to flag potentially false annotations and manually inspect them. Furthermore, SVM rejection showed the highest performance and confidence in the cells it annotated (>0.96 F1 score). An additional study further reinforced these findings by recommending linear SVM and logistic regression as the classifiers with the best performance and lowest run time when tested against nine other baseline classifiers[162]. SVMs are generally versatile classifiers that apply kernel functions to transform the data and then establish decision boundary hyperplanes that can be used to separate classes. Their ability to adapt to complex class borders with hyperparameter tuning makes them less likely to overfit. This is perhaps one of the reasons why in our study, SVM was the highest-performing tool for cell type annotation across a variety of dataset types including prenatal datasets which have not been used in prior benchmarking studies (Figures 12-13).

We also studied the effect of dataset complexity on performance (F1 scores and rejection rates). We found a significant correlation between complexity metrics and F1 scores for several classifiers such as NeuCA, scHPL, and scPred (Figures 14 and 15). Moreover, we found that rejection rates of classifiers also have a significant correlation with the N1 metric, which measures the fraction of borderline cells (Figure 16). This correlation may be due to the relationship between class or decision boundary complexity and model accuracy. Several studies have reported this positive correlation of boundary complexity with performance[163]. One survey of microarray data found that complexity measures, such as N1 that capture dataset complexity are correlated with classifier performance[164]. This study used feature-based complexity metrics such as F2. However, F2 was non-representative as it suffered from an underflow (0 value in all cases). This underflow was due to using all features in the metric calculations. Nevertheless, the authors recommended to use complexity metrics to analyze datasets before selecting a classifier. In our study, we used top 50 PCs, which has helped us avoid this underflow. We concluded that N1 and F2 captured substantial complexity that contributed to the decreased performance of cell type prediction tools when tasked with annotating prenatal datasets. These findings support our hypothesis that prenatal dataset characteristics can complicate annotation and hamper automated cell type annotation tools when they are trained on prenatal references.

While our work highlights the significant difference in performance between cell type annotation tools across prenatal and postnatal scRNA-seq dataset types, this benchmarking study has certain limitations. We used many prenatal and mixed datasets that our lab annotated. In some instances, using these datasets prevented external biases and confounders and allowed us to observe the effect of developmental age on performance. However, one caveat to this method is that we reduced the diversity of our data by generating different subsets from the same dataset (Jessa 2019). Limiting ourselves to a subset of one dataset, may limit the generalization of our benchmarking analysis to different tissues and time points. Using several prenatal datasets from different organs and sources should increase the generalizability and statistical significance of our findings. Another limitation of our study is the variance in the number of tools across

classifier classes. Certain classes of cell type annotation tools, such as the neural network methods, have been underrepresented. We included scIAE for the cortical time course assessment. While this classifier showed potential, we refrained from adding new tools to all of our analyses due to computational limitations and runtime errors. Using tools such as scIAE, scDeepSort, and other classifiers that incorporate prior knowledge such as gene regulatory networks will help us to better assess the performance of this class of classifiers. Our Snakemake-based pipeline ensures that addition of new tools can be easily incorporated for future studies.

Altogether, we found that cell type annotation tools perform exceptionally well in well-separated postnatal scRNA-seq datasets for which they were designed and tested. Nevertheless, we report that prenatal and mixed datasets possess complexity characteristics that hamper the performance of classifiers. We reaffirm previous works' findings that SVM outperforms other types of classifiers across different datasets. This result is interesting because the base linear SVM classifier outperforms tools that use SVM as scHPL and scPred. In the following sections, we will assess these classifiers in data-driven scenarios.

5.3. Cross species projections remain a challenge for automated cell type annotation

Cross-species cell type annotation is a promising application of automated cell type classifiers. The abundance of sequencing data from multiple species offers a unique opportunity to train classifiers on a larger number of high-quality references from multiple species. This application may be particularly important in fields where significant efforts are made to generate mouse models due to a lack of available human tissues. Human data is often more challenging to acquire due to both technical and ethical concerns. We are often limited by the size and number of samples we can obtain, a major obstacle to generating high quality data.

In contrast, mouse data is much more abundant, easier to sample and manipulate, and does not have as many ethical restrictions. The abundance of data generated by mouse

models would allow for finer annotations and the identification of new cell types, which can be used to annotate and interpret human data through cross-species annotation. Currently, some cell type annotation tools, such as SingleCellNet, promise good cross-species performance[76]. However, a systematic assessment of this application of cell type annotation has, to our knowledge, not been done. In our work, we aimed to assess the feasibility of cross-species annotation by testing 12 cell type annotation tools on paired mouse-human datasets from the adult pancreas and the adult cortex.

Our experimental data indicates that most classifiers have poor performance when annotating scRNA-seq data of one species using an scRNA-seq reference of another species (Figures 18-21). Although some tools, such as SciBet, SVM, and SingIR, performed relatively well, most tools did not correctly identify similar cell types. For example, CHETAH annotated human pancreatic immune cells as gamma cells. This annotation is not biologically relevant as these two cell types are distinct and not related through any known lineage. We suspect SciBet works well because it uses a feature selection technique to keep only cell type discriminant genes. This might have led to less overfitted models (i.e. with fewer features), which can adjust for significant reference-query differences[165]. However, most cell type annotation tools with a rejection option left the majority of cells unannotated in all four experiments. scPred, scmap-cell, scmap-cluster, and SVM rejection all performed poorly (Figure 20). This result is significant and shows that these tools have low confidence in their predictions and thus cannot assign a definitive cell type. Lastly, we utilized the consensus approach to obtain a majority vote on the cell type in each of the four experiments to annotate cells. We noticed significant high concordance with ground truth labels for cell types common between the reference and the query. As for the cell types that were not common, predictions were biologically relevant. For example, in the Allen brain datasets, human OPCs and microglia were annotated as mouse oligodendrocytes and macrophages, respectively. In this case, the consensus was to label the cells as the most similar available cell type. We observed that this consensus approach (Figure 22) outperformed any individual tool in almost all scenarios (Figures 18 to 21). We also noticed that many cell types such as microglia, macrophages, and OPCs were consistently well-annotated regardless of the classifier

(Figure 22). This effect might be due to the conserved transcriptomic gene expression controlling the identity of these cell types. In evolution, many cell types conserve vital features due to a conserved function[166,167].

There are several reasons that may explain the poor performance of cell type annotation tools for cross-species predictions. Firstly, many biological and experimental factors such as different experimental protocols, different sequencing depths, and different coverage can negatively affect the performance of a classifier and increase prediction uncertainty. For example, classifiers are usually trained only on the common features between a reference and a query dataset. Some of these factors can be adjusted for, such as batch effects, but certain factors, such as coverage and dropouts, remain a challenge. Secondly, cross-species cell type annotation is based on the orthologous genes between the species. Most tools utilize the one-to-one gene homolog between species, which might lead to the loss of critical cell type identity information conveyed by the one-to-many or many-to-one homolog that might have appeared during evolution due to gene duplications[168,169]. Thus, although homologous cells of different species often have conserved marker genes, they can still express necessary cell type-specific signatures that are lost when keeping only one-to-one homologous genes. In addition, cell type annotation tools often rely on quantifying the similarity of gene expression profiles which is sensitive to normalization and gene selection protocols[170]. Lastly, cell type divergence between species often occurs due to transcriptional changes of gene modules controlled by transcription factors which can have widespread effects on gene expression[170,171].

The consensus approach performed well in the cross-species setting. Consensus learning or vote is similar to ensemble methods that combine multiple base classifiers using bagging or boosting to produce one optimal solution. Random forests and boosted trees are one example of ensemble methods that perform well on multiple kinds of data[172,173]. The consensus approach is, however, different in that it uses independent and different types of classifiers to predict compared to the ensemble methods that use a collection of weak classifiers[174–176]. The consensus approach focuses on using heterogeneous classifiers to detect gene-cell type relationships. Rather than focus on one

representation of the data, we utilize classifiers that represent the data differently (neural networks, kernel methods, random forests) which should, in theory, explore the solution space more efficiently. If all classifiers agree on a prediction, this agreement should increase the confidence that the prediction is accurate. A consensus of different tools consistently outperforms a single classifier as it will average the solution and reduce the risk of choosing the wrong label. Moreover, most classifiers work by performing a local search for a solution and risk getting stuck in a local optimum. This limitation is avoided by having multiple starting points for the different classifiers[177]. This consensus approach has been used extensively in bioinformatics, particularly proteomics. For example, a consensus of Naïve Baye, random forests, and KNN was used to identify ligand binding to androgen receptors[178]. In addition, a consensus of SVM, random forests, and neural networks were used to predict protein-linked adverse drug reactions accurately[179]. The consensus approach was also used for non-bioinformatics applications such as tweet classification and had promising results[174].

Although our results show significant complexity in choosing a suitable classifier for cross-species annotation, we acknowledge there are some limitations to our experimental design. First, we used only four non-fully matching datasets for this experiment. It is difficult to attain statistical significance with this low number of datasets. Future studies focusing on incorporating more datasets from more species will be necessary to validate our findings. Human and mouse cell types are more divergent than human and primate ones[180]. As this divergence might also have a significant role in cross-species annotation[170], the inclusion of primate dataset may clarify some discrepancies that we observed in our experiments. Moreover, we only used 12 classifiers in this assessment due to computational restrictions. Some of the newer tools such as scGCN promise good cross-species performance using label transfer techniques[181]. Including these recent tools in our pipeline may also improve accuracy when annotating across species. Lastly, we only used one-to-one homologous genes in our training of the classifiers. This process potentially removes cell type-specific information, which can decrease classifiers' performance.

Altogether, our results support a consensus model for cell type annotation between species using scRNA-seq data. We find that most classifiers have difficulty identifying cell types between species, which is expected given the evolutionary barrier. Various factors may contribute to this problem. However, with the current expansion of cell type annotation tools, some tools are improving cross-species projections by using new techniques from deep learning. Future work identifying the strengths and weakness of these tools will help to move forward the technique of cross-species annotation.

5.4. Cell type identification in high grade gliomas using a consensus approach of automated classifiers

Histone mutant pediatric brain tumors are diseases of development thought to arise from cells that are stalled during development [33,34,153,154]. Identifying the cell types of origin in these cancers can help researchers to understand the underlying oncogenic mechanisms. This information would, in turn, help oncologists to exploit the weaknesses of these tumors using targeted therapy. Traditionally, cell types in cancer were identified using cell type markers and surface proteins[182,183]. However, with the advent of single-cell transcriptomics and the growth of single-cell references, automated approaches to identify cell types without manual curation became necessary. Here, we present an automated approach to identifying cell types in pediatric high-grade gliomas.

We used a consensus approach of three classifiers to project cell types in histone mutant high-grade gliomas and posterior fossa ependymomas (PFA-EP). As expected, the consensus projections identified ependymal-like cell types in PFA-EP, whereas H3.3K27M mutant HGGs showed oligodendroglial cell types (Figure 24), including OPC-like, oligodendrocyte-like, and astrocyte-like cell states[33,34,153,154]. The cell types identified for each tumor type are in agreement with published data. Interestingly, the H3.1K27M mutant showed a more divergent phenotype, with some samples showing ependymal-like states while others showed oligodendrocytic and astrocytic-like states. These findings were validated with other data analyses, including unsupervised data analysis and scATAC-seq data[34]. Importantly, our approach allowed for fast and objective cell type projections in cancer datasets.

We then sought to verify our cross-species mouse projections by comparing them to the projections obtained using a human fetal thalamic reference[143,144]. We performed this analysis on the thalamic subset of high-grade gliomas. Our findings showed significant concordance between the consensus labels obtained using a mouse reference and the labels obtained using the human fetal reference (>90% agreement, Figure 28). However, we saw a higher uncertainty rate (cells with no consensus cell type) using the mouse reference (2x the number of uncertain cells). This high percentage of unlabeled cells could be explained by the cross-species differences and the fact that these cells are malignant. Cell type annotation tools, as discussed above, encounter several challenges in the cross-species setting. In addition, malignant cells have aberrant expression of genes that can further distinguish them from the reference dataset.

The consensus approach implemented with scCoAnnotate delivered biologically relevant predictions validated using prenatal reference datasets. Although we only used three classifiers for cancer projections, the consensus approach outperformed the correlation method when used alone. Using cross-species references to annotate malignant cells may yield a high percentage of cells labeled as uncertain. However, we expect that expanding the consensus approach and choosing non-overfitting classifiers can help mitigate these drawbacks to yield accurate annotations of cancer datasets.

6. Chapter VI: Conclusions and future implications

6.1. Conclusions

scRNA-seq is a valuable tool that has revolutionized biomedical research. In single cell studies, identifying cell types in a sample is a primary objective that can be challenging. Traditional methods for cell type annotation using manual curation with marker genes found in the literature are exhausting and can lead to inconsistency in cell type labels across studies. Automated cell type annotation tools that implement machine-learning algorithms have been developed to address these challenges. However, automated cell type annotation tools have limitations in complex scenarios such as prenatal datasets, cross-species annotation, and cell type projection of malignant cells. In this thesis, we found that prenatal datasets are more complex than postnatal datasets, and cell type annotation tools had lower F1 scores and higher rejection rates for prenatal datasets. We also found that almost all cell type annotation tools perform poorly for cross-species annotation, and the consensus approach provided accurate cross-species annotations. Lastly, we used scCoAnnotate to project cell types in high-grade glioma cancer samples, and the consensus approach yielded accurate projections consistent with the known biology of these tumors. Overall, the thesis highlights the potential and limitations of automated cell type annotation in various complex scenarios and suggests that further development and improvements are necessary for accurate cell type annotation.

6.2. Future Directions

6.2.1. Applying unsupervised complexity measures

Our work sheds light on the complexity of prenatal and postnatal datasets and the impact of dataset complexity on the performance of cell type annotation tools. In this study, the method we used to score datasets for complexity involved supervised metrics that depend on the quality of labels. This dependency on label quality can lead to over- or under-estimation of complexity. This effect occurs, for example, with N1 when label noise is present. Therefore, for the purposes of our study, we assumed that the provided labels were the ground truth. However, this assumption might not be valid in all cases. To overcome this assumption, a future study could be designed using unsupervised

complexity metrics based on graph theory or unsupervised dimensionality reduction techniques, such as the average density of a graph. These unsupervised metrics may better capture the intrinsic properties of the dataset.

6.2.2. Diversifying prenatal datasets

In this thesis, I presented various analyses using all or a portion of 17 scRNA-seq datasets. These datasets were derived from different organs, different time points and different species. However, one caveat to our compilation of datasets is that the diversity of the data was higher for postnatal datasets as compared to the prenatal datasets. The postnatal datasets came from different studies, used different sequencing techniques and implemented various annotation schemes. In contrast, both the prenatal and mixed datasets were derived largely from the brain and were annotated using similar strategies by our lab members and collaborators. Moreover, we generated synthetic datasets from one dataset to be treated as distinct datasets. While this practice is helpful for our analyses, it may reduce our ability to generalize the findings. Although obtaining prenatal datasets is difficult, further studies incorporating a higher diversity of prenatal datasets from different organs with multiple annotation techniques, may increase the significance and build upon the findings of our study.

6.2.3. Incorporating more deep learning methods in the cross-validation and cross-species experiments

In our assessment, we included multiple classes of annotation tools and demonstrated that some classes of tools perform better than others in certain tasks. However, not all classes of tools were fully represented. Both computational restrictions and the need to input prior knowledge such as gene regulatory networks limited tool selection. We included only two distinct classifiers in the neural network class. This underrepresentation limits our conclusions about neural network performance for cell annotation. In our study, neural network classifiers did not perform well when analyzing complex datasets. However, these deep learning techniques are in early stages of development and are beginning to be applied and tested more frequently to analyze scRNA-seq data. Recent studies use both adversarial and recursive neural networks for cell type annotation.

Studies using these recent tools demonstrate that they have good cross-species performance and label transfer but require large training datasets and prior knowledge that can aid the annotation process[181,184]. Given their performance in cross-species annotation, expanding our analysis and pipeline to include these tools, even if it requires extracting prior knowledge from the literature, would improve upon our study. Deep learning approaches can drive scientific discoveries in biology, as illustrated by Alphafold2, an artificial intelligence-based tool that predicts protein structure[185]. Given their potential, a systematic assessment of deep learning cell annotation tools is essential.

6.2.4. Updating and improving the scCoAnnotate pipeline

In this thesis, we designed a parallelized and automated cell type annotation pipeline that offers flexibility, speed, and ease of use for bioinformaticians. This pipeline was used extensively in this study and has demonstrated the ability to deliver accurate annotations of large datasets using minimal resources and requiring little user intervention. Going forward, this pipeline can be updated by incorporating additional tools as they become available. Furthermore, adding custom consensus protocols that includes or excludes particular tools based on the task being performed may improve the accuracy of annotations. Moreover, adding plotting and post-analysis modules that can output the predictions and perform validity analysis can streamline the cell type annotation, ultimately making this process more objective and systematic.

7. Chapter VII: 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 Med.* **9**, 1–12 (2017).
2. Li, M. *et al.* DISCO: a database of Deeply Integrated human Single-Cell Omics data. *Nucleic Acids Res.* **50**, (2022).
3. Marioni, J. C. & Arendt, D. How Single-Cell Genomics Is Changing Evolutionary and Developmental Biology. *Annu. Rev. Cell Dev. Biol.* **33**, 537–553 (2017).
4. Hedlund, E. & Deng, Q. Single-cell RNA sequencing: Technical advancements and biological applications. *Molecular Aspects of Medicine* (2018) doi:10.1016/j.mam.2017.07.003.
5. Eberwine, J. *et al.* Analysis of gene expression in single live neurons. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 3010 (1992).
6. Brady, G., Barbara, M. & Iscove, N. N. Representative in Vitro cDNA Amplification From Individual Hemopoietic Cells and Colonies. *METHODS Mol. Cell. Biol.* **2**, 17–25 (1990).
7. Tietjen, I. *et al.* Single-cell transcriptional analysis of neuronal progenitors. *Neuron* **38**, 161–175 (2003).
8. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *DETAILS ONLINE* **6**, (2008).
9. Mortazavi, A., Williams, B. A., Mccue, K., Schaeffer, L. & Wold, B. RECEIVED 2 May; aCCEPTED 27 May; PUBLISHED ONLINE. *Nat. MeThodS* | **5**, 621 (2008).
10. Liu, Y., Fan, Z., Qiao, L. & Liu, B. Advances in microfluidic strategies for single-cell research. *TrAC Trends Anal. Chem.* **157**, 116822 (2022).
11. Tavakoli, H. *et al.* Recent advances in microfluidic platforms for single-cell analysis incancer biology, diagnosis and therapy. *Trends Analyt. Chem.* **117**, 13 (2019).
12. Rakszewska, A., Tel, J., Chokkalingam, V. & Huck, W. T. S. One drop at a time: toward droplet microfluidics as a versatile tool for single-cell analysis. *NPG Asia Mater. 2014 610* **6**, e133–e133 (2014).
13. Espina, V. *et al.* Laser-capture microdissection. *Nat. Protoc.* **1**, 586–603 (2006).

14. Hashimshony, T. *et al.* CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 1–7 (2016).
15. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
16. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
17. McGinnis, C. S. *et al.* MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods 2019 167* **16**, 619–626 (2019).
18. Regev, A. *et al.* The Human Cell Atlas. *bioRxiv* (2017) doi:10.1101/121202.
19. Domínguez Conde, C. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science (80-.).* **376**, (2022).
20. Ardini-Poleske, M. E. *et al.* The Molecular Atlas of Lung Development Program. *Am J Physiol Lung Cell Mol Physiol* **313**, 733–740 (2017).
21. The Lancet Neurology. The International Brain Initiative: collaboration in progress. *Lancet Neurol.* **20**, 969 (2021).
22. Sunkin, S. M. *et al.* Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. doi:10.1093/nar/gks1042.
23. Bhaduri, A., Nowakowski, T. J., Pollen, A. A. & Kriegstein, A. R. Identification of cell types in a mouse brain single-cell atlas using low sampling coverage. doi:10.1186/s12915-018-0580-x.
24. Nguyen, Q. H. *et al.* Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. *Nat. Commun. 2018 91* **9**, 1–12 (2018).
25. Busslinger, G. A. *et al.* Human gastrointestinal epithelia of the esophagus, stomach, and duodenum resolved at single-cell resolution. *Cell Rep.* **34**, 108819 (2021).
26. Molnár, Z. *et al.* New insights into the development of the human cerebral cortex. *J. Anat.* **235**, 432 (2019).
27. Nelson, A. C., Mould, A. W., Bikoff, E. K. & Robertson, E. J. Single-cell RNA-seq reveals cell type-specific transcriptional signatures at the maternal–foetal interface during pregnancy. *Nat. Commun. 2016 71* **7**, 1–12 (2016).
28. Keren-Shaul, H. *et al.* A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease. *Cell* **169**, (2017).

29. Sobue, A. *et al.* Microglial gene signature reveals loss of homeostatic microglia associated with neurodegeneration of Alzheimer's disease. *Acta Neuropathol. Commun.* **9**, (2021).
30. Olah, M. *et al.* Single cell RNA sequencing of human microglia uncovers a subset associated with Alzheimer's disease. *Nat. Commun. 2020 111* **11**, 1–18 (2020).
31. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332 (2019).
32. Chen, C. C. L. *et al.* Histone H3.3G34-Mutant Interneuron Progenitors Co-opt PDGFRA for Gliomagenesis. *Cell* (2020) doi:10.1016/j.cell.2020.11.012.
33. Jessa, S. *et al.* Stalled developmental programs at the root of pediatric brain tumors. *Nat. Genet.* (2019) doi:10.1038/s41588-019-0531-7.
34. Jessa, S. *et al.* K27M in canonical and noncanonical H3 variants occurs in distinct oligodendroglial cell lineages in brain midline gliomas. *Nat. Genet.* **54**, 1865–1880 (2022).
35. Chen, H. *et al.* Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. doi:10.1038/s41467-019-09670-4.
36. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol. 2019 375* **37**, 547–554 (2019).
37. He, Z., Peng, C., Li, T. & Li, J. Cell Differentiation Trajectory in Liver Cirrhosis Predicts Hepatocellular Carcinoma Prognosis and Reveals Potential Biomarkers for Progression of Liver Cirrhosis to Hepatocellular Carcinoma. *Front. Genet.* **13**, 527 (2022).
38. Karaayvaz, M. *et al.* Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. doi:10.1038/s41467-018-06052-0.
39. Bridges, K. & Miller-Jensen, K. Mapping and Validation of scRNA-Seq-Derived Cell-Cell Communication Networks in the Tumor Microenvironment. *Front. Immunol.* **0**, 1765 (2022).
40. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).
41. Bacher, R. & Kendziorski, C. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.* **17**, 1–14 (2016).

42. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **50**, 96 (2018).
43. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, 8746 (2019).
44. Liao, J. *et al.* Single-cell RNA sequencing of human kidney. *Sci. Data 2020 71* **7**, 1–9 (2020).
45. Zheng, G. X. Y. *et al.* ARTICLE Massively parallel digital transcriptional profiling of single cells. (2017) doi:10.1038/ncomms14049.
46. Jiang, P. Quality control of single-cell RNA-seq. *Methods Mol. Biol.* **1935**, 1–9 (2019).
47. Wu, Y. & Zhang, K. Tools for the analysis of high-dimensional single-cell RNA sequencing data. *Nat. Rev. Nephrol. 2020 167* **16**, 408–421 (2020).
48. Hong, R. *et al.* Comprehensive generation, visualization, and reporting of quality control metrics for single-cell RNA sequencing data. *Nat. Commun. 2022 131* **13**, 1–9 (2022).
49. Imoto, Y. *et al.* Resolution of the curse of dimensionality in single-cell RNA sequencing data analysis. *Life Sci. alliance* **5**, (2022).
50. Heimberg, G., Bhatnagar, R., El-Samad, H. & Thomson, M. Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. *Cell Syst.* **2**, 239–250 (2016).
51. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093–1098 (2013).
52. Jollife, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **374**, (2016).
53. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
54. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol. 2018 371* **37**, 38–44 (2018).
55. Andrews, T. S. & Hemberg, M. Identifying cell populations with scRNASeq. *Mol. Aspects Med.* **59**, 114–122 (2018).
56. Zhang, X. *et al.* CellMarker: a manually curated resource of cell markers in human

and mouse. *Nucleic Acids Res.* **47**, D721–D728 (2019).

57. Franzén, O., Gan, L. M. & Björkegren, J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database* **2019**, (2019).
58. Wang, J. *et al.* Tracing cell type evolution by cross-species comparison of cell atlases. *Cell Rep.* **34**, 108803 (2021).
59. Elmentaite, R., Domínguez Conde, C., Yang, L. & Teichmann, S. A. Single-cell atlases: shared and tissue-specific cell types across human organs. *Nat. Rev. Genet. 2022 237* **23**, 395–410 (2022).
60. Ilicic, T. *et al.* Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* **17**, 1–15 (2016).
61. Levitin, H. M. *et al.* De novo gene signature identification from single-cell RNA-seq with hierarchical Poisson factorization. *Mol. Syst. Biol.* **15**, e8557 (2019).
62. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* (2019) doi:10.15252/msb.20188746.
63. Dong, D. *et al.* Structural basis of assembly of the human T cell receptor–CD3 complex. *Nat. 2019 5737775* **573**, 546–552 (2019).
64. Young, M. D. *et al.* Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science (80-.).* **361**, 594–599 (2018).
65. Savage, P. *et al.* A Targetable EGFR-Dependent Tumor-Initiating Program in Breast Cancer. *Cell Rep.* **21**, 1140–1149 (2017).
66. Lambrechts, D. *et al.* Phenotype molding of stromal cells in the lung tumor microenvironment. *Nat. Med.* **24**, 1277–1289 (2018).
67. Darmanis, S. *et al.* Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma. *Cell Rep.* **21**, 1399–1410 (2017).
68. Duò, A., Robinson, M. D. & Soneson, C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research* **7**, 1141 (2018).
69. Wang, Z., Ding, H. & Zou, Q. Identifying cell types to interpret scRNA-seq data: how, why and more possibilities. *Brief. Funct. Genomics* **19**, 286–291 (2020).
70. Pasquini, G., Rojo Arias, J. E., Schäfer, P. & Busskamp, V. Automated methods for cell type annotation on scRNA-seq data. *Comput. Struct. Biotechnol. J.* **19**, 961–

969 (2021).

71. Ma, F. & Pellegrini, M. ACTINN: Automated identification of cell types in single cell RNA sequencing. *Bioinformatics* (2020) doi:10.1093/bioinformatics/btz592.
72. Li, C. *et al.* SciBet as a portable and fast single cell type identifier. *Nat. Commun.* (2020) doi:10.1038/s41467-020-15523-2.
73. 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 Biol.* **20**, 1–17 (2019).
74. Johnson, T. S. *et al.* LAmbDA: label ambiguous domain adaptation dataset integration reduces batch effects and improves subtype detection. *Bioinformatics* **35**, 4696 (2019).
75. Lieberman Id, Y., Rokach, L. & Shay, T. CaSTLe – Classification of single cells by transfer learning: Harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. (2018) doi:10.1371/journal.pone.0205499.
76. SingleCellNet: A Computational Tool to Classify Single Cell RNA-Seq Data Across Platforms and Across Species | Elsevier Enhanced Reader. https://reader.elsevier.com/reader/sd/pii/S2405471219301991?token=DDCF263B5C31FCD97402BE32D2DD73DF89FE04EDCE6E5685A94175DB8654D7B7F92AA3FAF89C695EBC7EA833C0BE3285&originRegion=us-east-1&originCreation=20220209194557.
77. Schölkopf, B. SVMs - A practical consequence of learning theory. *IEEE Intell. Syst. Their Appl.* **13**, 18–21 (1998).
78. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol. 2019 202* **20**, 163–172 (2019).
79. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods 2018 155* **15**, 359–362 (2018).
80. Pliner, H. A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *Nat. Methods 2019 1610* **16**, 983–986 (2019).
81. Zhang, Z. *et al.* SCINA: A Semi-Supervised Subtyping Algorithm of Single Cells and Bulk Samples. *Genes (Basel).* **10**, 531 (2019).

82. Domanskyi, S., Hakansson, A., Bertus, T. J., Paternostro, G. & Piermarocchi, C. Digital Cell Sorter (DCS): a cell type identification, anomaly detection, and Hopfield landscapes toolkit for single-cell transcriptomics. *PeerJ* **9**, e10670 (2021).
83. Michielsen, L., Reinders, M. J. T. & Mahfouz, A. Hierarchical progressive learning of cell identities in single-cell data. *Nat. Commun. 2021 121* **12**, 1–12 (2021).
84. Bernstein, M. N. & Dewey, C. N. Annotating cell types in human single-cell RNA-seq data with CellO. *STAR Protoc.* **2**, (2021).
85. Lin, Y. *et al.* scClassify: sample size estimation and multiscale classification of cells using single and multiple reference. *Mol. Syst. Biol.* **16**, (2020).
86. de Kanter, J. K., Lijnzaad, P., Candelli, T., Margaritis, T. & Holstege, F. C. P. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Res.* **47**, e95 (2019).
87. Zappia, L., Phipson, B. & Oshlack, A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLOS Comput. Biol.* **14**, e1006245 (2018).
88. Ng, A. Y. & Jordan, M. I. On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes.
89. Li, Z. & Feng, H. A neural network-based method for exhaustive cell label assignment using single cell RNA-seq data. *Sci. Reports 2022 121* **12**, 1–12 (2022).
90. Ongsulee, P. Artificial intelligence, machine learning and deep learning. *Int. Conf. ICT Knowl. Eng.* 1–6 (2018) doi:10.1109/ICTKE.2017.8259629.
91. Fortelny, N. & Bock, C. Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. *Genome Biol.* **21**, 1–36 (2020).
92. Yin, Q., Wang, Y., Guan, J. & Ji, G. scIAE: an integrative autoencoder-based ensemble classification framework for single-cell RNA-seq data. *Brief. Bioinform.* **23**, (2022).
93. Sarker, I. H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput. Sci.* **2**, 1–21 (2021).
94. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA sequencing data are corrected by matching mutual nearest neighbours.

Nat. Biotechnol. **36**, 421 (2018).

95. Zhang, Y., Zhang, F., Wang, Z., Wu, S. & Tian, W. scMAGIC: accurately annotating single cells using two rounds of reference-based classification. *Nucleic Acids Res.* **50**, E43 (2022).
96. Zeebaree, D. Q., Haron, H. & Abdulazeez, A. M. Gene Selection and Classification of Microarray Data Using Convolutional Neural Network. *ICOASE 2018 - Int. Conf. Adv. Sci. Eng.* 145–150 (2018) doi:10.1109/ICOASE.2018.8548836.
97. Imoto, Y. *et al.* Resolution of the curse of dimensionality in single-cell RNA sequencing data analysis. *Life Sci. alliance* **5**, (2022).
98. Debie, E. & Shafi, K. Implications of the curse of dimensionality for supervised learning classifier systems. *Pattern Anal. Appl.* **22**, 519–536 (2019).
99. Köhler, N. D., Büttner, M., Andriamanga, N. & Theis, F. J. Deep learning does not outperform classical machine learning for cell type annotation. *bioRxiv* 653907 (2021) doi:10.1101/653907.
100. Abdelaal, T. *et al.* A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.* **20**, 1–19 (2019).
101. Ma, W., Su, K. & Wu, H. Evaluation of some aspects in supervised cell type identification for single-cell RNA-seq: classifier, feature selection, and reference construction. *Genome Biol.* **22**, 1–23 (2021).
102. Huang, Q., Liu, Y., Du, Y. & Garmire, L. X. Evaluation of Cell Type Annotation R Packages on Single-cell RNA-seq Data. *Genomics. Proteomics Bioinformatics* **19**, 267 (2021).
103. Xie, B., Jiang, Q., Mora, A. & Li, X. Automatic cell type identification methods for single-cell RNA sequencing. *Comput. Struct. Biotechnol. J.* **19**, 5874–5887 (2021).
104. Zhao, X., Wu, S., Fang, N., Sun, X. & Fan, J. Evaluation of single-cell classifiers for single-cell RNA sequencing data sets. *Brief. Bioinform.* **21**, 1581–1595 (2020).
105. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst.* **3**, 346-360.e4 (2016).
106. Muraro, M. J. *et al.* A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst.* **3**, 385-394.e3 (2016).

107. Tian, L. *et al.* Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat. Methods 2019 166* **16**, 479–487 (2019).
108. Mulas, R. & Casey, M. J. Estimating cellular redundancy in networks of genetic expression.
109. Cardoso-Moreira, M. *et al.* Gene expression across mammalian organ development. *Nature* **571**, 505 (2019).
110. Chen, R. C., Dewi, C., Huang, S. W. & Caraka, R. E. Selecting critical features for data classification based on machine learning methods. *J. Big Data* **7**, 1–26 (2020).
111. Weinreb, C., Wolock, S., Tusi, B. K., Socolovsky, M. & Klein, A. M. Fundamental limits on dynamic inference from single-cell snapshots. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E2467–E2476 (2018).
112. Van den Berge, K. *et al.* Trajectory-based differential expression analysis for single-cell sequencing data. *Nat. Commun. 2020 111* **11**, 1–13 (2020).
113. Song, D. & Li, J. J. PseudotimeDE: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell RNA sequencing data. *Genome Biol.* **22**, 1–25 (2021).
114. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498 (2015).
115. Casey, M. J., Stumpf, P. S. & MacArthur, B. D. Theory of cell fate. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **12**, (2020).
116. Stanley, G., Gokce, O., Malenka, R. C., Südhof, T. C. & Quake, S. R. Continuous and Discrete Neuron Types of the Adult Murine Striatum. *Neuron* **105**, 688-699.e8 (2020).
117. Tasic, B. *et al.* Shared and distinct transcriptomic cell types across neocortical areas. *Nat. 2018 5637729* **563**, 72–78 (2018).
118. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 (2014).
119. Sachewsky, N., Xu, W., Fuehrmann, T., van der Kooy, D. & Morshead, C. M. Lineage tracing reveals the hierarchical relationship between neural stem cell populations in the mouse forebrain. *Sci. Reports 2019 91* **9**, 1–10 (2019).
120. Eze, U. C., Bhaduri, A., Haeussler, M., Nowakowski, T. J. & Kriegstein, A. R. Single-

cell atlas of early human brain development highlights heterogeneity of human neuroepithelial cells and early radial glia. *Nat. Neurosci. 2021 244* **24**, 584–594 (2021).

121. Vladoiu, M. C. *et al.* Childhood cerebellar tumours mirror conserved fetal transcriptional programs. *Nature* (2019) doi:10.1038/s41586-019-1158-7.
122. Monje, M. *et al.* Hedgehog-responsive candidate cell of origin for diffuse intrinsic pontine glioma. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 4453–4458 (2011).
123. Tirosh, I. *et al.* Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**, 309 (2016).
124. Fan, J., Slowikowski, K. & Zhang, F. Single-cell transcriptomics in cancer: computational challenges and opportunities. *Exp. Mol. Med. 2020 529* **52**, 1452–1465 (2020).
125. Han, Y., Liu, D. & Li, L. PD-1/PD-L1 pathway: current researches in cancer. *Am. J. Cancer Res.* **10**, 727 (2020).
126. Dohmen, J. *et al.* Identifying tumor cells at the single-cell level using machine learning. *Genome Biol. 2022 231* **23**, 1–23 (2022).
127. Mahin, K. F. *et al.* PanClassif: Improving pan cancer classification of single cell RNA-seq gene expression data using machine learning. *Genomics* **114**, 110264 (2022).
128. Wei, I. H., Shi, Y., Jiang, H., Kumar-Sinha, C. & Chinnaiyan, A. M. RNA-Seq Accurately Identifies Cancer Biomarker Signatures to Distinguish Tissue of Origin. *Neoplasia* **16**, 918 (2014).
129. Wilson, C. M., Fridley, B. L., Conejo-Garcia, J. R., Wang, X. & Yu, X. Wide and deep learning for automatic cell type identification. *Comput. Struct. Biotechnol. J.* **19**, 1052 (2021).
130. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* (2011).
131. Duan, B. *et al.* Learning for single-cell assignment. *Sci. Adv.* **6**, (2020).
132. Lin, Y. *et al.* scClassify: sample size estimation and multiscale classification of cells using single and multiple reference. *Mol. Syst. Biol.* **16**, 9389 (2020).
133. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and

Model Selection. (1995).
134. Lorena, A. C., Garcia, L. P. F., Lehmann, J., Souto, M. C. P. & Ho, T. K. How Complex is your classification problem? A survey on measuring classification complexity. (2021).
135. Maillo, J., Triguero, I. & Herrera, F. Redundancy and Complexity Metrics for Big Data Classification: Towards Smart Data. *IEEE Access* **8**, 87918–87928 (2020).
136. Suzuki, R. & Shimodaira, H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22**, 1540–1542 (2006).
137. Mölder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Research* **10**, (2021).
138. Anderson, A. G., Kulkarni, A., Harper, M. & Konopka, G. Single-Cell Analysis of Foxp1-Driven Mechanisms Essential for Striatal Development. *Cell Rep.* **30**, 3051-3066.e7 (2020).
139. Tasic, B. *et al.* Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72 (2018).
140. Mizrak, D. *et al.* Single-Cell Analysis of Regional Differences in Adult V-SVZ Neural Stem Cell Lineages. *Cell Rep.* **26**, 394-406.e5 (2019).
141. Dong, R. *et al.* Single-Cell Characterization of Malignant Phenotypes and Developmental Trajectories of Adrenal Neuroblastoma. *Cancer Cell* **38**, 716-733.e6 (2020).
142. Pijuan-Sala, B. *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis. *Nat. 2019 5667745* **566**, 490–495 (2019).
143. Eze, U. C., Bhaduri, A., Haeussler, M., Nowakowski, T. J. & Kriegstein, A. R. Single-cell atlas of early human brain development highlights heterogeneity of human neuroepithelial cells and early radial glia. *Nat. Neurosci. 2021 244* **24**, 584–594 (2021).
144. Bhaduri, A. *et al.* An atlas of cortical arealization identifies dynamic molecular signatures. *Nat. 2021 5987879* **598**, 200–204 (2021).
145. Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).
146. Yao, Z. *et al.* A taxonomy of transcriptomic cell types across the isocortex and

hippocampal formation. *Cell* **184**, 3222-3241.e26 (2021).

147. Ostrom, Q. T. *et al.* CBTRUS Statistical Report: Primary Brain and Central Nervous System Tumors Diagnosed in the United States in 2008-2012. *Neuro. Oncol.* **17 Suppl 4**, iv1–iv62 (2015).
148. Louis, D. N. *et al.* The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol.* **131**, 803–820 (2016).
149. Louis, D. N. *et al.* The 2007 WHO Classification of Tumours of the Central Nervous System. *Acta Neuropathol.* **114**, 97 (2007).
150. Jones, C. *et al.* Pediatric high-grade glioma: biologically and clinically in need of new thinking. *Neuro. Oncol.* **19**, 153–161 (2017).
151. Faury, D. *et al.* Molecular profiling identifies prognostic subgroups of pediatric glioblastoma and shows increased YB-1 expression in tumors. *J. Clin. Oncol.* **25**, 1196–1208 (2007).
152. Deshmukh, S., Ptack, A., Krug, B. & Jabado, N. Oncohistones: a roadmap to stalled development. *FEBS J.* **289**, 1315–1328 (2022).
153. Monje, M. *et al.* Hedgehog-responsive candidate cell of origin for diffuse intrinsic pontine glioma. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 4453–4458 (2011).
154. Filbin, M. G. *et al.* Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. *Science* **360**, 331–335 (2018).
155. Jessa, S. K27M in canonical and noncanonical H3 variants occurs in distinct oligodendroglial cell lineages in brain midline gliomas. *Nat. Genet.* (2022).
156. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nat. 2019 5707761* **570**, 332–337 (2019).
157. Li, Y. *et al.* Identification of Specific Cell Subpopulations and Marker Genes in Ovarian Cancer Using Single-Cell RNA Sequencing. *Biomed Res. Int.* **2021**, (2021).
158. Fraca, T. R., Miranda, P. B. C., Prudencio, R. B. C., Lorenaz, A. C. & Nascimento, A. C. A. A Many-Objective optimization Approach for Complexity-based Data set Generation. *2020 IEEE Congr. Evol. Comput. CEC 2020 - Conf. Proc.* (2020) doi:10.1109/CEC48606.2020.9185543.
159. Okimoto, L. C., Savii, R. M. & Lorena, A. C. Complexity Measures Effectiveness in

Feature Selection. (2017) doi:10.1109/BRACIS.2017.66.

160. Garcia, L. P. F., Lorena, A. C., De Souto, M. C. P. & Ho, T. K. Classifier Recommendation Using Data Complexity Measures. *Proc. - Int. Conf. Pattern Recognit.* **2018-August**, 874–879 (2018).
161. Raj, B. *et al.* Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol. 2018 365* **36**, 442–450 (2018).
162. Huang, Y. & Zhang, P. Evaluation of machine learning approaches for cell type identification from single-cell transcriptomics data. *Brief. Bioinform.* **22**, 1–11 (2021).
163. Atashpaz-Gargari, E., Sima, C., Braga-Neto, U. M. & Dougherty, E. R. Relationship between the accuracy of classifier error estimation and complexity of decision boundary. *Pattern Recognit.* **46**, 1315–1322 (2013).
164. Morán-Fernández, L., Bolón-Canedo, V. & Alonso-Betanzos, A. Can classification performance be predicted by complexity measures? A study using microarray data. *Knowl. Inf. Syst.* **51**, 1067–1090 (2017).
165. Decuyper, M., Stockhoff, M., Vandenberghe, S., -, al & Ying, X. An Overview of Overfitting and its Solutions You may also like Artificial neural networks for positioning of gamma interactions in monolithic PET detectors Analysis of overfitting in the regularized Cox model An Overview of Overfitting and its Solutions. 22022 (2019) doi:10.1088/1742-6596/1168/2/022022.
166. Affinati, A. H. *et al.* Cross-species analysis defines the conservation of anatomically segregated VMH neuron populations. *Elife* **10**, (2021).
167. Gargareta, V.-I. *et al.* Conservation and divergence of myelin proteome and oligodendrocyte transcriptome profiles between humans and mice. *Elife* **11**, (2022).
168. Gabaldón, T. & Koonin, E. V. Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.* **14**, 360 (2013).
169. Magadum, S., Banerjee, U., Murugan, P., Gangapur, D. & Ravikesavan, R. Gene duplication as a major force in evolution. *J. Genet.* **92**, 155–161 (2013).
170. Liu, X., Shen, Q. & Zhang, S. Cross-species cell type assignment of single-cell RNA-seq by a heterogeneous graph neural network. *bioRxiv* 2021.09.25.461790 (2021) doi:10.1101/2021.09.25.461790.

171. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
172. Song, Y. Y. & Lu, Y. Decision tree methods: applications for classification and prediction. *Shanghai Arch. Psychiatry* **27**, 130 (2015).
173. Shahzad, R. K., Fatima, M., Lavesson, N. & Boldt, M. Consensus decision making in random forests. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **9432**, 347–358 (2015).
174. Cui, R., Agrawal, G., Ramnath, R. & Khuc, V. Ensemble of Heterogeneous Classifiers for Improving Automated Tweet Classification. *IEEE Int. Conf. Data Min. Work. ICDMW* **0**, 1045–1052 (2016).
175. Velusamy, D. & Ramasamy, K. Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset. *Comput. Methods Programs Biomed.* **198**, (2021).
176. Alshdaifat, E., Al-hassan, M. & Aloqaily, A. Effective heterogeneous ensemble classification: An alternative approach for selecting base classifiers. *ICT Express* **7**, 342–349 (2021).
177. Plewczynski, D. Brainstorming: Consensus Learning in Practice. *Front. Neuroinform.* **3**, (2009).
178. Grisoni, F., Consonni, V. & Ballabio, D. Machine Learning Consensus To Predict the Binding to the Androgen Receptor within the CoMPARA Project. *J. Chem. Inf. Model.* **59**, 1839–1848 (2019).
179. Galletti, C., Aguirre-Plans, J., Oliva, B. & Fernandez-Fuentes, N. Prediction of Adverse Drug Reaction Linked to Protein Targets Using Network-Based Information and Machine Learning. *Front. Bioinforma.* **0**, 70 (2022).
180. Shami, A. N. *et al.* Single-Cell RNA Sequencing of Human, Macaque, and Mouse Testes Uncovers Conserved and Divergent Features of Mammalian Spermatogenesis. *Dev. Cell* **54**, 529-547.e12 (2020).
181. Song, Q., Su, J. & Zhang, W. scGCN is a graph convolutional networks algorithm for knowledge transfer in single cell omics. *Nat. Commun.* **12**, (2021).
182. Racila, E. *et al.* Detection and characterization of carcinoma cells in the blood. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 4589 (1998).

183. Kikuta, K. *et al.* Pancreatic stellate cells promote epithelial-mesenchymal transition in pancreatic cancer cells. *Biochem. Biophys. Res. Commun.* **403**, 380–384 (2010).
184. Lotfollahi, M. *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol. 2021 401* **40**, 121–130 (2021).
185. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nat. 2021 5967873* **596**, 583–589 (2021).