TUM SCHOOL OF LIFE SCIENCES

TECHNISCHE UNIVERSITÄT MÜNCHEN

Charting the neural organoid landscape and the impact of glucocorticoids on human brain development through single-cell genomics

Leander Paul Friedrich Dony

March 2024

HELMHOLTZ MUNICH

MAX PLANCK INSTITUTE
OF PSYCHIATRY

Technische Universität München
TUM School of Life Sciences

Charting the neural organoid landscape and the impact of glucocorticoids on human brain development through single-cell genomics

Leander Paul Friedrich Dony

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Markus List

Prüfende der Dissertation:

1. Prof. Dr. Dr. Fabian J. Theis
2. Prof. Dr. Dr. Elisabeth B. Binder
3. Prof. Dr. Alexander Meissner

Die Dissertation wurde am 27.03.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 02.10.2024 angenommen.

Acknowledgements

The work presented in this thesis was only possible with the support, collaboration, motivation, and input of countless kind, caring, and intelligent people. I want to thank everyone who has been part of this journey, whether or not they are explicitly named in this section.

First and foremost, I would like to thank my supervisors, Fabian Theis and Elisabeth Binder, for their ongoing guidance and support during my PhD. Being part of your labs has provided me with an exceptional learning environment. I greatly benefitted from your scientific input, our ongoing research discussions, your flexibility, your patience, and the academic freedom you gave me. Furthermore, I would like to thank Maria Colomé-Tatché for being part of my thesis advisory committee and providing valuable input on the multi-omics aspects of my projects. I want to thank Markus List for chairing my examination committee. I am very grateful to the ICB and MPI admin teams. Anna, Sabine, Dani, Karina, and Heike, your ongoing support made my PhD life so much easier.

I have been fortunate to profit from the knowledge, experience, and mentorship of several senior ICB and MPI colleagues throughout my PhD. Thank you, Cristiana, for getting me up to speed with organoids, glucocorticoids and psychiatry and coordinating the wet lab work that generated all our data. Thank you, David, for introducing me to the world of single-cell method development, explaining DE analysis over and over again, mentoring me, and being a constant source of project ideas. Thank you, Anthi and Simone, for helping me understand neuroscience and bioinformatics during many insightful discussions and contributing to the wet and dry lab side of my projects. Thank you, Malte, for being a reference for literally everything related to single-cell analysis.

I am grateful to the IMPRS-TP research school and the Joachim Herz Foundation for providing additional courses, funding, networking opportunities, and scientific exchange. One such transforming and memorable experience was my visit to the Treutlein and Camp labs in Basel. Thank you, Barbara and Gray, for hosting me and enabling this fantastic and fun learning experience with the Munich/Basel Organoid Atlassing team. Thank you both, Zhisong and Jonas, for sharing your organoid and neurobiology wisdom with me and introducing the whole Theislab Organoid Atlas team to Basel.

I want to thank my great colleagues and friends at the ICB and MPI. In particular, Marius, Giovanni, Leon, Laura, Lisa, Anna, Lea, Anna-Lena and Dominik. I really enjoyed the time we spent together during cooking evenings, bike rides, ski trips, science discussions, home office Wednesdays, or HIIT morning sessions. I am also so grateful for all my friends outside of work who created countless unforgettable memories and made me stay sane during the past years, including the current and former *Rubinsalamander*, the *Ottolenghi Gang*, and many, many more: you are amazing. Thank you, Mama und Papa, for always being there for me and unconditionally supporting me in every possible way for the last 30 years. Lea - thank you for being by my side and always making me feel at home, no matter where we are.

Abstract

Fetal neurodevelopment is a complex and tightly controlled process where disruptions can affect postnatal health. Single-cell transcriptomic technologies have enabled the detailed molecular characterisation of such dynamic developmental processes over the past years. The curation of single-cell transcriptomics data into harmonised cellular reference atlases has further unveiled the diversity of cellular states in the human brain. While the quantity of single-cell transcriptomics datasets has been growing continuously, data curation remains a laborious and largely manual task. Neural organoids have emerged as powerful three-dimensional models of human neurodevelopment, recapitulating the intricate interplay of different cellular states and fate transitions to generate complex neural tissue landscapes *in vitro*. Despite the potential and the increasing adoption of organoids as model systems, a comprehensive neural organoid reference is lacking to evaluate their shortcomings and strengths. A reference of this kind could additionally serve as the starting point for gaining deeper insights into human neurodevelopment. An in-depth understanding of human-specific neurodevelopmental processes forms the foundation for studying the onset and development of many psychiatric and neurodevelopmental disorders and ultimately developing novel treatments. While genetic risk factors for these diseases have been widely studied, the mechanisms underlying the effects of environmental risk factors are still poorly understood.

I address the abovementioned challenges through three interconnected contributions in this thesis. First, I develop a data curation tool for more efficient construction of cellular reference atlases. Second, using this tool, I construct a neural organoid reference atlas to uncover the current limitations of this model system and ease contextualising new organoid datasets from health and disease. Last, using the organoid atlas as a reference, I analyse the effects of glucocorticoid (GC) exposure on brain development in organoids to improve our understanding of the environment's impact on brain development and disease.

My first contribution introduces a data and model zoo called sfaira. It facilitates the large-scale curation and reuse of single-cell genomics data and enables their automated analysis with pre-trained end-to-end models. The sfaira software provides an easily extendable and structured

collection of single-cell transcriptomic datasets using controlled metadata vocabularies and an interactive interface for streamlined data curation. These curation functionalities of sfaira greatly facilitated the collection and streamlining of the datasets for the Human Neural Organoid Cell Atlas (HNOCA), the second contribution of this thesis. The HNOCA is a large-scale integrated reference atlas that captures the cellular diversity from 36 neural organoid datasets based on 26 differentiation protocols. In a comparison with fetal brain data, the HNOCA provides insight into cellular states under-represented in neural organoids, highlights systematic transcriptomic differences between organoids and primary tissue and quantifies the transcriptomic fidelity of different organoid cell types. I showcase the value of using the HNOCA as a reference dataset for evaluating novel organoid differentiation protocols and contextualising a new organoid dataset, modelling the effects of environmental risk for psychiatric diseases on neurodevelopment, the third contribution of this thesis. In this dataset, I use synthetic glucocorticoids (sGCs), a commonly prescribed type of drug in pregnancies at risk for preterm birth, as an example of a widely encountered prenatal environmental risk factor for psychiatric disease. I find that chronic GC exposure in organoids affects the transcription of critical neurodevelopmental genes and predominantly primes the development of the inhibitory neuron lineage, resulting in an increased abundance of inhibitory neurons. I identify PBX3 as a possible mediator of the GC-induced lineage priming effect.

The work presented in this thesis advances the state of the art in single-cell genomics data infrastructure, improves our understanding of neural organoids as a model system, and provides insight into the molecular mechanisms mediating the effects of environmental disease risk on human brain development. Altogether, this work underscores the power of integrating large-scale single-cell transcriptomics data and neural organoid technologies to unravel the complexities of neurodevelopmental processes and their disruption in disease. It paves the way towards improved disease modelling in organoids, with the potential to support novel therapeutic strategies in psychiatry and beyond.

Zusammenfassung

Die pränatale Gehirnentwicklung ist ein komplexer und präzise gesteuerter Prozess, dessen Störung die spätere Gesundheit beeinträchtigen kann. Einzelzell-Transkriptomik-Technologien haben in den letzten Jahren eine detaillierte molekulare Charakterisierung solcher dynamischer Entwicklungsprozesse zugänglich gemacht. Harmonisierte Referenzatlanten aus Einzelzell-Transkriptomdaten ermöglichen es, die zelluläre Vielfalt des menschlichen Gehirns zu erfassen. Während die Menge der Einzelzell-Transkriptomik-Datensätze kontinuierlich zunimmt, bleibt die Kuratierung der Daten eine mühsame und weitgehend manuelle Aufgabe. Neurale Organoide haben das Potenzial, wirkungsvolle dreidimensionale Modelle der menschlichen Gehirnentwicklung zu sein. Sie bilden das vielschichtige Zusammenspiel verschiedener Zellzustände und Entwicklungsstadien ab und generieren komplexe neurale Gewebelandschaften. Trotz dieses Potenzials und der zunehmenden Verbreitung von Organoidmodellen fehlt ein umfassender Referenz-Datensatz, welcher deren Defizite und Stärken aufzeigt und weiterhin einen tieferen Einblick in die menschliche Gehirnentwicklung geben könnte. Ein tiefgehendes Verständnis der zugrunde liegenden neurobiologischen Prozesse ist wiederum entscheidend für die Entwicklung neuer Therapien psychiatrischer Erkrankungen und neurologischer Entwicklungsstörungen. Während die genetischen Risikofaktoren für diese Krankheiten bereits Gegenstand vieler Untersuchungen sind, ist unser Verständnis der Wirkmechanismen umweltbedingter Risikofaktoren ungleich geringer.

In dieser Arbeit adressiere ich die oben beschriebenen Forschungslücken durch drei ineinandergreifende Beiträge. Erstens entwickle ich ein Werkzeug zur Datenkuratierung, das eine effizientere Erstellung von zellulären Referenzatlanten ermöglicht. Zweitens verwende ich dieses Werkzeug zur Konstruktion eines neuralen Organoid-Referenzatlas. Dieser hilft einerseits, derzeitige Grenzen des Modellsystems aufzuzeigen und andererseits neue Organoid-Datensätze in Kontext zu setzen. Drittens nutze ich diesen Atlas als Referenz, um in einem weiteren Experiment den Effekt von Glukokortikoiden auf die Gehirnentwicklung in Organoiden zu analysieren. Dadurch trage ich zu einem besseren Verständnis der Wirkungsmechanismen bei, welche den umweltbedingten Risiken für psychiatrische Erkrankungen zugrunde liegen.

In meinem ersten Beitrag stelle ich die eigens entwickelte Daten- und Modellbibliothek sfaira vor. Sie unterstützt die automatisierte Datenanalyse mit zuvor trainierten End-to-End-Modellen. Weiterhin ermöglicht sfaira es, umfangreiche Einzelzell-Genomikdaten zu kuratieren und somit wiederverwendbar zu machen. Die sfaira Software bietet eine leicht erweiterbare strukturierte Datenbank mit definierten Metadatenvokabularen für Einzelzell-Transkriptomdaten sowie eine interaktive Schnittstelle für eine effiziente Datenkuratierung. Diese Kuratierungsfunktionen erleichterten es erheblich, die Datensätze für den HNOCA, den zweiten Beitrag dieser Arbeit, zusammenzustellen. Der HNOCA ist ein umfassender integrierter Referenzatlas, der die zelluläre Vielfalt von 36 neuralen Organoid-Datensätzen auf der Grundlage von 26 Differenzierungsprotokollen erfasst. Durch den direkten Vergleich mit fetalen Gehirndaten gibt der HNOCA Einblick in zelluläre Zustände, die in neuralen Organoiden aktuell noch nicht ausreichend repräsentiert sind. Weiterhin hebt der Vergleich systematische transkriptomische Unterschiede zwischen Organoiden und Primärgewebe hervor und quantifiziert die transkriptomische Güte der verschiedenen Organoid-Zelltypen. Ich demonstriere den Nutzen der Verwendung des HNOCA als Referenz-Datensatz, um neuartige Organoiddifferenzierungsprotokolle zu bewerten. Analog verwende ich den HNOCA, um einen neuen Organoid-Datensatz zu kontextualisieren, welcher die Auswirkungen umweltbedingter Risiken für psychiatrische Erkrankungen auf die Neuroentwicklung modelliert. In dieser Modellierung, dem dritten Beitrag dieser Arbeit, verwende ich synthetische Glukokortikoide, ein häufig verschriebenes Medikament in Schwangerschaften mit Frühgeburtsrisiko, als Beispiel für einen weitverbreiteten pränatalen Umweltrisikofaktor für psychiatrische Erkrankungen. Ich zeige, dass eine chronische Glukokortikoid-Exposition in Organoiden die Transkription wichtiger neurologischer Entwicklungsgene beeinflusst. Außerdem fördert diese überwiegend die Entwicklung inhibitorischer Nervenzellen, was wiederum zu einer erhöhten Anzahl selbiger Zellen führt. Ich identifiziere das Gen PBX3 als einen möglichen Vermittler dieses Glukokortikoid-induzierten selektierenden Effekts.

Zusammenfassend entwickelt die hier vorgestellte Arbeit die verfügbare Infrastruktur für Einzelzell-Genomikdaten weiter, verbessert unser Verständnis neuraler Organoide als Modellsystem und gibt Einblick in die molekularen Wirkmechanismen umweltbedingter Risikofaktoren auf die menschliche Gehirnentwicklung. Somit unterstreicht diese Arbeit die Bedeutung der Zusammenführung umfangreicher Einzelzell-Transkriptomdaten und neuraler Organoid-Technologien, um die Komplexität neurologischer Entwicklungsprozesse und deren Störung durch Krankheiten zu entschlüsseln. Sie ebnet den Weg für eine verbesserte organoidbasierte Krankheitsmodellierung mit dem Potenzial, neue therapeutische Strategien in der Psychiatrie und darüber hinaus zu ermöglichen.

Contents

1. Introduction

Understanding the development of the human brain, with its immense complexity, remains a prominent challenge in biomedical research. Studying the intricate mechanisms of neurodevelopment and how aberrations can lead to disease requires sophisticated models and finely resolved cellular profiling approaches. In this thesis, I aim to bridge this gap by using single-cell genomics technologies to deepen our understanding of neural organoids as models of the developing brain and to investigate the impact of an environmental risk factor for psychiatric disease on neurodevelopment.

I structure this introduction into four parts. First, I provide an overview of single-cell transcriptomics data generation, their computational analysis and curation into large-scale cellular atlases. Next, I introduce neural organoid models of human neurodevelopment and explore recent related single-cell ribonucleic acid sequencing (scRNA-seq) studies. I then describe how the environment can impact human brain development and touch upon GC exposure as an example of an early-life environmental challenge. I consequently introduce *in vitro* models of the developing brain that could be employed to study the effect of such environmental challenges. Lastly, I summarise the main research questions addressed by this thesis and outline my specific scientific contributions.

1.1. Single-cell transcriptomics

As the fundamental unit of all living organisms, every cell performs specific roles that contribute to an organism's overall functioning. The disruption of cellular processes can lead to illnesses, highlighting the importance of cellular studies in understanding health and disease. For example, over 100 genes can affect the risk for autism spectum disorder (ASD) by impairing critical neurodevelopmental pathways when disrupted [1]. In another example, lissencephaly, disruptions in one of only a few relevant genes can lead to severely altered brain structure due to malfunctioning cellular pathways underlying neuronal migration [2].

Cells are categorised into two types: prokaryotic and eukaryotic. While prokaryotic cells are the building blocks of bacteria and archaea, eukaryotic cells comprise all animals, plants, and fungi. Unlike prokaryotic cells, eukaryotic cells contain a nucleus beside their various organelles, concentrating the cell's genetic material, the DNA. In protein biosynthesis, genetic information stored in the nucleus is transcribed into messenger RNA (mRNA) and subsequently translated into proteins. Proteins are essential for most cellular functions, and cells continuously synthesise and degrade them to adapt their function to the specific requirements of their current environment and state. Quantifying the amount of mRNA per gene present in a cell can, therefore, inform about a cell's current identity and function.

The first approaches to quantifying mRNA molecules from individual cells and selected genes used, for example, single-cell quantitative polymerase chain reaction (PCR) [3, 4] or manual dissection of fixed tissues combined with reverse northern blotting [5] and were later followed by first micro-fluidic devices [6, 7]. In 2009, sequencing technologies enabled the first successful whole-transcriptome scRNA-seq study of seven individual murine primordial germ cells [8].

1.1.1. Single-cell transcriptomics technologies

Since 2009, technological advances have allowed for a dramatic scaling of scRNA-seq experiments, recently reaching three to four million cells per study [9, 10]. Despite these tremendous improvements in scale, a sizeable fraction of each mRNA quantification protocol is identical across approaches [11]. In the first step, as many mRNA molecules in a cell as possible are converted to complementary DNA (cDNA) via reverse transcription, followed by synthesis of the second strand. Next, the cDNA is amplified by either PCR [12] or *in vitro* transcription [3, 13]. In the tagmentation step, adapters essential for downstream sequencing are integrated, thus completing the single-cell library generation process (Fig. 1.1).

A critical differentiation among generated libraries is between full-length and tag-based approaches. Full-length protocols capture entire mRNA transcripts, enabling in-depth analyses such as alternative splicing and allele-specific expression [14]. In contrast, tag-based methods selectively target the mRNA's 3′ or 5′ ends. This is achieved by targeting the mRNA using oligo-dT primers for the 3′ end, which bind to the poly-A tail of the polyadenylated mRNA, or specific primers for the 5′ end, which bind to the capped start of the transcript. While tag-based methods do not allow for in-depth analyses such as isoform detection and hinder unambiguous mapping to the genome [15], they offer higher throughput and cost efficiency.

Importantly, tag-based protocols also enable the use of unique molecular identifiers (UMIs). These short nucleotide sequences are added to the cDNA before the amplification step, serving as a unique label for each transcript [16]. Using UMIs in single-cell protocols mainly helps to combat amplification bias, where specific sequences are preferentially duplicated in the amplification step [17], but also improves the accuracy of downstream data normalisation [18]. While the use of UMIs and capturing full-length transcripts were initially mutually exclusive, several newer approaches have been developed to accommodate both features [19, 20].

Beyond the distinction between full-length and tag-based scRNA-seq protocols, cell separation is the main differentiating factor between the different approaches to constructing a single-cell library. The commercial Fluidigm C1 platform, which uses a microfluidic chip for cell separation, has been employed together with some of the early protocols, including CEL-seq2 [21] and Smart-seq version 1 [22]. The most commonly used protocols nowadays are either based on microwell plates, where each cell is placed into a separate well, or microfluidic droplet techniques, primarily using water droplets suspended in oil as reaction chambers. Examples of plate-based protocols include MARS-seq [23], the commercial ICELL8 solution [24], Smart-seq2 [25], and Smart-seq3 [20]. Droplet-based protocols include, for example, inDrop [26], Drop-seq [27], and 10x Chromium [28], a prevalent commercial platform supporting a large number of scRNA-seq experiments at the time of writing this thesis. In addition to these approaches, split-pool barcoding protocols such as sci-RNA-seq3 [29] have recently allowed for cost-efficient scaling of scRNA-seq experiments to up to four million cells [9].

Over the past years, single-cell transcriptomics has become a ubiquitous approach in biomedical research. It could be considered a successor to bulk RNA sequencing as it offers cell- instead of sample-level resolution and enables, for example, unbiased detection of cell identities and previously unavailable insights into cellular heterogeneity [30]. It should be noted, however, that bulk RNA sequencing is still the preferred method in situations where cellular heterogeneity is not the focus of the study and reduced noise levels are required as bulk readouts are generally less prone to noise given that they can be considered an average over many cells [31].

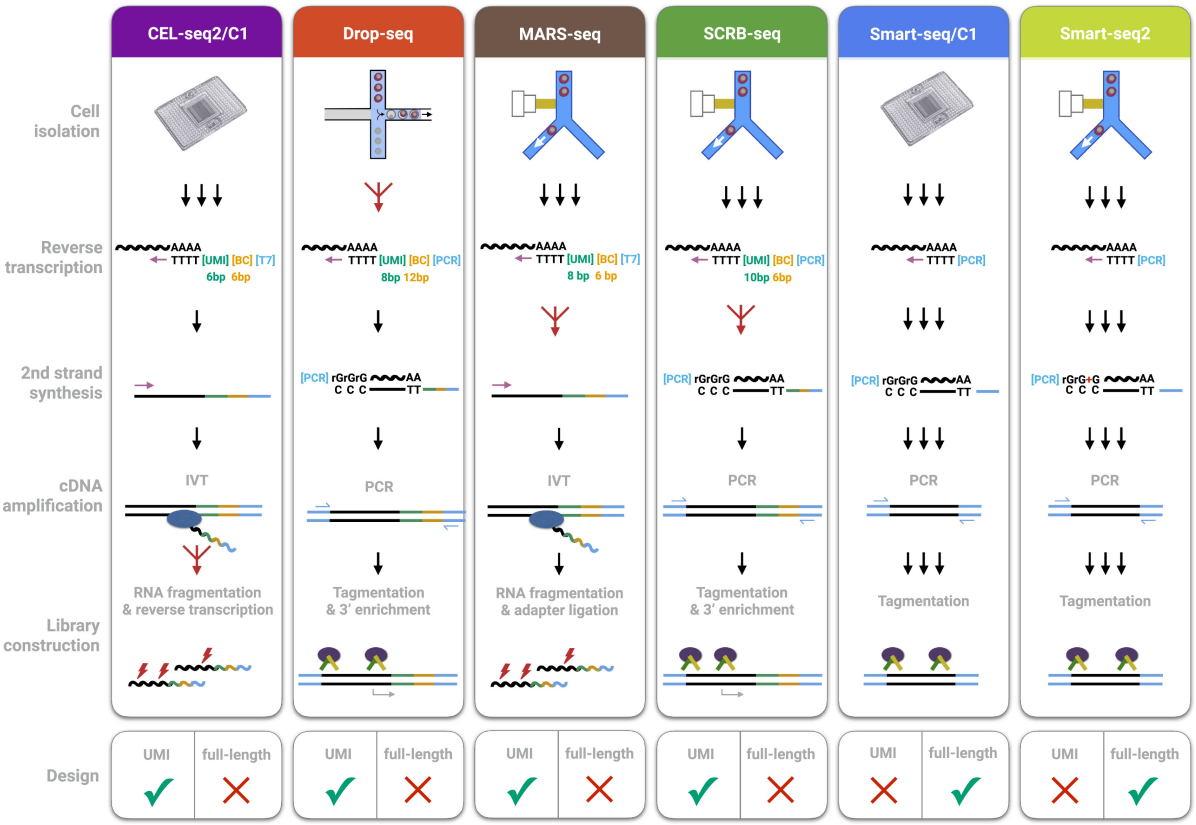

Figure 1.1.: **Single-cell transcriptomics protocols.** Selected protocols and their constituent steps used for scRNA-seq library preparation. The data presented in this thesis is mostly generated using the 10x Genomics Chromium platform, which is similar to the Drop-seq protocol, and the Smart-seq2 protocol for full-length transcript capture. IVT, *in vitro* transcription. UMI, unique molecular identifier. Reproduced from [32] with permission from Elsevier.

1.1.2. Computational analysis of single-cell transcriptomics data

The quantitative analysis of single-cell transcriptomics data is generally based on a cell-by-gene count matrix. Each entry represents the count of mRNA detections for a specific gene in a particular cell. A count matrix is obtained by demultiplexing the raw sequencing output, aligning sequencing reads to a reference genome or transcriptome, and quantifying the number of reads mapping to each gene in a cell-specific manner via the cell barcodes. Pipeline tools such as the 10x Genomics Cell Ranger software [28], STARsolo [33], or kallisto | bustools [34] efficiently automate these steps in a user-friendly manner.

Analysing scRNA-seq data requires specialised tools due to its high dimensionality, scale,

and complexity. While the most-used analysis tools in Python (scanpy [35]) and R (Seurat [36]) have been cited many thousand times (see `https://www.scrna-tools.org/table`), there are now over 1000 individual computational tools available, addressing all aspects of scRNA-seq analysis [37]. This proliferation partly reflects the complexity and diversity of the data and the studied biological questions but is also a result of multiple competing tools being developed for individual applications. This has created a need for structured benchmark studies, evaluating the performance and suitability of competing tools for a given task. Heumos and Schaar [38] offer an overview of the latest benchmarking results across tasks and derive a comprehensive list of current best-practise recommendations in single-cell genomics data analysis.

The abovementioned tools, scanpy and Seurat, are prominent representatives of pipeline-based analysis tools for scRNA-seq data. These pipelines typically encompass several steps: quality control (QC) to remove low-quality samples, normalisation to adjust for cell size and sequencing depth, feature selection to reduce noise, batch correction to mitigate effects from different sample origins and technical confounders, and dimensionality reduction to simplify data complexity while retaining essential information (Fig. 1.2a). [39]. While such a structured approach offers excellent flexibility and is widely adopted in the community, it requires careful tuning at each step to ensure optimal results and is not readily transferred to new datasets.

Downstream analysis steps generally focus on identifying cellular structure and attributing variance in the data to biological mechanisms. Cellular structure is mainly determined by grouping cells with similar transcriptomic profiles into clusters and using known marker genes to assign each cluster a cell type label (Fig. 1.2b). Depending on the biological question, attributing variance to biological mechanisms can take multiple forms. A common question is on the cell type-specific differences in gene expression between two groups of cells, such as treatment and control cells (differential expression (DE) analysis). The identified significant differentially expressed genes (DEGs) are often functionally annotated by statistically testing for enrichment of known functional gene sets (gene set enrichment). Testing for differences in composition, for example, between donors or treatment groups, evaluating covarying transcriptional changes between samples (cell-cell communication) or genes (gene regulatory network (GRN) inference), or predicting transcriptional effects of unseen perturbations (perturbation modelling) are other commonly applied downstream analyses (Fig. 1.2c) [38].

Autoencoder (AE)-based analysis approaches, such as scVI [40] or DCA [41], replace parts of the manual preprocessing and feature selection steps in scRNA-seq data analysis. These

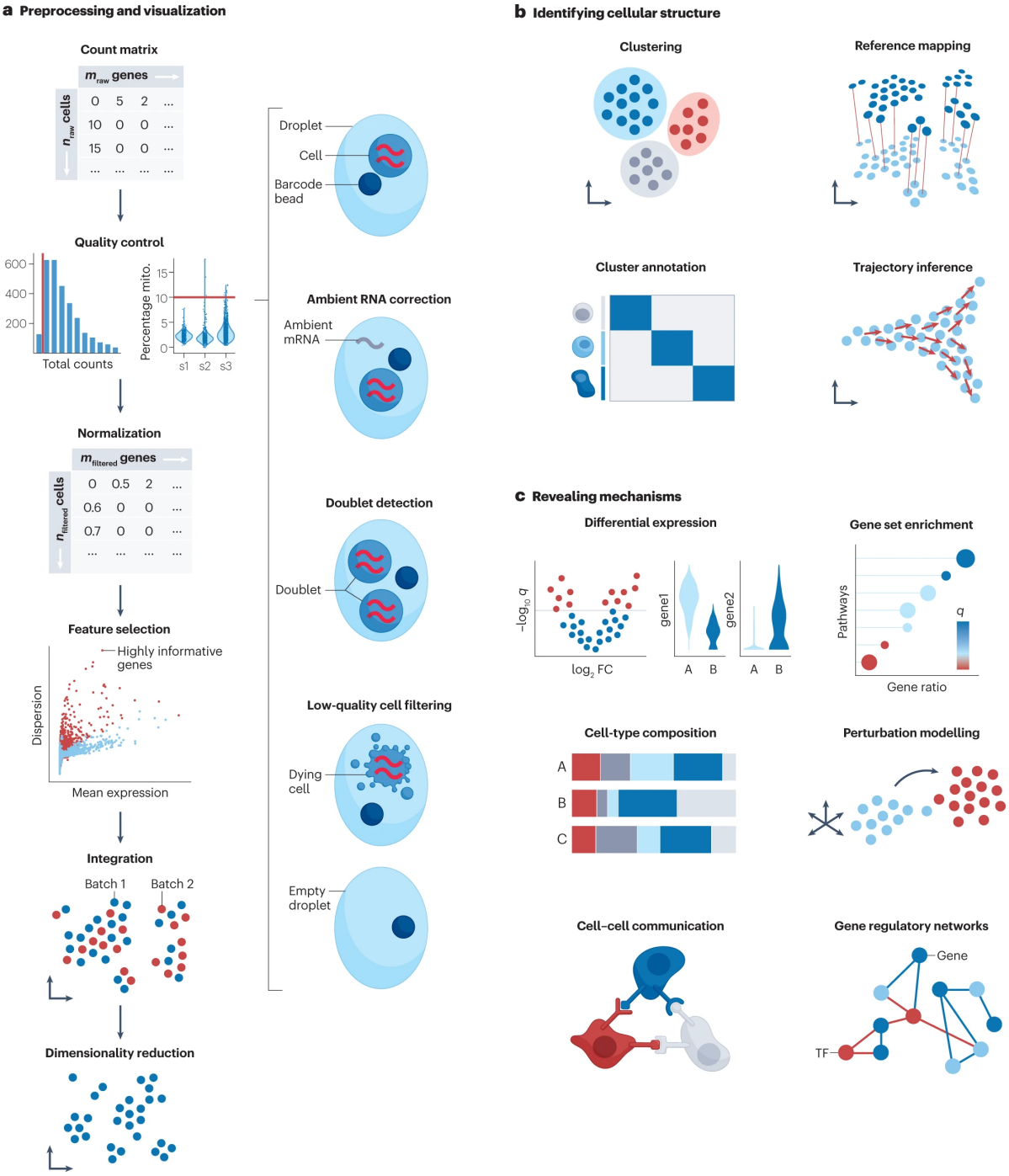

Figure 1.2.: **Example single-cell transcriptomics analysis pipeline.** **a.** Common preprocessing and batch correction steps of a scRNA-seq analysis pipeline. **b.** Typical components of cellular structure identification in scRNA-seq downstream analysis. **c.** Selected approaches for mechanistic interpretation and functional annotation of scRNA-seq data. Reproduced from [38] with permission from Springer Nature.

methods produce a low-dimensional latent representation directly from the raw transcript counts while implicitly performing preprocessing tasks, including normalisation, feature selection, and batch correction. While some (variational) AE-based approaches additionally offer certain downstream analysis functionality (for example, DE analysis in scVI [40]), many common downstream steps, such as cell type identification, DE analysis, or compositional analyses, are generally carried out in the same manual way as described above for pipeline approaches (Fig. 1.2c).

Notably, unlike pipeline approaches, parametric AE models can be, by design, distributed pre-trained on extensive datasets, facilitating the analysis of novel data. This has been widely adopted through the query-to-reference mapping paradigm. This powerful approach allows mapping novel datasets onto existing reference datasets for their analysis [42]. As such, query-to-reference mapping enables the rapid contextualisation of new data against a backdrop of well-characterised cell types or states, providing a framework for interpretation and discovery. Recently, foundational models have been suggested [43] [44] as a new concept in single-cell genomics data analysis. These large, usually transformer-based models are pre-trained on transcriptional data from tens of millions of cells and can be applied to many tasks by fine-tuning on small datasets, similar to how current large language models are operated [45].

Analysis of developmental single-cell transcriptomics data

Applying single-cell transcriptomics analysis to developmental biology has led to marked advances in our understanding of development across different human tissues, including, for example, the brain [46–48], the pancreas [49, 50], the lung [51–53], blood [54, 55], and across the whole fetus [9]. Analysing developmental scRNA-seq data can be challenging and requires tailored approaches. A primary challenge is the dynamic nature of developing tissues, characterised by ongoing continuous changes in gene expression patterns. In such settings, standard clustering methods may not adequately capture transient states or the progression of cells along developmental trajectories, complicating the assignment of cells to discrete types.

Trajectory inference methods have been developed to capture differentiation processes and their underlying transcriptional driving forces [56]. These techniques enable mapping continuous developmental pathways by ordering cells along a pseudo-temporal axis through, for example, computing diffusion processes on the cellular k-nearest neighbour (kNN) graph

[57, 58], comparing distances of mean transcriptional signatures of clusters [59], or leveraging additional data modalities such as the quantification of unspliced mRNA reads [60, 61]. A selection of trajectory inference approaches have been benchmarked for their accuracy, usability, and scalability [62]. Downstream trajectory inference tools (such as CellRank [63], Palantir [58], or FateID [64]) enable in-depth analyses of developmental processes, including the inference of gene dynamics over pseudo time and the detection of genes underlying developmental processes, so-called driver genes. CellRank, for example, computes a transition matrix from the output of the techniques mentioned above to automatically detect the initial and terminal states of a differentiation trajectory and uses this information to compute driver genes and gene expression dynamics.

1.1.3. Large-scale data curation in single-cell transcriptomics

With an ever-increasing number of scRNA-seq datasets publically available, new opportunities for machine-learning models trained on millions of cells to provide new biological insight emerge. Recently, so-called foundation models gained attention in single-cell genomics following their breakthroughs in language modelling. These models are pre-trained on data from tens of millions of cells to learn the fundamental relationships of gene expression. They can consequently be efficiently fine-tuned for specific tasks with only minimal extra data and have the potential to drive a paradigm shift in single-cell genomics data analysis [45]. Similarly, in single-cell genomics query-to-reference mapping, leveraging multiple existing datasets as a reference to augment or contextualise new experimental data can accelerate scientific discovery [42]. Overall, the reusability of data is becoming an increasingly important factor in enabling progress in the field.

The Findable, Accessible, Interoperable, and Reusable (FAIR) Data Principles [65] offer a structured framework to improve the reusability of scientific data. They ensure that data is not only stored in an accessible manner but is also comprehensible and easily reusable by other researchers. The data for most published scRNA-seq studies are *Findable* via weblinks or accessions in the Data Availability section of the associated manuscript. They are also usually *Accessible* through public repositories such as Zenodo or the Gene Expression Omnibus. While the first two FAIR principles (*Findable* and *Accessible*) are generally fulfilled in the field, this is much less the case for the *Interoperable* and *Reusable* principles. The main interoperability issue of published scRNA-seq datasets is the multitude of file types used for depositing the data. Some filetypes, such as loom (`https://loompy.org/`) or hdf5 (`https://www.hdfgroup.org/solutions/hdf5/`), require the installation of specific software

packages, while others, such as the R-specific RDS format, require the use of particular programming languages. Some of the universal file formats, including compressed text files (for example, using the .TSV.GZ extension), can also pose challenges in the case of large datasets: while filesize on disk can be relatively small through compression, these data formats store a dense count matrix, which can easily exceed the available system memory once loaded, even on powerful server instances. Formats storing sparse matrices on disk elegantly avoid memory limitations but can be slow to process. While these challenges in interoperability rarely entirely prevent data reuse, they make the process unnecessarily time-consuming for researchers. Free-text metadata, used in most published studies, further complicates data reuse. When several such studies are combined into a single reference dataset, two identical cell types might have slightly differently spelt labels and, without time-consuming manual intervention, will be treated as separate entities in downstream processing. This applies to cell type labels and almost all metadata, including tissue labels, experimental protocol, cell line of origin, and others, preventing efficient data reuse.

Streamlining available datasets through manual data curation is required to alleviate the abovementioned challenges, enable efficient data reuse, and unlock the full potential of the vast number of published scRNA-seq datasets. Data curation should include converting free-text metadata to controlled vocabularies as offered by the respective ontologies [66–68], available for almost all relevant metadata annotations. The generation of new datasets and the consequent insights into, for example, cell type heterogeneity should, in turn, lead to an updating of existing ontologies to maintain their utility [69]. Recent emerging streamlined data repositories, such as the *CZ CELLxGENE Discover Census* [70], the *Human Cell Atlas (HCA) Data Portal* (`https://data.humancellatlas.org`), or the EBI *Single cell expression atlas* [71], present a leap towards reusable data in single-cell genomics. Such repositories foster data reuse by providing streamlined data and metadata and offering multiple convenient data access options. Beyond these immediate advantages, their presence also encourages proper data curation by the authors at publication time, paving the way to a sustained improvement in data reusability in the field.

1.1.4. Single-cell transcriptomic atlases

Single-cell transcriptomic atlases are curated collections of datasets from the same biological system or tissue, aiming to provide a comprehensive reference map of the constituent cells. While the term *atlas* has been used to describe individual large-scale single-cell transcriptomic studies [72–75], it generally describes an integrated dataset comprised of several individual

studies by multiple research groups. Launched in 2016, the HCA initiative represents an ongoing effort to coordinate and harmonise the systematic generation of cell atlases from human tissues, aiming to characterise all human cell types based on their distinct transcriptomic signatures [76]. This endeavour has been enabled by the advent of single-cell transcriptomics, providing unprecedented resolution in studying cellular heterogeneity and revealing insights into cells' unique identities and states. Several atlases have thus far emerged from the constituent HCA Biological Networks. For instance, the Human Brain Cell Atlas [10] is a comprehensive neural atlas generated by an individual research group. It provides a detailed map of neuronal and glial cell types across brain regions, enhancing our understanding of the brain's complex architecture. In contrast, the Human Lung Cell Atlas [77] is an integrated and harmonised collection of 49 datasets that charts the cell landscape involved in respiratory function and pathology.

Besides data curation, described in the previous chapter of this thesis, data integration is a central element in constructing these atlases. It involves computationally blending data from multiple studies to create a cohesive and comprehensive integrated dataset. This is achieved by removing contributions of technical confounders such as sequencing technology or experiment site from the data. Many different data integration tools have been suggested in the literature and comprehensively benchmarked for their performance [78]. Some of the best-performing algorithms for scRNA-seq data integration included unsupervised [40, 79] and semi-supervised approaches [80, 81]. The latter can ingest prior information, such as cell type labels, to guide the integration process. Going one step further, a recently developed semi-supervised method has been designed to consider multiple levels of hierarchical cell type labels, potentially aiding the integration of datasets from particularly heterogeneous and complex tissues [82].

The opportunities presented by integrated cell atlases are vast. They can, for example, establish a community consensus on cell type labels for a given tissue [77], facilitating a common language for researchers to communicate their findings. Additionally, these atlases can enable the detection of novel, rare cell types that might be overlooked in individual studies [69] or inform about intra-cell type heterogeneity across different subsections of a tissue [10]. Perhaps most significantly, they provide a ground truth for the rapid annotation, contextualisation, and analysis of novel datasets [42].

1.2. Human neural organoids

The brain is the centre of human cognition and behaviour. It is critical to our overall health and well-being. Despite many advances in neuroscience, key aspects of human neurodevelopment still need to be fully understood. Gaining a better understanding of these processes is a prerequisite for comprehending the aetiology of neurological and psychiatric diseases and developing suitable treatments. These conditions represent a significant global health burden, affecting millions worldwide and posing substantial challenges in diagnosis, treatment, and long-term management. Understanding the complexity of brain development and the pathogenesis of neurological and psychiatric disorders is still challenging due to ethical, technical, and practical difficulties in studying human brain tissue directly. Besides the profound complexity of the brain, the scarcity of fetal post-mortem tissue and the limitations of post-mortem samples compared to *in vivo* tissue samples are the key limiting factors in modern neuroscience and psychiatry. Rodent models are valuable and accessible proxies to study human neural development. However, these models lack an abundance of outer radial glia (RG) characteristic of humans [83, 84] and differ in structural [85] and transcriptomic [86] aspects.

The emergence of neural organoids offers great potential in research areas where human-specific features are crucial, such as psychiatry [87]. Organoids are three-dimensional *in vitro* tissue models that allow for exploring previously inaccessible human-specific aspects of neurodevelopment and disease and open new avenues for developing targeted therapies to alleviate the global burden of brain disorders. They have, for example, already provided insights into the mechanisms underlying gyrification during neurodevelopment [88, 89], genetic diseases like microcephaly [90], and environmentally-conferred conditions such as Fetal Alcohol Spectrum Disorders [91].

1.2.1. Human neurodevelopment

The development of the human brain is a complex and tightly controlled process. Neural tube formation in the human embryo starts around postconceptional day 23 and leads to a tube-like structure comprising neuroepithelial cells. The neural tube is initially structured into three main segments (vesicles): the forebrain (prosencephalon), midbrain (metencephalon) and hindbrain (rhombencephalon) (Fig. 1.3a). These vesicles later subdivide to give rise to the different regions of the developed brain (Fig. 1.3b).

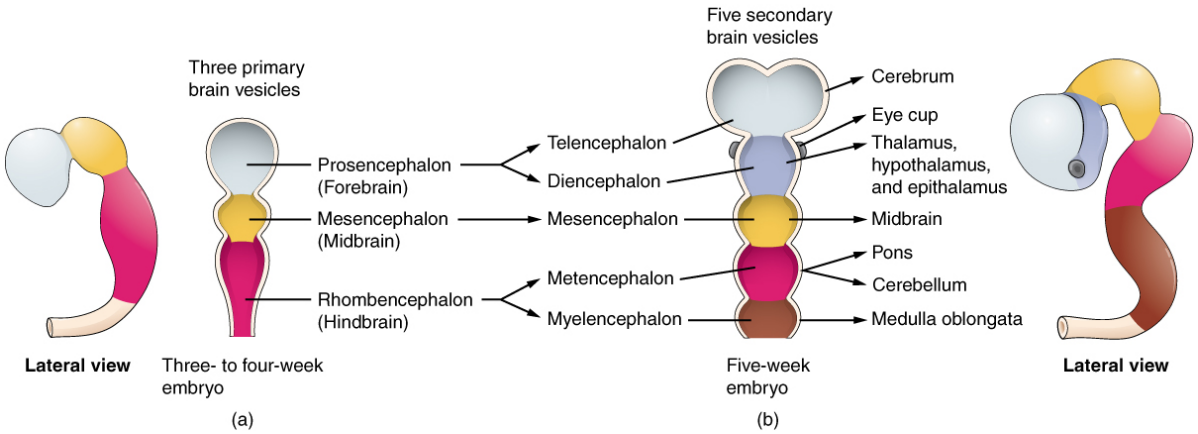

Figure 1.3.: **Schematic of the neural tube. a.** The three primary brain vesicles of the three- to four-week-old embryo. **b.** The five secondary brain vesicles of the five-week-old embryo. Reproduced from [92] Chapter 13.1 (Licensed under the Creative Commons Attribution 4.0 International license: `https://creativecommons.org/licenses/by/4.0/`).

In the further development of the human brain, neuroepithelial cells divide to create neural progenitor cells (NPCs) (such as apical RG found in the ventricular zone of the developing telencephalon), which further develop into neurons. This generation of neurons in the developing brain is called neurogenesis. In the developing telencephalon, approximately between postconceptional days 48 and 191, neurons are generated from apical RG by migration to the subventricular zone (then referred to as basal or outer RG) and consequent differentiation via the intermediate progenitor (IP) cell type. Following neurogenesis, newborn neurons, also known as neuroblasts, throughout the developing brain, migrate to their final locations in a process called neuronal migration. In the forebrain, newly born cortical projection neurons migrate radially to the cortical plate, while inhibitory neurons generated in the ganglionic eminences migrate laterally to the same location. Gliogenesis, the maturation of RG into macroglial cells (oligodendrocytes and astrocytes), occurs after neurogenesis and peaks around birth. Subsequent neurodevelopmental processes like synaptogenesis, myelination, and synaptic pruning also continue well into or occur exclusively in the postnatal period. Please refer to [93] for an in-depth introduction to neurodevelopment.

1.2.2. Neural organoids as models of human neurodevelopment

Derived from induced pluripotent stem cells (iPSCs) or embryonic stem cells, human neural organoids are three-dimensional self-organising systems that model the intricate process of

human brain development *in vitro*. Predominantly covering development during the first and second trimesters, they capture the critical neurodevelopmental process of neurogenesis. When cultured for 250 to 300 days, neural organoids have been shown to also resemble more advanced developmental (including postnatal) stages, with processes such as synapse formation occurring [94].

Structurally, neural organoids display several features reminiscent of the developing human brain. This includes the organisation of cells in layers around a central lumen, with RG lining the apical side (ventricular zone), IPs in the subventricular zone, and neurons positioned closer to what resembles a cortical plate [90, 95]. Notably, after about 150 days in culture, organoids can begin to display a form of cortical layering, distinguishing between upper and deeper layer neurons [96]. Furthermore, organoids can produce NPCs and neurons from various brain regions, even within a single organoid, highlighting their versatility as a model system [97]. They can contain a diverse array of neuronal cell types, partially resembling the cellular composition of the fetal brain while showing high transcriptional similarity in the cell types present [98, 99]. In older organoids, macroglial cells like oligodendrocyte precursor cells (OPCs) and astrocytes can also emerge [100–102].

However, neural organoids as a model system are not without limitations. The composition of cell types can significantly vary, not only across different organoids but also from primary tissue, with, for example, outer RG being underrepresented in organoids [103]. Structurally, the compact layering of neurons in organoids lacks the spatial fidelity observed in the primary brain and surrounding tissues, such as the meninges, are absent. Moreover, transcriptional analysis using scRNA-seq revealed a unique metabolic stress signature in organoids distinct from the cellular states encountered in primary fetal brain [103, 104]. Additionally, the absence of critical non-ectodermal-derived components like microglia [105, 106] and vasculature [107] highlights the limitations in fully recapitulating brain complexity, although recent attempts have been made to integrate these elements too [108–111]. Combining these ongoing endeavours with efforts to reduce cell stress and increase reproducibility in organoid differentiation protocols would lead the way to more accurately mimic the complexities and functions of the human brain, thereby enhancing the potential of neural organoids in basic research and clinical applications.

1.2.3. Neural organoid technologies

Generating neural organoids involves differentiating pluripotent stem cells into neural cell types within a three-dimensional culture system. Embryonic stem cells or iPSCs are initially aggregated to form embryoid bodies, facilitating cell-cell interactions essential for creating organoid structures. As organoids have presented profound self-patterning characteristics during their development [112], there are currently two main approaches for organoid differentiation from embryoid bodies, producing either unguided or regionalised organoids.

Unguided organoid differentiation protocols rely entirely on the system's inherent self-patterning abilities and do not add additional guiding molecules (morphogens) [113, 114]. This unbiased approach can generate organoids containing neural cells from a variety of brain regions, partially mimicking the complexity of the developing human brain (Fig. 1.4). It thereby enables studying the intricate self-organisational mechanisms underlying human brain development. At the same time, the lack of external guidance also leads to significant variability in regional composition across organoids from the same differentiation experiment [97], limiting the reproducibility of the approach.

The generation of regionalised organoids, in contrast, relies on adding morphogens to the organoid media to induce the development of specified brain regions. While the selection of morphogens is closely inspired by known signalling factors in early human neurodevelopment, such protocols provide an arguably less physiological environment for cellular differentiation and maturation. Besides the added control in organoid generation, such approaches offer a high degree of reproducibility and can also increase the maturation rate of the organoids. Morphogens have been used to generate organoid models of various brain regions, most prominently the dorsal telencephalon [99, 115–117], but also the ventral telencephalon [118, 119], the thalamus [120] and hypothalamus [121], the midbrain [122], the hindbrain [123], and the ChP [124]. Besides specific brain regions, neural organoids have been guided to enrich particular cell types such as oligodendrocytes [125].

Over the past years, significant efforts have been made to develop protocols that produce neural organoids that are more similar to their primary counterparts. Improvements have been suggested at different points of the protocol. Examples include the embedding in matrigel towards the beginning of the protocol to enable more complex tissue growth [90], adding specific maturation media to support development [126], or slicing of the organoids at later protocol stages to improve nutrient supply and reduce cell stress [96]. More involved approaches include the construction of assembloids [127], where organoids of different brain

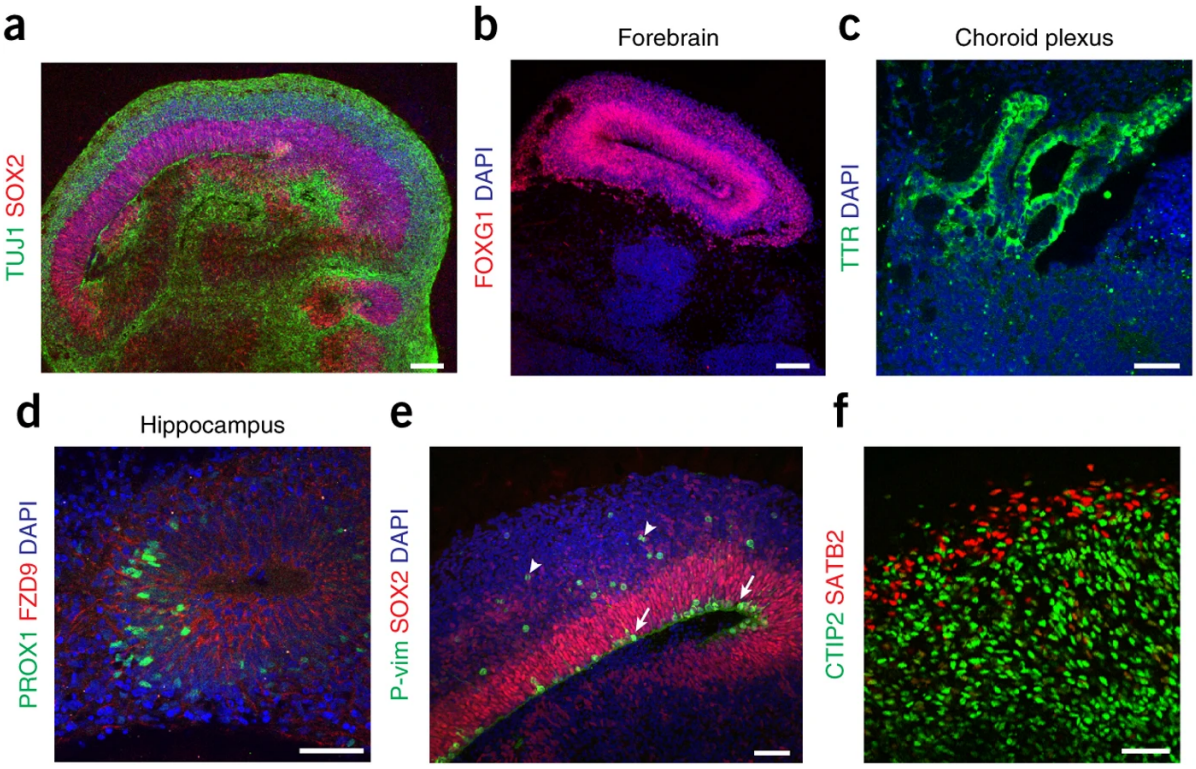

Figure 1.4.: **Immunohistochemistry stains of unguided neural organoids resembling different brain regions. a.** Structural organisation of apical NPCs (SOX2, red) and basally located neurons (TUJ1, green). **b.** Forebrain neurons (FOXG1, red). **c.** Choroid plexus (ChP) tissue (TTR, green). **d.** PROX1 (green) and FZD (red) mark hippocampal regions. **e.** SOX2 (red) marks RG. Mitotic RG (P-vimentin (P-vim), green) are present at the apical membrane (arrows) and outside the ventricular zone (arrowheads). **f.** Layering of more mature upper-layer neurons (SATB2, red) and younger deep-layer neurons (CTIP2, green). DAPI (blue) marks nuclei of cells. Reproduced from [113] with permission from Springer Nature.

regions are cultured in close proximity and functionally integrate to form basic neural circuits [118, 128]. More recently, the transplantation of neural organoids into living animals has been shown to boost organoid maturation while displaying an integration of neural circuits [129].

1.2.4. Neural organoids as disease models

Besides being powerful models of human-specific neurodevelopment [97], neural organoids offer unique disease-modelling capabilities due to their accessibility and flexibility. Either

by derivation from patient-specific cell lines or by exposing organoids to external stimuli and stressors, they can be used to model the contribution of genetic and environmental risk factors to the onset and development of disease. In the context of neurodevelopmental disorders (NDDs), organoids have helped study genetic diseases, for example, shedding light on altered neuronal migration [130] and the role of apical RG [131] in neuronal heterotopia and suggesting possible treatment options [132]. Further genetic diseases studied in organoids include Timothy syndrome [118], Rett syndrome [133, 134], and Miller-Dieker syndrome [135]. Organoids have also served as valuable models of neurodegenerative diseases, including Alzheimer's disease [136] and Parkinson's disease [137]. In the context of viral infections, organoids have helped to understand the effects of the Zika virus [138, 139] and human cytomegalovirus [140] on brain development.

In the context of psychiatry, organoids derived from schizophrenia patients have shown altered progenitor survival and impaired neurogenesis [141]. In ASD, perturbing three risk genes resulted in asynchronicity in the development of excitatory and inhibitory neuron classes [142], substantiated by a large-scale perturbation screen of ASD risk genes indicating priming of the ventral telencephalic inhibitory neuron lineage in ASD [143]. Concerning external toxin exposure modelling, neural organoids have provided insight into, for example, the neurotoxic effects of alcohol [91, 144] and methamphetamine [145], leading to apoptosis, impaired development, or neuroinflammation.

1.3. Neurodevelopment and psychiatric disease

Neurodevelopment is a highly intricate and tightly regulated process that lays the foundational framework for brain architecture, affecting an individual's susceptibility to disease. Deviations in this process, whether due to genetic, epigenetic, or environmental factors, can predispose individuals to psychiatric conditions later in life [146]. For instance, schizophrenia has been associated with disruptions in synaptic pruning, a critical process of neurodevelopment [147, 148]. Similarly, ASD exemplifies the profound impact of early neurodevelopmental anomalies during neurogenesis on cognitive and behavioural outcomes, as noted in Section 1.2.4.

Studying the genetic underpinnings of psychiatric conditions provides further evidence for the tight link between neurodevelopmental processes and the aetiology of psychiatric disease. In a recent study, several risk genes for ASD and other NDDs have been shown to affect interneuron generation and migration, thereby directly interfering with critical

neurodevelopmental processes [149]. Conversely, many variants linked to genes with known essential roles in neurodevelopmental pathways confer risk for ASD, schizophrenia or other psychiatric diseases [150–152]. Overall, these disorders often arise from a complex interplay of genetic vulnerabilities and early life experiences, emphasising the importance of the neurodevelopmental period in psychiatric disease manifestation [153].

1.3.1. Environmental impact on brain development and disease

The heritability of psychiatric conditions varies profoundly by disorder. For example, ASD is considered a largely genetic disease with heritability estimates around 75 % [154]. In contrast, for major depressive disorder, heritability is estimated to be below 50 % in the general population [155, 156]. Environmental factors, therefore, play an essential role in the onset and development of psychiatric disease. Early human brain development during pregnancy is marked by extensive tissue transformation and rapid brain maturation. Hence, it is a highly susceptible period for environmental perturbations interfering with the developmental process.

Birth weight, a coarse proxy for the environmental conditions experienced by the fetus during pregnancy, has been correlated with multiple neurobehavioural outcomes, including educational attainment and executive function, as well as disease-related outcomes, including attention deficit hyperactivity disorder and depressive symptoms [157]. Many forms of environmental interference with development have been associated with increased risk for psychiatric diseases, ranging from immune-induced perturbations caused by viral infections [158] to widespread chemicals with potentially neurotoxic properties [159]. Beyond exposure to pathogens or harmful substances, any difficulties arising during pregnancy and birth [160, 161] as well as impaired maternal health and well-being, also contribute to disease risk [162]. These factors can lead to measurable structural aberrations in the newborn child's brain [163, 164].

While the exact mechanisms underlying environmental interference with development are diverse and not fully understood, GCs have been suggested as one potential mediator of prenatal adversity in the unborn child [165].

1.3.2. Glucocorticoids in brain development

GCs are vital steroid hormones primarily produced by the adrenal cortex that play several important roles in the human body. Firstly, they are essential mediators of the human stress response as a component of the hypothalamic-pituitary-adrenal (HPA) axis. In case of a stress-inducing stimulus, the hypothalamus releases corticotropin-releasing hormone, which prompts the pituitary gland to secrete adrenocorticotropic hormone, which in turn stimulates the adrenal cortex to produce GCs, such as cortisol. GCs exert their function by binding to glucocorticoid receptors (GRs) and mineralocorticoid receptors present in the cytosol of most cells throughout the body. Upon binding GCs, the GR translocates to the nucleus, regulating the transcription of various genes as a transcription factor (TF) by binding to GC-response elements [166] and interacting with other TFs [167]. Besides their role in the human stress response, GCs play a crucial role in development.

GCs are critical for the development of several tissues, such as the lung [168] and the cardiovascular system [169]. Additionally, they regulate essential processes like synapse formation and elimination, neural migration, and neurogenesis in the developing brain [165]. GC levels are tightly regulated during development across many mammals, and in the unborn human child, GC levels rise sharply during the third trimester [170]. These hormonal signals are crucial for the timely execution of developmental milestones and for preparing the fetus for birth [171]. While the exact underlying molecular mechanisms are not fully understood, the role of GCs in neurodevelopment is a delicate balance, with aberrant levels having the potential to alter brain structure and function (Fig. 1.5a) and the fetal HPA-axis (Fig. 1.5b), affecting postnatal behaviour and increasing the risk of neuropsychiatric disorders [165, 172].

Maternal stress or diseases can lead to elevated maternal GC levels during pregnancy, which can profoundly affect fetal brain development [173]. While the placenta acts as a barrier, to a large extent protecting the fetus from maternal GC levels, its protective function can be reduced through stress and increased GC levels in the mother [174] (Fig. 1.5c), leading to altered GC levels in the fetus. Excessive GC exposure in the unborn child can disrupt neural development, leading to structural changes in the brain and altered postnatal behaviour [175]. Besides structural changes, epigenetic modifications, mainly on the DNA-methylation level, have been suggested to mediate the effects of excess GCs on behavioural outcomes [165, 176] (Fig. 1.5d), and to play a role in the intergenerational transmission of adverse experiences [173, 177].

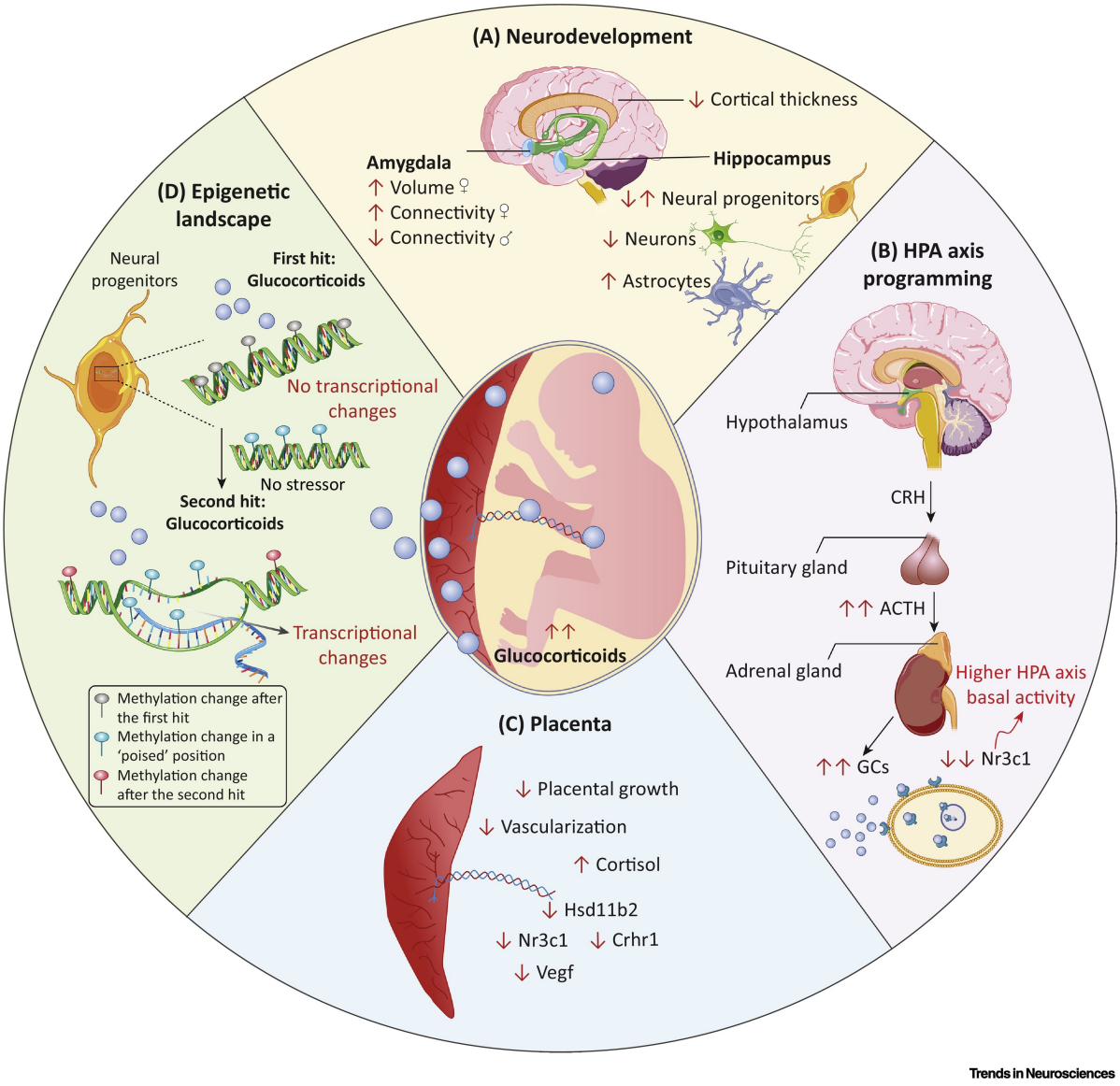

Figure 1.5.: **Effects of excess prenatal GC exposure. a.** GCs influence neurodevelopment by impairing neuronal differentiation and affecting key limbic system structures, including the amygdala and hippocampus. **b.** Excess GCs lead to increased fetal HPA axis basal activity. **c.** Elevated GC levels impact the placenta by hindering its vascular development and growth and suppressing the expression of HSD11B2. This enzyme protects the fetus from maternal GCs. **d.** Increased GC levels modify the fetus's epigenetic framework, boosting transcriptional reactions to stressors encountered later in life. ACTH, adrenocorticotropic hormone; CRH, corticotropin-releasing hormone; CRHR1, corticotropin-releasing hormone receptor 1; HSD11B2, hydroxysteroid 11-β dehydrogenase 2; NR3C1, nuclear receptor subfamily 3 group C member 1; VEGF, vascular endothelial growth factor. Reproduced from [165] with permission from Elsevier.

Beyond increased endogenous GC levels in the mother, the fetus can be exposed to non-physiological levels of GCs through the pharmacological use of sGCs, such as dexamethasone and betamethasone during pregnancy. SGCs are critical antenatal drugs, accelerating the maturation of the fetus, especially its lungs, in case of risk for preterm birth. Unlike their endogenous counterparts, sGCs readily cross the placenta and exert their pharmacological effects directly on the fetus [178]. Their use significantly reduces fetal morbidity and mortality [179] but also affects fetal neurodevelopment [165, 171]. Given that over ten per cent of babies (over 13 million) were born prematurely in 2020 [180], antenatal GC treatment is a sizable environmental factor for fetal brain development. Studies suggest that the long-term outcomes of antenatal GC treatment on the brain depend on the timing of administration. For extremely preterm births (before 28 gestational weeks), a beneficial effect of GC treatment is observed with a significantly reduced risk for impaired neurodevelopment [181, 182]. In contrast, administering GCs later in pregnancy was associated with adverse long-term outcomes, including an increased prevalence of mental and behavioural disorders and neurocognitive impairments [183–186].

1.3.3. Modelling glucocorticoid effects on brain development

Rodent models have provided valuable insights into GC effects on brain development, such as altered HPA-axis activity in rats [187] and altered hippocampal development in mice [188]. As discussed in Section 1.2, fundamental differences exist between the developing lissencephalic rodent brain and the gyrified human brain, limiting the capacity of rodents to model human GC-induced neurodevelopmental aberrations that could lead to disease. Larger mammals, such as sheep, have been used to study GC exposure in a gyrified brain [189, 190]; however, their handling as model animals is much more complex and expensive while they still not have the same underlying genetic information and regulation as humans. As discussed in Section 1.2.2, human neural organoids, while still far from perfect, are accessible three-dimensional models of human brain development and have already been used successfully to model the effect of psychiatric disease risk factors (Section 1.2.4). The remainder of this section focuses on neural organoid models. Please refer to [172] for an in-depth review of different *in vitro* modelling approaches of the neurobiological effects of GCs.

The mechanisms underlying the onset and development of psychiatric diseases are complex and likely involve multiple cell types. Modelling the effects of environmental risk factors for disease on development in neural organoids is, therefore, best done in models that accurately

recapitulate the cell type diversity and interactions occurring in the primary fetal brain. While some cell types, such as microglia, are inherently absent in current neural organoid models and are only added in specialised protocols [111], general protocol choice can affect the cell type diversity in the organoid. For example, organoids derived using strongly regionalised cortical protocols [96, 103] often lack a significant population of inhibitory neurons. Such a lack can limit the ability of the model to capture relevant developmental aberrations, such as an imbalance in excitatory and inhibitory neuronal lineage development, which is relevant for the aetiology of certain psychiatric diseases [191–194]. While having their own limitations (Section 1.2.3), unguided neural organoids generally produce a more comprehensive array of neural cell types and might be preferential models in that regard. Alternatively, the fusion of regionalised organoids of different brain regions into assembloids is a promising approach for modelling development in the context of psychiatric diseases with a specific focus on neural circuits [149].

Besides implicating multiple cell types, there is strong evidence for the involvement of multiple brain regions beyond the cortical regions in the onset and development of psychiatric diseases. To name some examples, dysregulation in the midbrain has been associated with schizophrenia and bipolar disorder [195, 196] and thalamic connectivity system deficits have been linked to various psychiatric conditions, including schizophrenia, bipolar disorder, major depressive disorder, and ASD [197, 198]. The hypothalamus, as part of the HPA-axis, has been associated with, for example, anxiety disorders [199, 200] and the hindbrain, specifically the cerebellum and pons, have been implicated in ASD [201–203]. Therefore, understanding the contribution of environmental risk factors, such as GCs, to the aetiology of disease requires the consideration of multiple brain regions. As in the previous section, unguided neural organoids arise as a particularly suitable model in this context, as they can generate neuronal cells from various brain regions and, therefore, can provide a broader view of GC exposure effects during brain development.

1.4. Research questions and contributions of this thesis

The main question I address in this thesis is how to improve our understanding of neural organoids as a model system of the developing brain to consequently use them to discover the mechanisms underlying the onset and development of psychiatric diseases. I structure this question into the following open challenges:

1. Single-cell genomics technologies have enabled many scientific breakthroughs over the past years, and the number of publicly available scRNA-seq datasets is growing rapidly. Public data can provide a rich resource to augment and contextualise new studies through integrated atlases. However, data curation and reuse have remained challenging and time-consuming due to a lack of metadata standardisation and consistent data formats in the field.

2. The curation of large-scale transcriptomic cell atlases can uncover the mechanisms underlying development and disease. The human brain is one of the most complex organs in the human body and is critical to our overall health and well-being. Neural organoids have emerged as powerful *in vitro* models of human neurodevelopment. Combining a large number of existing neural organoid scRNA-seq datasets into a comprehensive single-cell atlas would uncover underrepresented sections of the developing human brain in organoid models. It would further provide insight into systematic transcriptomic differences between organoids and primary fetal brain and pave the way towards more faithful organoid model systems. It would also enable the rapid and effective contextualisation of newly generated neural organoid datasets.

3. Improved *in vitro* models of human brain development in health and disease are the foundation for developing future treatment strategies. Novel model systems such as neural organoids have enabled promising advances in understanding the genetic contribution to the onset and development of psychiatric diseases in a human context. While it is widely accepted that environmental factors also strongly contribute to these pathological processes, the underlying mechanisms have been investigated to a significantly smaller extent.

In this thesis, I aim to facilitate the curation of large-scale transcriptomic cell atlases with a new software framework, uncover the current limitations of neural organoid models through an integrated neural organoid cell atlas and improve our understanding of the mechanisms underlying environmental risk for psychiatric diseases using neural organoids.

Specifically, I address the three abovementioned challenges with the following contributions:

1. In collaboration with David Fischer, I developed the SFAIRA package. SFAIRA is a data and model zoo for single-cell genomics that allows for efficient data curation and automated analysis. In particular, SFAIRA facilitates the curation of large-scale single-cell atlases using controlled metadata vocabularies and streamlining of feature spaces.

2. In collaboration with Zhisong He and Jonas S. Fleck, I constructed the HNOCA, a harmonised single-cell reference atlas encompassing a comprehensive list of public neural organoid scRNA-seq datasets. HNOCA revealed compositional and transcriptional differences between primary fetal brain and organoids and serves as a reference to contextualise novel scRNA-seq datasets.

3. I show that chronic exposure to GCs, an example of an environmental challenge during development, affects the expression of several key neurodevelopmental genes and leads to priming of the inhibitory neuron lineage in neural organoids.

These contributions are related to each other in that I used the SFAIRA framework to curate the datasets for the HNOCA. SFAIRA allowed for an efficient harmonisation of the comprising datasets' metadata, which was essential for several aspects of our analysis and visualisations presented in the HNOCA manuscript. In a further connection, I used the HNOCA as a reference atlas to assign cell type labels in our GC-treated neural organoid scRNA-seq dataset. The availability of the HNOCA as a reference atlas allowed me to identify the neuronal cells in the dataset as originating from non-telencephalic brain regions, a less common region of neuron origin in unguided neural organoids. Additionally, the HNOCA result highlighting cell stress as a universal property of neural organoid scRNA-seq data enabled me to improve the preprocessing of the newly generated organoid data. See Fig. 1.6 for a visual representation of these contributions and their connections.

Altogether, my contributions deepen our understanding of neurodevelopmental disorder development and pave the way towards improved disease modelling in organoids, which has the potential to drive novel therapeutic strategies in psychiatry and beyond.

The main findings of my work have already been published in a peer-reviewed journal or are under revision at the time of writing this thesis. Therefore, this thesis is based on and in parts identical to the following publications:

1. Fischer, D. S.*, **Dony, L.***, König, M., Moeed, A., Zappia, L., Heumos, L., Tritschler, S., Holmberg, O., Aliee, H. & Theis, F. J. Sfaira accelerates data and model reuse in single cell genomics. *Genome Biology* **22**, 248. doi:10.1186/s13059-021-02452-6 (2021)

2. He, Z.*, **Dony, L.***, Fleck, J. S.*, Szałata, A., Li, K. X., Sliškovic, I., Lin, H. -C., Santel, M., Atamian, A., Quadrato, G., Sun, J., Pasca, S. P., Camp, J. G., Theis, F. & Treutlein, B. An integrated transcriptomic cell atlas of human neural organoids. *In revision.* Preprint [204] doi: 10.1101/2023.10.05.561097 (2023)

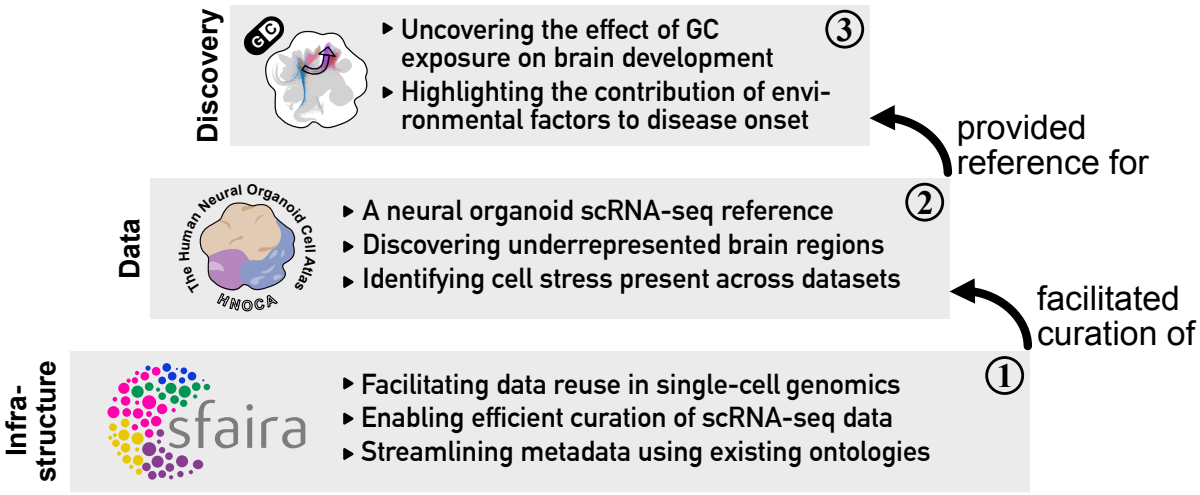

Figure 1.6.: **Thesis contributions overview.** A pyramidal schematic displaying the relationship between the contributions presented in this thesis. Robust and scalable infrastructure serves as a foundation for curating comprehensive datasets that provide a reference to contextualise novel datasets addressing specific biological questions.

3. **Dony, L.**, Krontira, A. C., Kaspar, L., Ahmad, R., Demirel, I. S., Grochowicz, M., Schäfer, T., Begum, F., Sportelli, V., Raimundo, C., Koedel, M., Labeur, M., Cappello, S., Theis, F. J., Cruceanu, C.*, Binder, E. B.* Chronic exposure to glucocorticoids amplifies inhibitory neuron cell fate during human neurodevelopment in organoids. *In review.* Preprint [205] doi: 10.1101/2024.01.21.576532 (2024)

Note that "*" denotes an equal contribution to the publication. Specifically, I have made the following contributions to these publications:

- **Publication 1:** Together with David Fischer, I wrote the package code and carried out the analysis for the publication. I lead the data curation and model training with support from David Fischer. I designed the organoid-specific metadata scheme with input from the other authors.

- **Publication 2:** Together with Zhisong He and Jonas S. Fleck, I collected and retrieved the scRNA-seq data included in the HNOCA with suggestions from Fabian Theis, Sergiu P. Pasca, J. Grayson Camp and Barbara Treutlein. Together with Zhisong He, I performed the DE and transcriptomic comparison analysis with support from Irena Sliškovic

and Katelyn X. Li. Together with Katelyn X. Li, Irena Sliškovic and Artur Szałata, I curated the HNOCA data and harmonised the associated metadata. I preprocessed and integrated the HNOCA datasets with support from Katelyn X. Li and Irena Sliškovic. Together with Zhisong He, Jonas S. Fleck and Artur Szałata, I developed the HNOCA data processing pipeline. Together with Katelyn X. Li, I performed the benchmark of integration methods. I provided input for the cell type hierarchy curation by Zhisong He and Jonas S. Fleck. I coordinated the work of Artur Szałata, Katelyn X. Li, and Irena Sliškovic by defining tasks and providing input where needed.

- **Publication 3:** Together with Elisabeth B. Binder, Cristiana Cruceanu, and Fabian Theis, I defined the main questions and analysis steps of this work. I analysed all scRNA-seq and single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq) data. I interpreted and visualised the results with input from my co-authors.

During my doctoral work, I contributed to the following additional peer-reviewed publications not included in this thesis (in chronological order):

1. Bastidas-Ponce, A., Tritschler, S., **Dony, L.**, Scheibner, K., Tarquis-Medina, M., Salinno, C., Schirge, S., Burtscher, I., Böttcher, A., Theis, F. J., Lickert, H. & Bakhti, M. Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development* **146**, dev173849. doi:10.1242/dev.173849 (2019)

2. **Dony, L.**, König, M., Fischer, D. S. & Theis, F. J. Variational autoencoders with flexible priors enable robust distribution learning on single-cell RNA sequencing data in *ICML 2020 Workshop on Computational Biology (WCB) Proceedings*, Paper 37. `https://icml-compbio.github.io/2020/papers/WCBICML2020_paper_37.pdf` (2020)

3. Lotfollahi, M.*, **Dony, L.***, Agarwala, H.* & Theis, F. J. Out-of-distribution prediction with disentangled representations for single-cell RNA sequencing data in *ICML 2020 Workshop on Computational Biology (WCB) Proceedings*, Paper 41. `https://icml-compbio.github.io/2020/papers/WCBICML2020_paper_41.pdf` (2020)

4. Cruceanu, C., **Dony, L.**, Krontira, A. C., Fischer, D. S., Roeh, S., Di Giaimo, R., Kyrousi, C., Kaspar, L., Arloth, J., Czamara, D., Gerstner, N., Martinelli, S., Wehner, S., Breen, M. S., Koedel, M., Sauer, S., Sportelli, V., Rex-Haffner, M., Cappello, S., Theis, F. J. & Binder, E. B. Cell-Type-Specific Impact of Glucocorticoid Receptor Activation on the Developing Brain: A Cerebral Organoid Study. *American Journal of Psychiatry* **179**, 375–387. doi:10.1176/appi.ajp.2021.21010095 (2022)

5. Heumos, L., Schaar, A. C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücken, M. D., Strobl, D. C., Henao, J., Curion, F., Single-cell Best Practices Consortium, Aliee, H., Ansari, M., Badia-i-Mompel, P., Büttner, M., Dann, E., Dimitrov, D., **Dony, L.**, Frishberg, A., He, D., Hediyeh-zadeh, S., Hetzel, L., Ibarra, I. L., Jones, M. G., Lotfollahi, M., Martens, L. D., Müller, C. L., Nitzan, M., Ostner, J., Palla, G., Patro, R., Piran, Z., Ramírez-Suástegui, C., Saez-Rodriguez, J., Sarkar, H., Schubert, B., Sikkema, L., Srivastava, A., Tanevski, J., Virshup, I., Weiler, P., Schiller, H. B. & Theis, F. J. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics* **24**, 550–572. doi:10.1038/s41576-023-00586-w (2023)

6. Sikkema, L., Ramírez-Suástegui, C., Strobl, D. C., Gillett, T. E., Zappia, L., Madissoon, E., Markov, N. S., Zaragosi, L. - E., Ji, Y., Ansari, M., Arguel, M. - J., Apperloo, L., Banchero, M., Bécavin, C., Berg, M., Chichelnitskiy, E., Chung, M.- i., Collin, A., Gay, A. C. A., Gote- Schniering, J., Hooshiar Kashani, B., Inecik, K., Jain, M., Kapellos, T. S., Kole, T. M., Leroy, S., Mayr, C. H., Oliver, A. J., Von Papen, M., Peter, L., Taylor, C. J., Walzthoeni, T., Xu, C., Bui, L. T., De Donno, C., **Dony, L.**, Faiz, A., Guo, M., Gutierrez, A. J., Heumos, L., Huang, N., Ibarra, I. L., Jackson, N. D., Kadur Lakshminarasimha Murthy, P., Lotfollahi, M., Tabib, T., Talavera-López, C., Travaglini, K. J., Wilbrey-Clark, A., Worlock, K. B., Yoshida, M., Lung Biological Network Consortium, Chen, Y., Hagood, J. S., Agami, A., Horvath, P.,Lundeberg, J., Marquette, C.- H., Pryhuber, G., Samakovlis, C., Sun, X., Ware, L. B., Zhang, K., Van Den Berge, M., Bossé, Y., Desai, T. J., Eickelberg, O., Kaminski, N., Krasnow, M. A., Lafyatis, R., Nikolic, M. Z., Powell, J. E., Rajagopal, J., Rojas, M., Rozenblatt-Rosen, O., Seibold, M. A., Sheppard, D., Shepherd, D. P., Sin, D. D., Timens, W., Tsankov, A. M., Whitsett, J., Xu, Y., Banovich, N. E., Barbry, P., Duong, T. E., Falk, C. S., Meyer, K. B., Kropski, J. A., Pe'er, D., Schiller, H. B., Tata, P. R., Schultze, J. L., Teichmann, S. A., Misharin, A. V., Nawijn, M. C., Luecken, M. D. & Theis, F. J. An integrated cell atlas of the lung in health and disease. *Nature Medicine* **29**, 1563–1577. doi:10.1038/s41591-023-02327-2 (2023)

7. Krontira, A. C., Cruceanu, C., **Dony, L.**, Kyrousi, C., Link, M. - H., Kappelmann, N., Pöhlchen, D., Raimundo, C., Penner-Goeke, S., Schowe, A., Czamara, D., Lahti-Pulkkinen, M., Sammallahti, S., Wolford, E., Heinonen, K., Roeh, S., Sportelli, V., Wölfel, B., Ködel, M., Sauer, S., Rex-Haffner, M., Räikkönen, K., Labeur, M., Cappello, S. & Binder, E. B. Human neurogenesis is altered via glucocorticoid-mediated regulation of ZBTB16 expression. *Neuron* **In press**. doi:10.1016/j.neuron.2024.02.005 (2024)

While these publications are not discussed further in this thesis, the following connections exist:

- **Additional Publication 1** describes a scRNA-seq analysis of Neurogenin3-positive cells in the developing mouse pancreas. Under the guidance of Sophie Tritschler, I analysed the scRNA-seq data and generated parts of the visualisations presented in the manuscript. These contributions equipped me with a good understanding of scRNA-seq analysis in a developing system focusing on lineage analysis, which helped me analyse the neural organoid scRNA-seq datasets presented in this thesis.

- **Additional Publication 2** suggests a more expressive prior for variational autoencoder (VAE)-based representation learning for scRNA-seq data. I conceived the project with input from David Fischer, defined the datasets and metrics for the manuscript and, together with Martin König, developed the method. Karin Hrovatin has since further explored this approach for the large-scale integration of scRNA-seq data in challenging scenarios similar to the HNOCA presented in this thesis.

- **Additional Publication 3** explores different approaches to obtaining disentangled latent spaces in scRNA-seq representation learning. I provided input on biologically relevant features for disentanglement and, together with Mohammad Lotfollahi, coordinated the project. We obtained promising results for disentangling perturbation effects from other biological variation, which could aid the analysis of future perturbation studies similar to the one presented in the context of GCs in this thesis.

- **Additional Publication 4** investigates the effects of acute GC exposure on gene expression during brain development in neural organoids. Together with Cristiana Cruceanu, I devised the scRNA-seq analysis steps. I analysed all scRNA-seq data and produced parts of the visualisations presented in the manuscript. This work laid the foundation for the chronic GC exposure studied in this thesis. It showed that neural organoids are responsive to GC treatment and a suitable model system for this application. Furthermore, it highlighted that the transcriptional changes induced by GC exposure are significantly enriched for genes associated with ASD.

- **Additional Publication 5** devises best practice recommendations for analysing single-cell data across different modalities. Together with Lukas Heumos, I contributed to ensuring reproducibility and accessibility to the materials by developing a containerised version of each chapter. The recommendations formulated in this work informed many of the analysis choices taken in the work presented in this thesis.

- **Additional Publication 6** constructs an integrated human lung cell atlas. Together with other authors, I contributed to the data curation by collecting datasets and associated metadata for the atlas. This contribution provided me with valuable learnings for later curating datasets for the HNOCA.

- **Additional Publication 7** identified ZBTB16 as a crucial mediating factor of the GC response in the developing brain. I analysed the scRNA-seq data presented in this publication. This work is closely connected to the work on chronic GC exposure in neural organoids, presented here. Specifically, it provides mechanistic insight into the increased neuronal differentiation drive reported in this thesis.

2. Methods

This chapter describes the materials and methods my collaborators and I have used to generate the results presented in Chapter 3. I have detailed my contributions to each of the following sections in Section 1.4 of this thesis.

2.1. Sfaira accelerates data and model reuse in single-cell genomics

This method section corresponds to, and is in part identical to, the one presented in the following publication:

Fischer, D. S.*, **Dony, L.***, König, M., Moeed, A., Zappia, L., Heumos, L., Tritschler, S., Holmberg, O., Aliee, H. & Theis, F. J. Sfaira accelerates data and model reuse in single cell genomics. *Genome Biology* **22**, 248. doi:10.1186/s13059-021-02452-6 (2021)

"*" denotes an equal contribution.

2.1.1. Data and model zoo implementation

In SFAIRA, we represent datasets using data loader classes, which inherit generic data loading capabilities from a parent class. These classes are a structured version of data-loading scripts, which are commonly used in data analysis workflows. The classes allow lazy metadata queries through automatic metadata storage, meaning count matrices can be loaded into memory on demand. This feature allows users to interactively subset large instance lists of these classes. In cases where data repositories directly provide streamlined processed datasets, individual loading scripts are not required. Instead, we interface these existing data zoos with a single class that can be instantiated for all datasets in the zoo. SFAIRA maintains the universe of all contributed data loader classes, and users can dynamically build libraries

of a subset of these datasets through the sfaira data API. We provide all the code for the sfaira data and model zoo in the sfaira package. Each model in the zoo is built around its individual model class, which can be accessed through a streamlined interface, similar to the established genomics model zoo kipoi [206].

2.1.2. Serving pre-trained model weights and topologies

The sfaira model zoo provides access to model weights stored in public versioned repositories like Zenodo [207] or locally for private models. These parameter files are versioned, which means they can be accessed in a reproducible manner. The sfaira model zoo is designed to host various model architectures. Here, we only discuss the models underlying the presented analysis results. It is important to note that we expect the model population in sfaira to expand beyond this initial set of models in the future.

2.1.3. Intrinsic preprocessing of count data

We have added a common input data transform to all embedding and cell type prediction models. This transform helps to reduce variability in the data, requiring less model complexity and fewer training steps to adjust the models' internal normalisation of the data. We chose a transform that can be evaluated irrespective of the batch size, depending only on an individual observation (cell). We scaled the data points x per cell i and gene j to $10,000$ and log-transformed this scaled vector to reduce variation caused by the number of UMIs observed per cell. This variation depends on technical factors such as library depth and stochasticity in mRNA capture during the sequencing experiment.

$$x_{ij} \;=\; \log\left(\frac{x_{ij}}{\sum_{n=0}^{N} x_{in}} \times 10^4\right) \tag{2.1}$$

It is important to note that this processing does not necessarily need to be carefully benchmarked, unlike for principle component analysis (PCA) and downstream Uniform Manifold Approximation and Projection (UMAP) [208] computation in standard scRNA-seq analysis workflows, due to the innate ability of the first layers of the neural network to adjust to unwanted sources of variation. We chose to use a basic transform to improve training performance. With many datasets and sufficient training time, it is possible to imagine entirely

removing preprocessing from these networks in the long run.

2.1.4. Output and loss function of embedding models

In SFAIRA, we use AE-based embedding models to learn a low-dimensional representation of the data. Please refer to [209] for an in-depth review of deep-learning models in single-cell genomics. Our tool supports various model outputs and loss functions. The exact configurations are represented in the topology identifier. Studies have shown that AEs can learn embeddings of scRNA-seq data using a negative binomial reconstruction loss. This type of loss requires the estimation of a mean (μ) and a dispersion parameter (ϕ). In the first version of our tool, SFAIRA, we have included output states specifically geared towards the negative binomial distribution. We use an exponential inverse-linker function in the final layer to achieve this. Two types of output are available: one that estimates a fixed dispersion per gene and another that estimates one dispersion parameter per gene and cell. The negative log-likelihood over N samples and J genes is computed as follows:

$$\begin{aligned} ll_{\mathrm{NB}}(\mu, \phi; x) \;=\; & -\sum_{n=0}^{N}\sum_{j=0}^{J} \log\Gamma(\phi_j + x_{nj}) + \log\Gamma(x_{nj}+1) + \log\Gamma(\phi_j) \\ & - x_{nj}(\log(\mu_{nj}) + \log(\mu_{nj}+\phi_j)) - \phi_j(\log(\phi_j) + \log(\mu_{nj}+\phi_j)) \end{aligned} \tag{2.2}$$

2.1.5. Output and loss function of cell type prediction models

The standard SFAIRA cell type prediction model assumes that the output should be a probability distribution across cell types that are already known. The cross-entropy loss is typically used to evaluate the fit of such probability mass distributions. As an extension, we allow multiple output categories to be assigned to a single true set of labels, which we call aggregated cross-entropy loss. This is used when datasets differ in their cell type label granularity. For instance, one dataset may only annotate a few tissue-specific cell types and *lymphocytes*. In contrast, another dataset may contain the same four tissue-specific types and differentiate lymphocytes further into *T cells* and *B cells*. As the cell ontology [66] usually provides a mapping between different annotation levels, we propose aggregating the predicted class probabilities across all labels categorised under *lymphocytes* for the first dataset. This allows any probability mass distribution for a lymphocyte observation in the first dataset across

T and *B cells*, enabling the classifier to learn the differences between T and B cells from the second dataset. At the same time, the classifier can use the first dataset to improve its model of the difference between lymphocytes and the remaining cell types. We compare the resulting aggregated cross-entropy loss (cce_{agg}) to cross-entropy for a binary (cce_{binary}) and a multi-class ($cce_{multi-class}$) prediction problem. The transformations labelled with (*) hold if $y \subset \{0, 1\}$, i.e., if the labels lie on binary support.

$$cce_{binary} = -\sum_{n=0}^{N} y_n \log(p_n) + (1 - y_n) \log(1 - p_n) \overset{(*)}{=} -\sum_{n=0}^{N} \sum_{k \subset K^+} \log(p_{nk}) \tag{2.3}$$

$$cce_{multi-class} = -\sum_{n=0}^{N} \sum_{k \subset K} y_{nk} \log(p_{nk}) \overset{(*)}{=} -\sum_{n=0}^{N} \sum_{k \subset K^+} \log(p_{nk}) \overset{(|K|=2)}{=} cce_{binary} \tag{2.4}$$

$$cce_{agg} \overset{(*)}{=} -\sum_{n=0}^{N} \log \left(\sum_{k \subset K^+} p_{nk} \right) \overset{\left(\forall n: \sum_{k=0}^{K} y_{nk} = 1\right)}{=} cce_{multi-class} \tag{2.5}$$

where K^+ is the set of positive classes with $y_k = 1$ and K^- is the set of positive classes with $y_k = 0$ and N is the number of observations. In the above example with lymphocytes that are split into T cells and B cells, $\sum_{k \subset K} y_k = 2$ for observations assigned as *lymphocytes*, as the label is $y_k = 2$ for both the T cell and B cell class which make up the set of lymphocytes K^+. Similarly, the predicted probability mass for an observation n labelled *lymphocytes* is the sum of probability masses predicted for T and B cells $\sum_{k \subset K^+} p_{nk}$. In contrast, *T cell* is a leaf node label, and its set of positive classes K^+ only contains the label *T cell*. Here, the predicted probability mass for the label *T cell* is $\sum_{k \subset K^+} p_{nk} = p_{nl}$ where l is the T cell class. The accuracy metric acc_{agg} corresponding to cce_{agg} is:

$$acc_{agg} = -\frac{1}{N} \sum_{n=0}^{N} I \left[\left(\sum_{k \subset K^+} p_{nk} y_{nk} \right) \succ \max_{k \subset K^-} (p_{nk}) \right] \tag{2.6}$$

The indicator function $I[]$ determines whether the aggerated probability mass predicted for a specific cell type label is greater than the probability mass assigned to any leaf node of the

ontology that is not a subclass of the class in question. Alternatively, one could use sigmoid transforms of independent cell type predictions. However, this approach does not take into account the prior knowledge that a cell can only belong to one class in an adequately defined cell type ontology. Therefore, we do not support this setting.

2.1.6. Model architectures

Our multilayer perceptron models for predicting cell types from gene expression data utilise fully connected layer stacks. For instance, a multilayer perceptron used in this study was trained on all protein-coding genes from either mouse or human, had one hidden layer with 128 units, was trained without *L1* and *L2* penalties on the parameters, and Scaled Exponential Linear Unit (SELU) activations.

We also defined a marker gene-dominated model for predicting cell types from gene expression data. In this model, a sigmoid function based on a gene-specific linear embedding of the gene expression values models an expression threshold. A fully connected layer then pools information from all genes to the cell type prediction.

AEs with fully connected layers and count noise distributions were proposed in [41] to learn embeddings of scRNA-seq data. The complete architectures are documented in the SFAIRA code. An example AE used in this study was trained on all protein-coding genes from either mouse or human, had three hidden layers of sizes 512, 64, and 512, was trained without *L1* and *L2* penalties on the parameters or input dropout, was trained with batch normalisation between fully connected layers, with SELU activations, and with a single trained dispersion parameter per gene in the output for the negative binomial reconstruction loss.

AEs with fully connected layers on count noise data were proposed in [40] to learn embeddings of scRNA-seq data. Here, we imposed a unit Gaussian prior on the latent space activations. The full architectures are reported in the SFAIRA code. An example VAE used in this study was trained on all protein-coding genes from either mouse or human, had three hidden layers of sizes 512, 64, and 512, was trained without *L1* and *L2* penalties on the parameters or input dropout, was trained with batch normalisation between dense layers, with a SELU activation function, and with a single trained dispersion parameter per gene in the output for the negative binomial reconstruction loss.

We used random projection as a baseline embedding model to contextualise our reported model performance. For this, we utilised the *SparseRandomProjection()* method from SCIKIT-

LEARN (v0.24.1) [210]. As with all other models, we fit the model on the training data and project the test data to reduced dimensions (64 in this case). We then reconstructed the original dimensionality of the data by multiplying the reduced data with the components of the fitted random projection model. To allow computation of the losses, we considered any negative values in the reconstruction invalid and converted them to a small positive number (1e-10). We did the same for any zero values in the reconstruction. We then computed the mean squared error of the reconstruction and the negative log-likelihood of the negative binomial distribution with a constant scale of 1.0.

2.1.7. Data curation

The data included in this study at the time of publication, including human [9, 47, 74, 97, 211–243] and mouse [244–247] datasets, were obtained in the least-processed expression matrix format provided by the dataset authors. The processing details are documented in the respective data loaders within SFAIRA. For instance, the datasets used for zero-shot analyses (Fig. 3.4b, Supplementary Fig. B.1) were obtained using SCANPY [35] or from CELLxGENE data collections [248–251], as described in the accompanying notebooks. No processing was done except for applying the preprocessing layers discussed in Section 2.1.3. We selected the protein-coding genes from the *Mus_musculus.GRCh38.102* genome assembly for mice and *Homo_sapiens.GRCh38.102* for humans as the feature space.

It is important to note that the datasets listed here use many different cell type label conventions. We mapped the cell type annotation from each dataset to the cell ontology and identified the label space of each cell type predictor model per anatomic location based on the most fine-grained cell types observed in that dataset. These cell type labels are leaf nodes of a sub-graph that describes all cells observed in a given tissue and their ontological relationships if one considers the directed acyclic graph of the ontology. We evaluated the accuracy and loss of coarser labels during testing and evaluation using the aggregated cross-entropy and accuracy metrics described in Section 2.1.5.

Where available, we held out complete datasets to evaluate model performance test metrics. However, some organs were only represented by a single dataset in the data zoo. In such cases, we held out a random set of 20 % of all cells as test data.

2.2. An integrated transcriptomic cell atlas of human neural organoids

This method section corresponds to, and is in part identical to, the one presented in the following publication:

He, Z.*, **Dony, L.***, Fleck, J. S.*, Szałata, A., Li, K. X., Sliškovic, I., Lin, H. -C., Santel, M., Atamian, A., Quadrato, G., Sun, J., Pasca, S. P., Camp, J. G., Theis, F. & Treutlein, B. An integrated transcriptomic cell atlas of human neural organoids. *In revision.* Preprint [204] doi: 10.1101/2023.10.05.561097 (2023)

"*" denotes an equal contribution.

2.2.1. Collecting, curating, and harmonising 36 human neural organoid scRNA-seq datasets

We collected data from 33 public human neural organoid datasets, obtained from 25 publications [96, 97, 99, 101, 103, 104, 116–125, 133, 142, 193, 252–257], and combined them with three unpublished datasets to create our atlas (Supplementary Table 1). To curate these neural organoid datasets, we used the SFAIRA framework [258] (GITHUB *dev* branch, 18th April 2023). We obtained scRNA-seq count matrices and associated metadata from each publication's data availability section or directly from the authors for unpublished data. We harmonised the metadata according to SFAIRA standards (`https://sfaira.readthedocs.io/en/latest/adding_datasets.html`) and added an extra metadata column called *organoid_age_days*, describing the number of days the organoid had been cultured for before collection.

Afterwards, we excluded any non-applicable subsets from the published datasets. These included diseased samples and samples with disease-associated mutations (*Huang et al.* [121], *Sawada et al.* [193], *Khan et al.* [252], *Bowles et al.* [253], *Paulsen et al.* [142]), fused organoids (*Birey et al.* [118]), primary fetal data (*Bhaduri et al.* [103], *Uzquiano et al.* [257]), hormone-treated samples (*Kelava et al.* [256]), data collected before neural induction (*Kanton et al.* [97], *Fleck et al.* [254]), and share-seq data (*Uzquiano et al.* [257]). We then used all the remaining datasets to create a single *AnnData* object [259] by harmonising them to a common feature space using any genes of the *protein_coding* and *lncRNA* biotypes from ENSEMBL release 104 [260] while filling any missing genes in a dataset with zero counts.

2.2.2. Data preprocessing

All data processing and analysis were done using SCANPY [35] (v1.9.3), except where otherwise stated. For QC and filtering of the HNOCA, we removed any cells expressing less than 200 genes. We then removed outlier cells based on two QC metrics: the number of expressed genes and the percentage of mitochondrial counts. We first applied a z-transformation to the relevant values across all cells to identify outlier cells. Cells with any z-transformed metric less than -1.96 or greater than 1.96 were identified as outliers. We fitted a Gaussian distribution to the histogram denoting the number of expressed genes per cell for any dataset collected using the v3 chemistry by 10x Genomics and containing more than 500 cells after filtering. If a bimodal distribution was detected, we removed any cell with fewer genes expressed than defined by the valley between the two maxima of the distribution. We normalised the raw read counts of all Smart-seq2 data by dividing it by the maximum gene length for each gene provided by BIOMART [261]. Next, we multiplied these normalised read counts by the median gene length across all genes in the datasets. We treated those length-normalised counts equivalently to raw counts from the UMI-based datasets in our downstream analyses.

We then log-normalised the expression matrix by dividing the raw counts for each cell by the total counts in that cell and multiplying by a factor of 1,000,000 before taking the natural logarithm of each count+1. We computed 3000 highly variable features in a batch-aware manner using the SCANPY *highly_variable_genes()* function (flavor = "seurat_v3", batch_key = "bio_sample"). Here, *bio_sample* represents biological samples as provided in the original metadata of the datasets. We used these 3000 features to compute a 50-dimensional PCA representation of the data, which we used to compute a kNN graph (n_neighbours = 30, metric = "cosine"). Using the kNN graph, we computed a two-dimensional UMAP embedding [208] of the data and a coarse (resolution = 1) and fine (resolution = 80) clustering of the unintegrated data using *Leiden* clustering [262].

2.2.3. Automatic marker-based cell type annotation with snapseed

We developed an auto-annotation strategy to obtain initial annotations for semi-supervised integration. Firstly, we created a hierarchy of cell types, including progenitors, neurons, and non-neural types, defined by a set of marker genes (available together with the analysis code on GITHUB, as stated in Section 3.2). We then used the reference similarity spectrum (RSS) method [97] to anchor our data in a recently published human fetal brain cell atlas [46] and constructed a kNN graph (k = 30) in the RSS space. We clustered the dataset using the *Leiden*

algorithm [262] (resolution = 80). We used the GPU-accelerated RAPIDS [263] implementation provided through SCANPY for both steps.

We computed the area under the receiver operating characteristic curve (AUROC) and the detection rate across clusters for all cell type marker genes on a given level in the hierarchy. We then calculated a score for each cell type by multiplying the maximum AUROC with the maximum detection rate among its marker genes. Each cluster was assigned to the cell type with the highest score. We performed this procedure recursively for all levels of the hierarchy. We used the same procedure for the fine (resolution = 80) clustering of the unintegrated data to obtain cell type labels as a ground-truth input for downstream benchmarking integration methods.

We implemented this auto-annotation strategy in the SNAPSEED PYTHON package, which is available on GITHUB (`https://github.com/devsystemslab/snapseed`). SNAPSEED is a package to enable scalable marker-based annotation for atlas-level datasets where manual annotation is not feasible. The package consists of three main functions: *annotate()* for non-hierarchical annotation of a list of cell types with defined marker genes, *annotate_hierarchy()* for annotating more complex, manually defined cell type hierarchies, and *find_markers()* for fast discovery of cluster-specific features. All functions are based on a GPU-accelerated implementation of AUROC scores using JAX (`https://github.com/google/jax`).

2.2.4. Semi-supervised data integration with scPoli

Conditional variational autoencoder (CVAE)-based methods have become a popular tool for scRNA-seq data integration. They are based on a VAE architecture, where the high-dimensional gene expression data input is first compressed into a low-dimensional latent representation by the encoder network using a standard-normal prior and afterwards reconstructed by the decoder network. To remove confounding batch effects from the data, CVAEs introduce a condition covariate that is fed into the encoder and decoder network, identifying the corresponding batch of each cell. Because of this additional information, the latent representation no longer contains the batch information and can be used as an integrated data embedding. SCPOLI [82] is a recent CVAE-based data integration tool, which uses learnable conditional embeddings and prototypes for biological covariates to be conserved during training, such as cell type labels. We used the SCPOLI [82] model from the SCARCHES [42] package to integrate the HNOCA datasets. To define the batch covariate for integration, we concatenated the dataset identifier (*id*) with the annotation of biological replicates (*bio_sample*)

and technical replicates (*tech_sample*), resulting in 396 batches. The model represents the batch covariate as a learned vector of size 5. We used the top three RSS-based SNAPSEED cell type annotation levels as the cell type label input for the scPOLI prototype loss. We chose a hidden layer size of 1024 for the one-layer scPOLI encoder and decoder, respectively, and the latent embedding dimension as 10. We used a value of 100 for the *alpha_epoch_anneal* parameter. We did not use the unlabelled prototype pretraining. We trained the model for seven epochs, including five pretraining epochs.

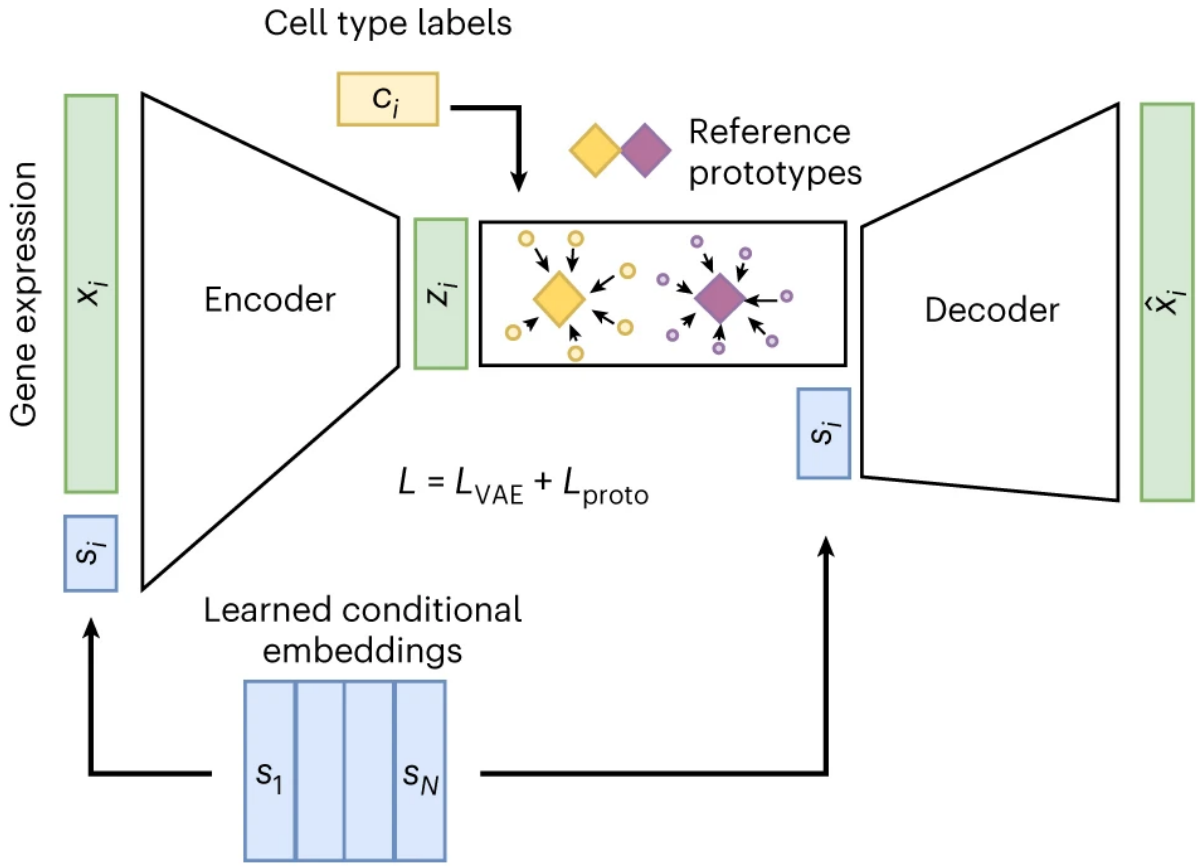

Figure 2.1.: **The scPoli model architecture enables semi-supervised data integration.** scPOLI uses an extended CVAE architecture to remove confounding batch variation with learnable conditional embeddings. An additional prototype loss enables the inclusion of cell type label information in the training procedure to create a final embedding that groups the same cell types together across datasets. L, total loss; L_{VAE}, CVAE loss; L_{proto}, prototype loss. Adapted from [82] (Licensed under the Creative Commons Attribution 4.0 International license: `https://creativecommons.org/licenses/by/4.0/`).

2.2.5. Benchmarking of data integration

Due to the plethora of available scRNA-seq data integration tools, we resorted to a structured benchmark to identify the most suitable model for our data. Please refer to [78] for an in-depth explanation of the applied benchmark. We used the GPU-accelerated scib-metrics Python package (v0.3.3, `https://github.com/YosefLab/scib-metrics`) to quantitatively compare the results of integrating the organoid atlas with different tools. We selected the method with the highest overall performance. We compared the data integration performance across the following integration approaches: unintegrated PCA, RSS [97] integration, scVI [40] integration (default parameters except for using 2 layers, latent space of size 30, and negative binomial likelihood), scANVI [81] integrations using either the snapseed level 1, 2, or 3 annotation as cell type label input, scPoli [82] integrations using either the snapseed level 1, 2, or 3 annotation or all three annotation levels at once as cell type label input, scPoli integrations of meta-cells aggregated with the *aggrecell* algorithm (first employed as *pseudocell* [97]) using either the snapseed level 1 or 3 annotation as cell type label input to scPoli.

We used several scores [78] provided by the scib-metrics package to determine integration quality, including *Leiden normalised mutual information score*, *Leiden adjusted rand index*, *average silhouette width* per cell type label, *isolated label score* (*average silhouette width*-scored), and cell type *local inverse Simpson's index* to quantify conservation of biological variability. To quantify batch effect removal, we used *average silhouette width* per batch label, integration *local inverse Simpson's index*, *kNN batch-effect test* (also known as *kBET)* score, and *graph connectivity*. We ranked integration approaches by an aggregate total score of all metrics, individually normalised into the range between zero and one.

Before benchmarking, we removed any cells from the dataset with the identical latent representation as another cell across each integrated embedding. We repeated this process until none of the latent representations contained any more duplicate rows, removing a total of 3293 duplicate cells (0.002 % of the entire dataset). This was necessary for the benchmarking algorithm to complete without errors. We used the snapseed level 3 cell type annotation, computed on the unintegrated PCA embedding as ground-truth labels for the benchmark.

2.2.6. Pseudo time inference

Neural optimal transport is a powerful new framework that can be applied to compute trajectories in scRNA-seq data by transporting cells across time points while minimising

displacement costs. Please refer to [264] for an in-depth explanation of this framework. We used a real-time-informed pseudo time based on neural optimal transport [264] in the scPoli latent space to infer a global ordering of differentiation state. We grouped organoid age into seven bins using the number of days in culture: (0, 15], (15, 30], (30,60], (60, 90], (90, 120], (120, 150], (150, 450]). We solved a *temporal neural problem* using moscot [265]. We used the *score_genes_for_marginals()* method to get proliferation and apoptosis scores for each cell and scored the marginal distributions based on expected proliferation rates. We then calculated marginal weights as $\exp(4(prolif\_score - apoptosis\_score))$.

We solved the optimal transport problem using the following parameters: iterations = 25000, compute_wasserstein_baseline = False, batch_size = 1024, patience = 100, pretrain = True, train_size = 1. To determine displacement vectors for each cell in age bin i, we used the subproblem corresponding to the $[i, i+1]$ transport map. We used the subproblem $[i-1, i]$ for the last age bin. We calculated displacement vectors by subtracting the original cell distribution from the transported distribution. We used the velocity kernel from CellRank [63] to calculate a transition matrix from displacement vectors. We then used this matrix as an input for computing diffusion maps [57]. Lastly, we determined pseudo-temporal ordering by ranking the negative of the first diffusion component.

2.2.7. Preprocessing primary fetal brain data

We obtained the scRNA-seq data of the primary fetal brain atlas [46] processed with Cell Ranger from GitHub (`https://storage.googleapis.com/linnarsson-lab-human/human_dev_GRCh38-3.0.0.h5ad`). We removed cells with fewer than 300 detected genes for QC. We normalised the transcript counts for each cell by the total count, scaled them by a factor of 10,000, and log-transformed the resulting values. The feature set was intersected with all genes detected in the organoid atlas, and using *Donor* as the batch key, we selected the 2000 top highly variable genes (HVGs) with the scanpy function *highly_variable_genes()*. Moreover, we created an additional column, *neuron_ntt_label*, to represent the neurotransmitter transporter (NTT) subtype labels that were automatically classified based on the cell cluster metadata *AutoAnnotation* column (`https://github.com/linnarsson-lab/developing-human-brain/files/9755350/table_S2.xlsx`).

2.2.8. Mapping the organoid atlas to the primary reference data

We compared our organoid atlas with data from the abovementioned primary human brain atlas [46] using scArches [42]. scArches is a query-to-reference mapping tool that uses a method called architecture surgery on CVAE-type embedding models to project query data onto reference atlases. See Fig. 2.2 for an illustration and outline of the architecture surgery workflow. First, we pre-trained a scVI model [40] on the primary atlas with *Donor* as the batch key. The model was constructed with the following parameters: n_latent = 20, n_layers = 2, n_hidden = 256, use_layer_norm = "both", use_batch_norm = "none", encode_covariates = True, dropout_rate = 0.2, and trained with a batch size of 1024 for a maximum of 500 epochs with an early stopping criterion. Next, we finetuned the model with scANVI [81], using *Subregion* and *CellClass* as cell type labels, with a batch size of 1024 for a maximum of 100 epochs, with an early stopping criterion and n_samples_per_label = 100. We used the scArches [42] implementation provided by scVI-tools [266, 267] to project the organoids atlas to the primary atlas. The query model was finetuned with a batch size of 1024 for a maximum of 100 epochs, with an early stopping criterion and weight_decay = 0.0.

2.2.9. Construction of the bipartite weighted kNN graph

We used the joint embedding of the primary reference [46] and query (HNOCA) data to compute an unweighted bipartite kNN graph by identifying the 100 nearest neighbours of each query cell in the reference data. This computation was done using either PyNNDescent or RAPIDS-cuML (`https://github.com/rapidsai/cuml`) in Python, depending on the availability of GPU acceleration. Similarly, we built a reference kNN graph by identifying the 100 nearest neighbours of each reference cell in the reference data. We used the Jaccard index ($J(A, B) = |AB||AB|$) to represent the similarity of the reference neighbours of two connected cells for each edge in the reference-query bipartite graph. We used the square of the Jaccard index as the weight of the edge between reference and query in the bipartite weighted kNN graph.

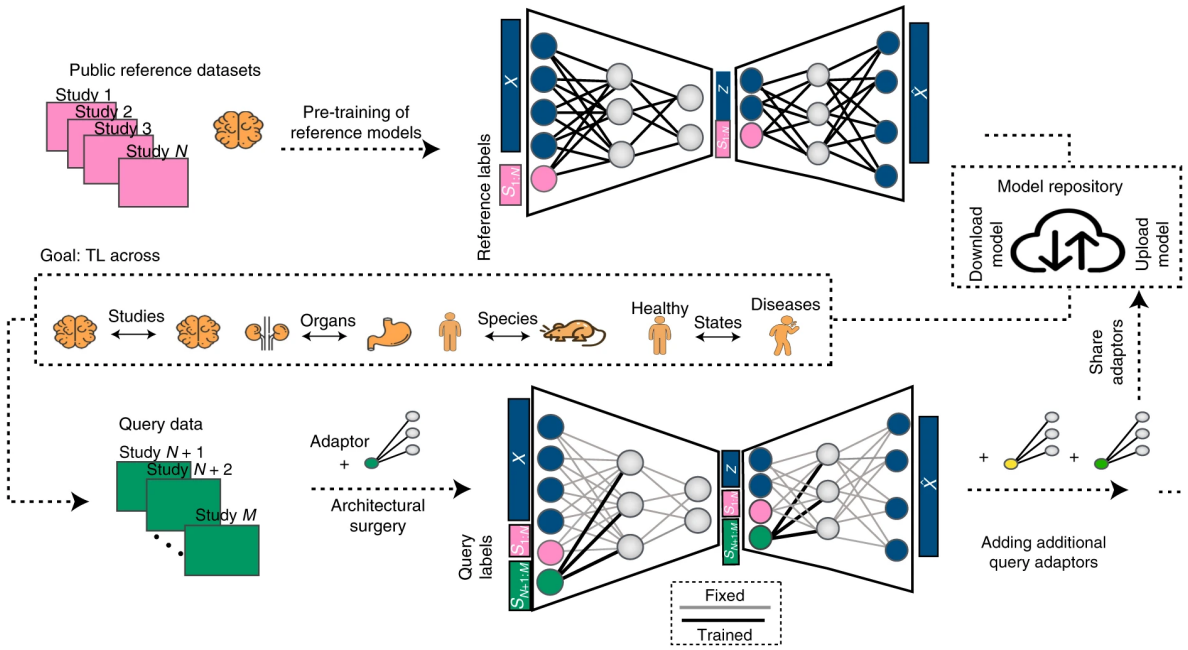

Figure 2.2.: **The scArches query-to-reference mapping pipeline.** scArches uses a CVAE-based model (as chosen by the user) that is trained with the reference data. When mapping query data onto this reference using architecture surgery, the reference network weights are fixed, and a query adapter is added to the network architecture. During query mapping, only the weights of the adapter are trained. After query training, the reference and query data can be projected together through the network, yielding an integrated embedding. Adapted from [42] (Licensed under the Creative Commons Attribution 4.0 International license: `https://creativecommons.org/licenses/by/4.0/`).

2.2.10. Transferring labels from primary to organoid data using the weighted kNN graph

The weighted kNN graph estimated between the primary reference [46] and query (HNOCA) allows transferring categorical metadata labels of reference to query cells via majority voting. We calculated the support of each category for every query cell as the sum of weights of edges that link to reference cells in that category. The query cell was assigned the category with the largest support.

We constrained the transfer procedure for the final regional labels of non-telencephalic NPCs and neurons and the NTT labels for non-telencephalic neurons. For regional labels, only non-telencephalic regions with the finest resolution, such as the hypothalamus, thalamus, midbrain dorsal, midbrain ventral, cerebellum, pons, and medulla, were considered as valid categories for transfer. To transfer NTT labels, we first identified valid region-NTT label pairs in the reference based on the provided NTT labels in the reference *neuroblast* and *neuron* clusters. Their most common regions were then re-estimated hierarchically to the abovementioned finest resolution. While transferring NTT labels, we considered only valid NTT labels for the repective brain region during majority voting for each non-telencephalic neuron with the regional label transferred.

2.2.11. Computing organoid presence scores for primary developing brain cell types

In simple terms, the presence score is a score given to each cell in the reference dataset that describes the chance of that cell type or state appearing in the query dataset. Here, we calculated the presence scores of primary atlas cells in each HNOCA dataset to determine how often a cell type or state represented by each primary cell appeared in each HNOCA dataset.

Specifically, we subset the weighted kNN graph for each HNOCA dataset and calculated the raw weighted degree for each cell in the primary atlas as the sum of the weights of the remaining edges linked to the cell. We applied a random walk with restart procedure to smooth the raw scores across the kNN graph of the primary atlas. For this, we first row-normalised the reference data kNN graph adjacency matrix (**A**), to convert it into a transition probability matrix (**P**). With the raw scores represented as a vector s_0, in each iteration t, we generated s_t as $s_t = as_0 + (1-a)P^T s_{t-1}$.

We repeated the described procedure 100 times to obtain smooth presence scores, which we then log-transformed. We trimmed scores lower than the 5th or higher than the 95th percentile. The trimmed scores were normalised to fall within the range of 0 to 1, and these became the final presence scores for the respective HNOCA dataset. The maximum presence score for the entire HNOCA dataset was calculated as the highest of all the presence scores for each cell in the primary atlas. A high maximum presence score (close to 1) indicates that the cell type or state appeared frequently in at least one HNOCA dataset. In contrast, a low maximum presence score (close to 0) suggests that the cell type or state was under-represented in all HNOCA datasets.

2.2.12. Morphogen effects on cell type composition

We used the PERTPY (`https://github.com/theislab/pertpy`) implementation of the scCODA algorithm [268] to test for changes in cell type composition upon administering morphogens from different organoid differentiation protocols. scCODA is a Bayesian model to detect and quantify compositional changes in scRNA-seq data. We extracted information about the added morphogens from each differentiation protocol. We grouped them into 15 broad molecule groups based on their role in neural differentiation (Supplementary Table S1) and used them as covariates in the model. We used the region labels transferred from the primary atlas as *cell_type_identifier* labels in the analysis, using the coarse level 1 cell type labels for cell types without regional identity. We removed the very early developmental cell types, pluripotent stem cells and neuroepithelium cells, from the dataset for this analysis. We used our *bio_sample* annotation as the *sample_identifier* in scCODA. We ran the algorithm sequentially with default parameters, using No-U-turn sampling (*run_nuts()* function). We used each cell type once as a reference to apply a majority vote-based system for identifying the cell types that were credibly differential in more than half of the iterations.

Additionally, we developed a second approach to test for differential composition using regularised linear regression. Our technique involved fitting a generalised linear model with the region composition matrix as the response variable Y and molecule usage as the independent variable X: $Y \sim X\beta$

We used lasso regularisation ($\alpha = 1$) with Gaussian noise and an identity link function. The regularisation parameter λ was determined automatically through cross-validation, implemented in the function *cv.glmnet()* from the GLMNET [269] R package. We considered all non-zero coefficients β as indications of enrichment and depletion.

2.2.13. Identifying systematic pathway differences between organoids and primary fetal cells

To compare the transcriptomes of organoid and primary cells, we analysed the HNOCA dataset by selecting only the cells labelled *Neuron* based on the final level 1 annotation. We also used the human developing brain atlas, focusing on cells with a valid label in the *neuron_ntt_label* annotation column. We also included two other fetal cortical cell datasets [48, 98]. The data from *Camp et al.* [98] was further subsetted to cells labelled *fetal*, and we estimated the transcripts per million reads for each gene in each cell using RSEM [270] based on the STAR [271] mapping results. We performed PCA, kNN graph construction, UMAP [208], and *Leiden* clustering [262] (resolution 0.2) using SCANPY and identified the cluster with the highest STMN2 and NEUROD6 expression as the cortical neuron cluster and used only those cells. The data from *Eze et al.* [48] was subsetted to cells annotated as *Neuronal* in the publications' Supplementary Table 5 (*Cortex Annotations*). We then performed PCA, kNN graph construction, and UMAP [208] to visualise the dataset. We found that only samples from the individuals *CS14_3*, *CS20*, *CS22*, and *CS20* contained detectable expression of STMN2 and NEUROD6, so we further subsetted the dataset to cells from those individuals.

To compare the gene expression differences between HNOCA cells and their primary counterparts, we combined cells of the same regional neuronal cell type into pseudobulk samples using the PYTHON implementation of DECOUPLER [272] (v1.4.0). We added up the counts for each sample and discarded any samples with less than ten cells or a total count of fewer than 1000 (sample annotation columns: *batch* for HNOCA, *SampleID* for the human developing brain atlas, *sample* for *Camp et al.* and *individual* for *Eze et al.*). We then selected the intersection of features in all datasets and removed cells with fewer than 200 genes expressed. We also removed any genes expressed in less than 1 % of neurons in HNOCA and any genes located on the X and Y chromosomes. We removed samples from the pseudobulk data associated with an organoid differentiation assay that contributed less than two total samples or fewer than 100 total cells for each regional neural cell type.

We used the EDGER [273] gene-wise negative binomial generalised linear model with quasi-likelihood F-tests to iteratively compute DEGs between each organoid differentiation protocol and primary cells of the matching regional neural cell types while correcting for organoid age in days, the number of cells per pseudobulk sample, as well as median and standard deviation of the number of detected genes per pseudobulk sample. We used the human developing brain atlas data, *Eze et al.* data, and *Camp et al.* data as primary data for the DE comparison in the cell type *Dorsal Telencephalic Neuron NT-VGLUT*. We used the human

developing brain atlas as the sole reference for all other cell types. To determine which genes were significantly differentially expressed, we checked their false discovery rate (FDR) (Benjamini-Hochberg)-corrected p-value, which had to be below 0.05, and their absolute log2 fold-change (log2FC), which had to be above 0.5.

We used the GSEApy [274] Python package with the *GO_Biological_Process_2021* gene set to perform a functional enrichment analysis on our DE results.

2.2.14. Comparing the transcriptomic fidelity of different neuronal organoid cell types to their primary counterparts

To estimate the similarity of neurons in HNOCA and their primary counterparts, we started by computing the average gene expression for each neural cell type in the reference data [46] and the HNOCA. We only considered the neural cell types of an HNOCA dataset if it contained at least 20 cells. We used a Chi-squared test-based variance ratio test on the generalised linear model with a Gamma distribution (identity link) to identify HVGs across the neural cell types in the primary reference. We set the coefficient of variance of transcript counts across neural cell types as the response and the reciprocal of average transcript counts across neural cell types as the independent variable. We considered genes with Benjamini-Hochberg adjusted p-values less than 0.01 as highly variable. Finally, we calculated the similarity between one neural cell type in the primary atlas and its counterpart in each HNOCA dataset as the Spearman correlation coefficient across the identified HVGs.

We determined the similarity of core transcriptomic identity between organoid and fetal cells by comparing the co-expression of TFs obtained from the AnimalTFDB 4.0 database [275]. We focused on TFs in the HVG list and computed Spearman correlations for the comparison.

We identified cells under metabolic stress using the scanpy *score_genes()* function with default parameters to calculate scores for the *canonical glycolysis* gene set obtained from the enrichR [276, 277] *GO_Biological_Process_2021* database. We included all neuronal cells from HNOCA and primary references of *Braun et al.* [46], *Eze et al.* [48], and *Camp et al.* [98] in this analysis.

In order to determine the robustness of the identified difference between the correlation of glycolysis scores and the similarity of the entire transcriptome versus the correlation of glycolysis scores and the similarity of the core transcriptome (based on TFs), we created 100 sets of HVGs, each of the same size as the highly variable TFs. We computed transcriptomic

similarities using these sets and then evaluated their correlation with the glycolysis scores.

2.2.15. Reference mapping of the neural organoid morphogen screen scRNA-seq data to the human developing brain atlas and HNOCA

We mapped scRNA-seq data from the neural organoid morphogen screen to two of our models with scArches [42]: the scANVI [81] model of the human developing brain atlas [46] and the scPoli [82] model of the HNOCA. See Fig. 2.2 for an overview of the scArches reference mapping workflow. We used the *dataset* field of the screen data as the batch covariate, containing the following categories: *organoid screen*, *secondary organoid screen*, or *fetal striatum 21pcw*. To map to the primary reference, we used the scVI-tools [266] implementation of scArches without using cell type annotations and trained the model for 500 epochs with weight_decay = 0 and otherwise default parameters. To map to the HNOCA, we utilised scArches through scPoli and trained the model for 500 epochs without unlabelled prototype training.

2.3. Chronic exposure to glucocorticoids amplifies inhibitory neuron cell fate during human neurodevelopment in organoids

This method section corresponds to, and is in part identical to, the one presented in the following publication:

Dony, L., Krontira, A. C., Kaspar, L., Ahmad, R., Demirel, I. S., Grochowicz, M., Schäfer, T., Begum, F., Sportelli, V., Raimundo, C., Koedel, M., Labeur, M., Cappello, S., Theis, F. J., Cruceanu, C.*, Binder, E. B.* Chronic exposure to glucocorticoids amplifies inhibitory neuron cell fate during human neurodevelopment in organoids. *In review.* Preprint [205] doi: 10.1101/2024.01.21.576532 (2024)

"*" denotes an equal contribution.

2.3.1. Stem cell culture

We used two human iPSC lines for this study. The first cell line was reprogrammed using human iPSCs from skin fibroblasts (*HPS0076:409b2*, RIKEN BRC cell bank, female) [278, 279]. Here, we refer to it as *Line 409b2*. The second cell line was reprogrammed using a plasmid-based protocol for integration-free human iPSCs from peripheral blood mononuclear cells (PBMCs) from a female donor through the *BeCOME* study [280]. Here, we refer to it as *Line FOK4*. MTA approvals were obtained for both human iPSC lines. Human iPSCs were cultured in Matrigel-coated (1:100 diluted in *DMEM-F12* (Gibco™, 31330-038), Corning Incorporated, 354277) *Costar®* 6-well cell culture plates (Corning Incorporated, 3516) in *mTESR1 Basal Medium* (STEMCELL Technologies, 85851) supplemented with *1x mTESR1 Supplement* (STEMCELL Technologies, 85852) at 37 °C with 5 % CO_2. Passaging was performed with *Gentle Cell Dissociation Reagent* (STEMCELL Technologies, 07174). *RevitaCell Supplement* (1:100 diluted, Gibco™, A2644501) was added for 24 hours after passaging to promote cell survival.

2.3.2. Generating neural organoids

With some modifications, we generated unguided human neural organoids as previously described [113]: Human iPSCs were dissociated in *StemPro Accutase Cell Dissociation Reagent*

(Life Technologies, A1110501). Single cells (n = 9000) were dispensed into each well of an Ultra-low attachment 96-well plate with round bottom wells (Corning Incorporated, 7007) in *human embryonic stem cell medium* (*DMEM/F12-GlutaMAX* (Gibco™, 31331-028) with 20 % *Knockout Serum Replacement* (Gibco™, 10828-028), 3 % *Fetal Bovine Serum* (Gibco™, 16141-061), 1 % *non-essential amino acids* (Gibco™, 11140-035), 0.1 mM *2-mercaptoethanol* (Gibco™, 31350-010)) supplemented with 4 ng/ml human recombinant *Fibroblast Growth Factor* (Peprotech, 100-18B) and 50 μM *Rock inhibitor Y27632* (Millipore, SCM075) for 4 days and in human embryonic stem cell medium without bFGF and Rock inhibitor for an additional 2 days to form embryoid bodies. On day 6, the medium was changed to neural induction medium (*DMEM/F12 GlutaMAX* supplemented with 1:100 *N2 supplement* (Gibco™, 15502-048), 1 % *non-essential amino acids* and 1 μg/ml *Heparin* (Sigma, H3149)) and cultured for an additional 6 days. On day 12, the embryoid bodies were embedded in *Matrigel* (Corning Incorporated, 354234) drops and transferred to 10-cm cell culture plates (TPP, 93100) in neural differentiation medium without vitamin-A (*DMEM/F12GlutaMAX* and *Neurobasal* (Gibco™, 21103-049) in a 1:1 ratio, additionally supplemented with 1:100 *N2 supplement* 1:100 *B27 without Vitamin A* (Gibco™, 12587-010), 0.5 % *non-essential amino acids*, *insulin* 2.5 μg/ml (Gibco™, 19278), 1:100 *Antibiotic-Antimycotic* (Gibco™, 15240-062) and 50 μM *2-mercaptoethanol*) for 4 days. On day 16, Organoids were transferred onto an orbital shaker in NDM+A medium (same composition as *neural differentiation medium without vitamin-A* with the addition of *B27 with Vitamin A* (Gibco™, 17504-044) in place of *B27 without Vitamin A*) and were grown in these conditions at 37°C with 5 % CO_2. NDM+A medium was changed twice weekly until the organoids were collected for cryopreservation, single-cell dissociation, or fixation in paraformaldehyde.

To validate the inhibitory-excitatory neural lineage effects of GCs, we generated guided ventral organoids as previously described [281]: Embryoid bodies were formed starting from iPSCs dissociated into single cells using *Accutase* (Sigma-Aldrich, A6964) (n = 9,000). Five days later, to induce brain regionalisation during the neuronal induction, embryoid bodies were treated individually with *SAG* (1:10,000) (Millipore, 566660) and *IWP-2* (1:2,000) (Sigma-Aldrich, I0536) for ventral identity and with *cyclopamine A* (1:500) (Calbiochem, 239803) for dorsal identity. All other culture parameters were identical to the ones described above for unguided organoids.

2.3.3. Generating and validating a neuron-specific fluorescent reporter cell line

We used Line 409b2 human iPSCs to generate an enhanced green fluorescent protein (eGFP)+/GAD1+ heterozygous iPSC line. Guide RNA (gRNA) (crRNA and tracrRNA,

IDT) for editing with the recombinant S.p. HiFi Cas9 Nuclease V3 protein (IDT) was selected to cut efficiently at a short distance from the ATG start codon of the GAD1 gene by using the Benchling web tool (`https://benchling.com`). A 1611 nt donor single-stranded oligo DNA nucleotides (ssODNs) (IDT) for homology-directed recombination was designed to have homology arms of 222-300 nt on either side of the insert DNA, a 717 nt sequence encoding for eGFP followed by the 3'UTR and the polyA signal. Lipofection (reverse transfection) was performed using the alt-CRISPR manufacturer's protocol (IDT) with a final concentration of 10 nM of the gRNA, ssODN donor, and Cas9. In brief, 0.75 μL *RNAiMAX* (Invitrogen, 13778075) and the RNP mix (gRNA, ssODN, and Cas9 protein) were separately diluted in 25 μL *OPTI-MEM* (Gibco, 1985-062) each and incubated at room temperature for 5 min. Both dilutions were mixed to yield 50 μL of *OPTI-MEM*. The lipofection mix was incubated for 20–30 min at room temperature. Cells were dissociated with Accutase (Life Technologies) for 6 min and counted during incubation. The lipofection mix, 100 μL containing 50,000 dissociated cells in mTeSR1 supplemented with *RevitaCell* (1:100, Gibco) and the 2 μM *M3814 NHEJ inhibitor* [282] was thoroughly mixed and placed in 1 well of a 96-well plate covered with *Matrigel matrix* (Corning, 35248). The media was exchanged to regular *mTeSR1 media* (StemCell Technologies) containing the NHEJ inhibitor after 24 h. Single-cell–derived clonal cell lines were analysed and genotyped by PCR using genomic DNA isolated with *Quick-Extract DNA Extraction Solution* (Lucigen) and primers binding within and downstream the modified region (Primer 1) or in the HAs (Primer 2).

Appendix A.1 lists the exact nucleotide sequences used for this work.

2.3.4. Dexamethasone exposure

The organoids were subjected to GCs by dissolving dexamethasone in dimethyl sulfoxide and mixing it with the NDM+A culture medium. Dexamethasone was first diluted in dimethyl sulfoxide at 100 μM and then further diluted in the NDM+A culture medium to a final concentration of 100 nM. The organoids in the vehicle control group (referred to as *Veh* in this thesis) received equal amounts of dimethyl sulfoxide. We replaced the supplemented media every two days during chronic exposures (days 60-70 in culture). Some organoids were collected on day 70 after being exposed for ten days. Other organoids from the same batch were further cultured in regular unsupplemented media for a 20-day wash-out period. This was followed by an acute exposure (at day 90 in culture) to 100nM dexamethasone or dimethyl sulfoxide for 12 hours.

2.3.5. Immunohistochemistry

We fixed organoids with 4 % paraformaldehyde for 45 minutes at 4°C, cryopreserved them with 30 % sucrose, fixed them in optimal cutting temperature compound (Thermo Fisher Scientific), and stored them at -20°C before cutting and preparation of 16 μm cryosections on *SuperFrost* slides. For immunofluorescence, we postfixed sections using 4 % PFA for 10 mins and permeabilised them with 0.3 % Triton for 5 minutes. We subsequently blocked sections with 0.1 % TWEEN, 10 % Normal Goat Serum, and 3 % BSA. We diluted primary and secondary antibodies in blocking solution and visualised and analysed the fluorescent stainings with a Leica laser-scanning confocal microscope. Before fixing with paraformaldehyde, we retrieved antigens for staining with green fluorescent protein (GFP) and PBX3 or PAX6 and SATB2. More specifically, we incubated the slides in citric buffer (0.01 M, pH 6.0) for 1 min at 720 watts and 10 mins at 120 watts, left them to cool down at room temperature for 20 minutes, and washed them once with PBS. *Alexa-anti-chicken-488* and *Alexa-anti-rabbit-647* were used as secondary antibodies. All secondary antibodies are diluted to 1 μg/ml or 1:1000.

Appendix A.2 lists the specific antibodies used in this section.

2.3.6. Cell imaging and counting

We imaged the immunofluorescence slides using the confocal mode on either the *MICA Microhub* microscope (Leica) with 20x and 63x lenses or the *AxioScan.Z1* Slide Scanner (Zeiss) with a 20x lens. We used the *Leica Application Suite X* software (v1.4.4.26810) for imaging on the *MICA Microhub* microscope. To identify cells that were positively stained for PBX3 and GFP-GAD1, as well as cells that were double-positive for PBX3 and GFP-GAD1, two independent experimenters manually counted cells in two separate staining experiments. We counted cells using the *Cell Counter Tool* in ImageJ. We reported the counts as cells/mm^2 normalised by tissue surface area. We quantified the *AxioScan.Z1* Slide Scanner images as total counts per entire organoid slice (n = 5 per group). We quantified the *MICA Microhub* images as one representative selected tile per mosaic organoid slide (n = 10 tiles per group). We reported the statistical analyses as 2-sided unpaired T-tests.

2.3.7. scRNA-seq library preparation and sequencing

We used *StemPro Accutase Cell Dissociation Reagent* (Life Technologies) to break down individual cells, followed by filtration through 30 μM and 20 μM filters (Miltenyi Biotec). We then removed any debris using a *Percoll gradient* (Sigma). The single cells were resuspended in ice-cold Phosphate-Buffered Saline supplemented with 0.04 % Bovine Serum Albumin and prepared for single-cell separation. We conducted experiments in a paired case-control design with 2 or 4 conditions on day 70 and day 90, respectively, run in parallel. We ran the single cells through the 10x Chromium controller to form gel emulsion beads containing barcoded single cells, which were then prepared into single-cell libraries using the *Chromium Single Cell 3' Reagent Kit v2* according to the manufacturer's recommendations without any modifications (10x Genomics). To ensure an optimal cell number, we loaded 10,000 cells per sample onto a channel of the 10x chip. We evaluated all libraries using a *High Sensitivity DNA Analysis Kit* for the *2100 Bioanalyzer* (Agilent) and *KAPA Library Quantification kit for Illumina* (KAPA Biosystems). The sequencing of the 10x Genomics scRNA-seq was performed on an *Illumina NovaSeq 6000* (Illumina, San Diego, CA) by the sequencing core facility of the Max Planck Institute for Molecular Genetics (Berlin, Germany).

2.3.8. scATAC-seq library preparation and sequencing

We isolated individual cells from organoids using the scRNA-seq protocol described in the previous section. Subsequently, we prepared the nuclei and generated scATAC-seq libraries using the *Chromium Single Cell ATAC Library & Gel Bead Kit* (16 reactions, PN-1000110) without any modifications, following the manufacturer's instructions (10x Genomics). The sequencing of the 10x Genomics scATAC-seq libraries was carried out on an *Illumina NovaSeq 6000* (Illumina, San Diego, CA) by the sequencing core facility at the Max Planck Institute for Molecular Genetics (Berlin, Germany).

2.3.9. scRNA-seq quality control and filtering

We produced count matrices from FASTQ files using 10x Genomics CELL RANGER v3.0.2 [28] and the transcriptome *hg38_ensrel94* [283]. The count matrices of Line 409b2 and Line FOK4 acutely treated and control organoids (*Veh-Veh* and *Veh-Acu* conditions), aged 90 days, were previously used in a separate analysis [284] and are available from the *Gene Expression*

Omnibus repository (accession: GSE189534). We reprocessed and reanalysed all data using the SCVERSE [267] packages SCANPY [35] v1.9.3 and ANNDATA [259] v0.9.1 with PYTHON v3.10.12, unless stated otherwise.

For QC, we removed cells with less than 1200 UMI counts, cells with more than 150,000 UMI counts, cells with less than 700 genes expressed, and cells with 25 % or more mitochondrial UMI counts. Additionally, we removed any genes expressed in less than 20 cells after cell filtering.

For the second QC step, we computed an initial clustering of the entire dataset using *Louvain* clustering (LOUVAIN PYTHON package v0.8.0, `https://github.com/vtraag/louvain-igraph`) with appropriate preprocessing and a resolution of 0.5. We removed two samples (*409b2-D70-Chr-V1*, *409b2-D70-Veh-V2*) that mainly clustered separately from all other samples. Furthermore, we removed three additional samples (*409b2-D70-Veh-C1*, *FOK4-D90-Veh-Veh-C1*, *FOK4-D90-Veh-Veh-C2*) that had low numbers of expressed genes or a sequencing saturation below 35 %, as reported by 10x Genomics CELL RANGER.

For the final QC step, we reclustered the data using the same procedure as before and computed marker genes using the *rank_genes_groups()* function of SCANPY with default parameters. We removed any clusters containing mostly mesenchymal cells, epithelial cells, myocytes, neuroectoderm, neural stem cells, macrophages, or fibroblasts based on the marker gene signature.

2.3.10. Preprocessing scRNA-seq data and annotating cell types

We applied the following steps separately to the data obtained from Line 409b2 and Line FOK4.

We computed normalisation size factors using the SCRAN [285] R package v1.22.1 (R v4.1.2) with the appropriate preprocessing to obtain an initial coarse clustering of the data as required for this approach. We normalised the raw counts of each cell by the respective size factor and consequently $\log(1 + x)$-transformed them. We computed four thousand HVGs based on the log-normalised counts using the *cell_ranger* flavour [28]. We computed a PCA representation [210], kNN graph [208], force-directed graph drawing (`https://github.com/bhargavchippada/forceatlas2`) [286], *Louvain* clustering (`https://github.com/vtraag/louvain-igraph`) [262], and partition-based graph abstraction [59] with default parameters. We used the layout

obtained from plotting the partition-based graph abstraction results with a threshold of 0.05 as initialisation to compute a UMAP [208] with default parameters.

We ranked genes for each cluster and then sub-clustered or merged the clusters where appropriate. We assigned cell type identities to the clusters by comparing the top-ranking genes per cluster with known marker genes from developmental neurobiology. This resulted in identifying eight cell types and one cluster of unknown identity.

2.3.11. Filtering non-viable cells

We applied the following steps separately to the data obtained from Line 409b2 and Line FOK4.

To identify cells in non-viable metabolic states, we scored the following Gene Ontology Biological Process genesets [287, 288] using the respective SCANPY function with default parameters in every cell. As previously suggested [104], we used *glycolytic process* (GO:0006096) and *response to endoplasmic reticulum stress* (GO:0034976) as negative markers of cell viability and *gliogenesis* (GO:0042043), *neurogenesis* (GO:0022008), and previously reported marker genes of the ChP [124] as positive markers of cell viability. We scaled each score to the range (0,1). We computed a joint cell viability score by adding all scaled positive viability scores, subtracting the sum of all scaled negative viability scores, and scaling the final score to the range (0,1). Based on the final score, we identified the *Unknown* cell cluster as mostly non-viable and removed it from the dataset. Additionally, we identified cells with a final viability score of less than or equal to 0.4 as non-viable cells and removed them from the dataset.

Following the removal of non-viable cells, we computed HVGs, a PCA representation, a kNN graph, a force-directed graph drawing, partition-based graph abstraction (threshold 0.001 for computing the layout), and UMAP, following the same procedure described in the section above.

2.3.12. Reference mapping to the Human Neural Organoid Cell Atlas

We used SCPOLI [82] from the SCARCHES [42] package v0.5.9 to project the scRNA-seq data acquired in our study to the HNOCA using the query-to-reference-mapping. See Fig. 2.1

for an illustration of the scPoli architecture and Fig. 2.2 for an illustration of the scArches workflow. We obtained the HNOCA data and the scPoli integration model weights used in the original study from `https://github.com/theislab/neural_organoid_atlas`. We subset the feature space of our datasets to the feature space used in the obtained model and filled any missing genes with zero expression. We then trained the query model for five pre-training epochs and one training epoch with unlabelled prototype training enabled. We annealed the model hyperparameter *alpha* over 10 epochs and set the model hyperparameter *eta* to 5. After that, we fed the data acquired in our study and the HNOCA data through the trained model, which produced a mapped latent representation. We used this mapped latent representation as input for the kNN graph and UMAP computation. To contextualise the data generated in our study, we used cell type annotations from the HNOCA annotation column *annot_level_2*, as shown in Fig. 3.8 of this thesis.

2.3.13. Computing differentially expressed genes

We used the R tool MAST [289] v1.20.0 (R v4.1.2) to compute DEGs per cell type between treatment conditions on log-normalised expression data. We carried out this analysis separately for the two source cell lines and removed samples that contained less than ten cells of a given cell type. Additionally, we removed genes expressed in less than 5 % of cells of a given cell type to compute DEGs for that cell type. We fit a hurdle model for each cell type and source cell line according to the following formula:

$$\sim ngeneson + treatment\_acute + treatment\_chronic + treatment\_chronic : treatment\_acute$$

where *ngeneson* corresponds to the number of expressed genes in the sample, *treatment_chronic* corresponds to the 10-day treatment applied between day 60 and day 70 (*Veh* or *Chr*), and *treatment_acute* corresponds to the 12-hour treatment applied at day 90 (*Veh*, *Acu* or *None* for samples collected at 70 days in culture). We used a likelihood-ratio test to test for DEGs at D70 (chronic effect) and D90 (chronic and acute effect).

We considered only genes with a FDR-corrected p-value of less than 0.1 in both source cell lines and agreeing direction of log2FC as differentially expressed (*consensus DEGs*). This approach helped us reduce the number of false-positive DEGs. We used the UpSetPlot v0.8.0 Python package (`https://github.com/jnothman/UpSetPlot`) to visualize the DEGs.

2.3.14. Enrichment analyses

We computed consensus DEGs as described in the previous section and used them as input for the enrichment analysis. We performed the enrichment analysis using the Python implementation of enrichR [276, 277] via the GSEApy [274] package v1.0.5 with default parameters. For the annotation of biological function, we used the *GO_Biological_Process_2021* geneset provided by enrichR. We used the *ENCODE_and_ChEA_Consensus_TFs_from_ChIP-X* geneset provided by enrichR to perform TF target enrichments. For better coverage of TFs, we used a second database for this enrichment: *CollectTRI* [290], obtained via the Python implementation of decoupler [272] v1.4.0. We considered any hits with a FDR of less than 0.1 significantly enriched.

We summarised and visualised the Gene Ontology enrichment results using GO-Figure [291] v1.0.1. The *go.obo* version used was *releases/2021-05-01*, the *go.obo* version used to create GO relations was *releases/2023-04-01*, and the *similarity_cutoff* was 0.2.

2.3.15. Processing published neural organoid validation data

We downloaded count data, associated metadata, and gene names from ArrayExpress (accession: E-MTAB-7552) as stated in the data availability section of the associated publication [97]. We subsetted the dataset to cells from the study's 70-day-old organoid cell line comparison section and removed any cells from cell line 409b2 or without a cell type label. We also removed genes expressed in less than 10 remaining cells from the dataset. We then normalised raw counts per cell to the median total counts per cell in the dataset and $\log(1+x)$-transformed them. Next, we computed HVGs and a PCA representation, as with our original datasets. We computed an integrated kNN graph using the *batch-balanced kNN* algorithm [292] (bbknn Python package v1.5.1) using cell line as a batch key and neighbours_within_batch = 5 with otherwise default parameters of the scanpy *.external* implementation. We computed a UMAP representation and force-directed graph drawing from this integrated kNN graph with default parameters. In this dataset, we identified the *Cortical neurons* cluster and the *LGE interneurons* cluster as the excitatory and inhibitory neuron cell types, respectively.

2.3.16. Inferring trajectories and computing driver genes

We removed the *RGS5+ Neuron* cluster from the dataset for Line 409b2 and Line FOK4 data, followed by recomputing HVGs, a PCA representation, the kNN graph, a force-directed graph drawing, partition-based graph abstraction (threshold 0.001 for computing the layout), and a UMAP representation. We applied all the steps described in this section separately to the data derived from Line 409b2, Line FOK4, and the external validation data.

To compute a pseudo time, we used the SCANPY *.external* implementation of PALANTIR [58] following the manual selection of an *early cell* within a progenitor cluster for each dataset (RG for Line 409b2 and FOK4, Cortical NPCs for the validation data). First, we computed a PALANTIR diffusion map with five diffusion components and otherwise default parameters. Second, we computed a t-distributed stochastic neighbourhood embedding [293] representation on the first two components of the PALANTIR multi-scale data matrix with a perplexity of 150 and otherwise default parameters. We used the resulting embedding to compute the PALANTIR pseudo time, sampling 500 waypoints and otherwise default parameters.

For computing lineage probabilities based on the PALANTIR pseudo time, we used CELLRANK [63, 294]. We installed CELLRANK from the GITHUB *main* branch (`https://github.com/theislab/cellrank`) at commit *c3ced63* (earliest stable version including this commit: v2.0.1). We initiated the CELLRANK pseudo time kernel with the PALANTIR pseudo time and computed a transition matrix. This, in turn, was used to initiate the GPCCA [295] estimator, which allowed us to compute macrostates from the transition matrix. We manually selected inhibitory and excitatory neuron trajectory endpoints (plus an additional ChP endpoint in Line 409b2 and FOK4 data) from the computed macrostates and used them to compute each cell's fate probabilities and terminal state's fate. We further used CELLRANK to compute lineage drivers for each terminal state, correcting for FDR and discarding any drivers where the significance of the driver correlation could not be computed. We deemed a driver gene with an FDR below 5 % significant in all downstream analyses. We used the SCIPY [296] implementation of the t-test on two related samples of scores to compute the significance of the difference between the alignment of consensus DEGs and driver gene directionality between the excitatory and inhibitory neuron lineages across three datasets.

For visualising gene trends, we used the kNN-smoothing on the expression data, as implemented in the SCVELO [61] v0.2.5 *moments()* function, with 30 principal components and otherwise default parameters. Using this data, we fitted gene trends using the GAMR model

[297] with 7 knots and plotted them along PALANTIR pseudo time.

2.3.17. Processing published fetal brain reference data

We downloaded the CELL RANGER-processed count matrices from the recently published first-trimester fetal brain atlas [46] using the link provided by the authors (`https://storage.googleapis.com/linnarsson-lab-human/human_dev_GRCh38-3.0.0.h5ad`). We obtained the associated metadata for organoid age, 10x Chromium chemistry version, and NTT annotations from supplementary tables S1 and S2 of the publication: `https://github.com/linnarsson-lab/developing-human-brain/files/9755355/table_S1.xlsx` and `https://github.com/linnarsson-lab/developing-human-brain/files/9755350/table_S2.xlsx`. We removed any genes expressed in less than 20 cells and cells with less than 200 genes expressed in the dataset. Then, we normalised the total counts of each cell to 10,000 and $\log(1 + x)$-transformed them. Using the *CellClass* annotation provided by the authors, we subset the dataset to the clusters *Neuron*, *Neuroblast*, *Neuronal IPC*, and *Radial glia*.

Using the *Chemistry* annotation, we subset the dataset to cells collected by the 10x Genomics 3' v2 chemistry. We removed cells expressing neither the GABA NTT nor any glutamate NTTs from the dataset. We also removed cells expressing both the GABA NTT and any of the glutamate NTTs from the dataset. Finally, we created a new metadata column *NTT_simplified*, indicating whether the GABA NTT or any of the glutamate NTTs were expressed in each cell.

2.3.18. Preprocessing scATAC-seq data

We produced count matrices from FASTQ files using 10x Genomics CELL RANGER ATAC [298] v2.0.0 with the reference *GRCh38* (ENSEMBL release 94) [283]. Unless stated otherwise, we performed all downstream analyses using the R packages SIGNAC [299] v1.9.0 and SEURAT [219] v4.3.0 (R v4.1.2). We loaded the aggregated and filtered peak-barcode matrix from 10x Genomics CELL RANGER ATAC and the associated fragments file and metadata with SIGNAC. We discarded any features detected in less than 10 cells and any cells with less than 200 detected features from the dataset. We used gene annotations from the *EnsDb.Hsapiens.v86* package v2.99.0 (`https://bioconductor.org/packages/release/data/annotation/html/EnsDb.Hsapiens.v86.html`). We computed transcription start site enrichment, nucleosome signal, and the fraction of reads in peaks per cell with SIGNAC. We used AMULET [300]

from the GitHub *main* branch (`https://github.com/UcarLab/AMULET`) at commit *9ce413f* for detecting and removing doublet cells from the dataset.

We conducted QC on the cells in the dataset and retained only those that met all following criteria: over 1,000 fragments in peak regions, less than 100,000 fragments in peak regions, transcription start-site enrichment score greater than 2.7, transcription start-site enrichment score smaller than 10, over 30 % reads in peaks, *blacklist ratio* smaller than 0.66 and a nucleosome signal ratio below 10. This removed 7 % of cells, leaving us with 20,616 remaining cells. We then performed *TF-IDF normalisation* [301], identified top features (min.cutoff = 'q0'), conducted singular value decomposition, computed a kNN graph, calculated a UMAP representation [208], and finally clustered the data [302]. We computed gene activities and log-normalised ChromVar [303] activities using the *BSgenome.Hsapiens.NCBI.GRCh38* genome (`https://bioconductor.org/packages/release/data/annotation/html/BSgenome.Hsapiens.NCBI.GRCh38.html`) and motif position frequency matrices from the JASPAR2020 database (`https://bioconductor.org/packages/release/data/annotation/html/JASPAR2020.html`) [304].

2.3.19. Integrating scRNA-seq and scATAC-seq data

We carried out the integration described in this section individually for the GC-exposed and vehicle data (scRNA-seq and scATAC-seq) from 90 days-old organoids (*Veh-Veh* and *Chr-Veh* conditions) using the Python package GLUE [305] v0.3.2.

We saved the scATAC-seq data as an h5ad object by exporting it to Python using anndata2ri v1.1 (`https://github.com/theislab/anndata2ri`) automatic conversion. We then subset the data for the respective treatment condition and reduced it to 101 dimensions using 15 latent semantic indexing iterations, as implemented in GLUE. We discarded the first dimension as it usually correlates strongly with read depth. Using cosine similarity as a metric, we used the resulting representation to compute a kNN graph as implemented in the scanpy [35] package. This was followed by UMAP [208] computation also using the scanpy implementation.

Next, we processed the raw scRNA-seq count data by subsetting it to the respective treatment condition and using the scanpy package with default parameters unless stated otherwise for the following processing steps: HVG computation (n_top_genes = 2000, flavor = "seurat_v3"), count normalisation per cell to the median total counts per cell in the dataset, $\log(1 + x)$-transformation, scaling each feature to unit variance and zero mean, computation of 100

principle components, kNN graph computation using the cosine similarity metric, and UMAP computation.

We then computed a GLUE RNA-anchored guidance graph using the scRNA-seq and scATAC-seq data. We fitted a GLUE model using a negative binomial probability distribution and highly variable features from both data modalities. We used principal components as a reduced representation of the scRNA-seq data together with the latent semantic indexing embedding of the scATAC-seq data. We passed the data from both modalities through the trained GLUE model and used the resulting concatenated representation to compute a combined kNN graph and UMAP representation of the data. We used a bipartite matching approach [306], as implemented in the SCIM package (`https://github.com/ratschlab/scim`, *master* branch, commit *6392e65*), to match cells from both modalities one by one into *metacells*. If no scATAC-seq match was found for a scRNA-seq observation, we used only the scRNA-seq information. We calculated the GLUE latent vector of the cell as the average latent vector of the matched cells and used it for joint kNN graph and UMAP computation for data visualisation. Finally, we used the PYTHON implementation of MAGIC [307] (`https://github.com/KrishnaswamyLab/MAGIC`) to impute gene activities on the matched dataset using k = 15 neighbours, decay = 1, thresh = 1e-4, and four nearest neighbours for kernel bandwidth computation.

2.3.20. Gene-regulatory network inference

Inferring GRNs has been a longstanding challenge in systems biology. However, the emergence of single-cell genomics technologies has propelled the development of various GRN inference methods in recent years. These methods typically operate on scRNA-seq data, scATAC-seq data, or multimodal data combining gene expression and chromatin accessibility information. This information is then combined with prior information cis-regulatory elements and TF motifs to infer the GRN (Fig. 2.3). Including chromatin accessibility data in the inference task generally improves performance as this data bears relevant information on, for example, the accessibility of TF binding sites. In cases where only unpaired scRNA-seq and scATAC-seq data are available, their prior integration with tools like GLUE [305] enables the use of GRN inference tools relying on multimodal data input. Known limitations of GRN inference approaches from single-cell genomics data are, for example, the sparsity of the data and the lack of complete TF motif information. Please refer to [308] for an in-depth review of GRN inference in the era of single-cell genomics.

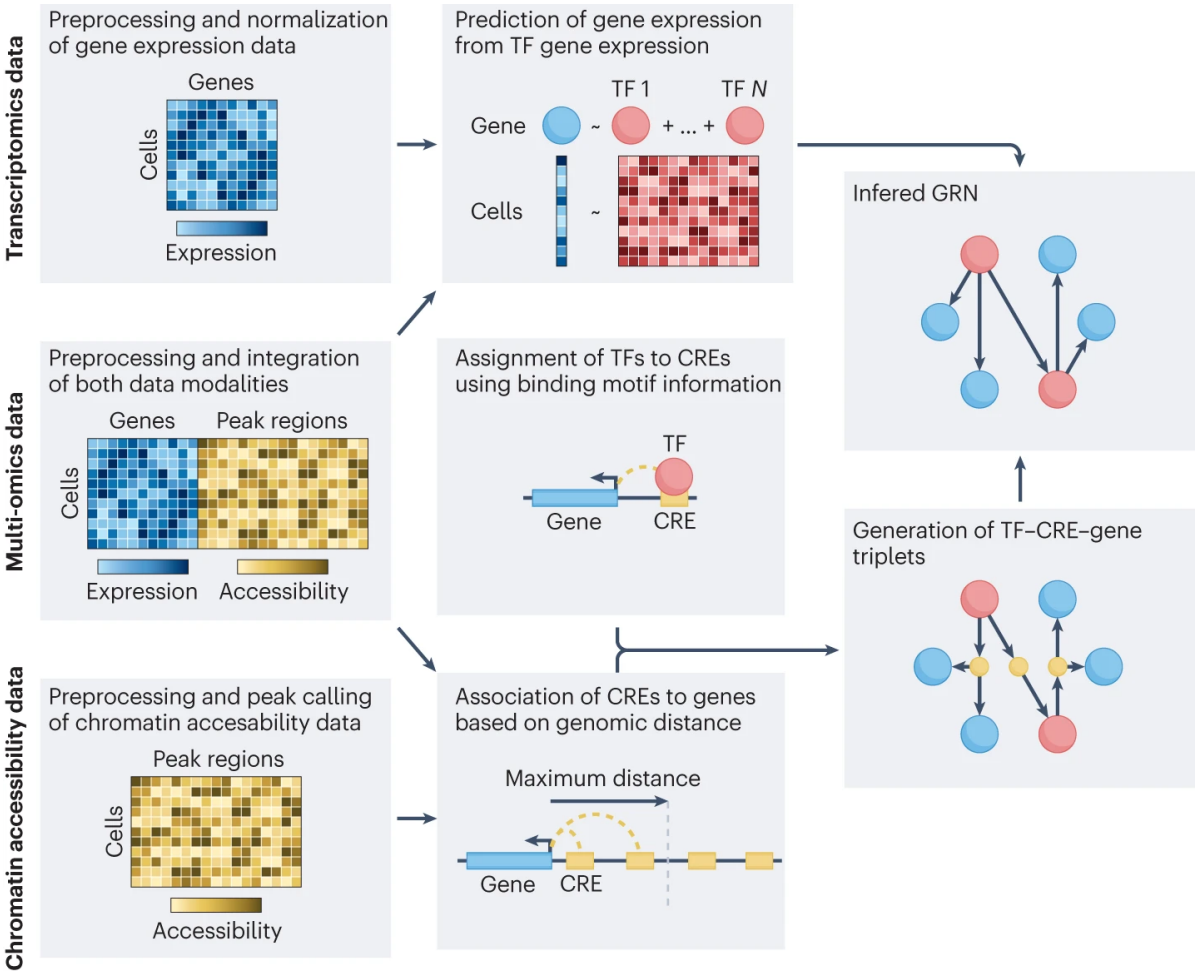

Figure 2.3.: **Common multimodal GRN inference pipeline.** Preprocessed scRNA-seq data provides gene expression information. In case of this being the only data available, the GRN is constructed by fitting a model which predicts gene expression from TF expression levels alone. Preprocessed scATAC-seq data provides chromatin accessibility information for cis-regulatory elements (CREs) that can be associated with genes by, for example, simple distance measures. TF motif databases associate TFs with CREs. Merging these data sources creates TF–CRE–gene triplets which can be aggregated into a GRN. In cases of multimodal data being available, gene expression information can further improve the prediction by informing the TF–CRE–gene triplet association. Reproduced from [308] with permission from Springer Nature.

The R tool Pando [254] has been developed for GRN inference in neural organoids and leverages multimodal data and prior information, including TF motifs, known cis-regulatory elements and evolutionary conservation. We used Pando v1.0.3 (`https://github.com/quadbio/Pando`) with R v4.1.2, in conjunction with Signac v1.9.0 and Seurat v4.3.0 (for preprocessing), to infer GRNs from the integrated multimodal data separately for the two treatment conditions (as in the integration step). First, we loaded the integrated metadata into a Seurat object and individually preprocessed the data of the two modalities. Next, we embedded the scATAC-seq peaks in low-dimensional space using *TF-IDF normalisation* [301], top feature identification (min.cutoff = 'q0'), and singular value decomposition. We log-normalised the scRNA-seq data (normalisation.method = "LogNormalize", scale.factor = 10000), identified top features (selection.method = "vst", features = 4000), scaled the data, and computed a PCA representation. We initiated the GRN using both data modalities and conserved regions from mammals as included in Pando (*phastConsElements20Mammals.UCSC.hg38*). Finally, we scanned candidate regions for TF binding sites, as provided by Pando (*motif2tf data*). We used the resulting data and initialised network to infer the GRN (peak_to_gene_method = 'Signac', method = 'glm') followed by gene module identification (p_thresh = 0.1, nvar_thresh = 2, min_genes _per_module = 1, rsq_thresh = 0.05). We used ggplot2 [309] v3.4.2 and ggraph (`https://github.com/thomasp85/ggraph`) v2.1.0 to generate TF-centered GRN visualisations.

As we conducted multimodal analyses exclusively in Line 409b2, we applied a stricter FDR cutoff of 5 % to define a DEG for all DE analyses presented in Section 3.3.6 of this thesis.

2.3.21. Statistical testing

We used the SciPy [296] implementation of the t-test for the means of two independent samples to test for significance throughout Section 3.3, unless stated otherwise. We computed correlation coefficients using the SciPy implementation of the Pearson correlation coefficient and p-value for testing non-correlation throughout Section 3.3, unless stated otherwise.

3. Results

In this chapter, I report the results of this thesis in three main sections. In Section 3.1, I describe the sfaira data and model zoo and outline how it can aid scRNA-seq data curation and analysis. sfaira greatly facilitated the data curation of the HNOCA. In Section 3.2, I detail the construction of HNOCA and the insights it provides into the current landscape of neural organoid scRNA-seq data. Using HNOCA as a neural organoid reference dataset helped contextualise the new scRNA-seq data generated in Section 3.3. This section describes a neural organoid treatment paradigm with GCs that unveils some of the effects of excess GC exposure on the developing brain.

My contributions to this chapter's work are detailed at the beginning of each main section.

3.1. Sfaira accelerates data and model reuse in single-cell genomics

This result section corresponds to, and is in part identical to, the one presented in the following publication:

Fischer, D. S.*, **Dony, L.***, König, M., Moeed, A., Zappia, L., Heumos, L., Tritschler, S., Holmberg, O., Aliee, H. & Theis, F. J. Sfaira accelerates data and model reuse in single cell genomics. *Genome Biology* **22**, 248. doi:10.1186/s13059-021-02452-6 (2021)

"*" denotes an equal contribution.

The sfaira package is available on GitHub: `https://github.com/theislab/sfaira`.

With respect to the aims of this thesis, this section addresses Challenge 1, as outlined in Section 1.4: Single-cell genomics technologies have enabled many scientific breakthroughs over the past years, and the number of publicly available scRNA-seq datasets is growing rapidly. Public data can provide a rich resource to augment and contextualise new studies

through integrated atlases. However, data curation and reuse have remained challenging and time-consuming due to a lack of metadata standardisation and consistent data formats in the field.

My contributions to the results presented here are as follows: Together with David Fischer, I wrote the package code and carried out the analysis for the publication. I also led the data curation and model training with support from David Fischer. With input from the other authors, I designed the organoid-specific metadata scheme.

SFAIRA is a comprehensive repository for streamlined scRNA-seq datasets and pre-trained end-to-end analysis models (Fig. 3.1). By integrating a model zoo with a data repository containing datasets from numerous sources with streamlined metadata, we facilitate access to datasets and pre-trained models (Section 3.1.1). The datasets provided use established ontologies for most available metadata annotations and can be easily queried through summary statistics to contextualise new observations (Section 3.1.2).

Through its unified interface, SFAIRA allows for the automated training of cell type classification and embedding models across a wide range of tissues and species, drastically reducing the time and work required for a first embedding and cell type annotation of new datasets (Section 3.1.3). It introduces a novel method to handle varying resolutions of cell type annotations during model training and a user-friendly interface, facilitating the curation of additional datasets (Section 3.1.5). SFAIRA is designed to be model-agnostic, serving as a versatile platform for deploying and accessing models. This openness and our structured model versioning system encourage the exchange of models between developers and users, facilitating model sharing (Section 3.1.4). Beyond practical benefits, SFAIRA addresses the need for model interpretability and generalisability. It uses a unified feature space that includes all genes, allowing for a comprehensive analysis of gene contributions to model outcomes without discarding low-variance features. This approach enables a deeper understanding of the latent space dimensions concerning gene activity (Section 3.1.8) and enables zero-shot inference on new datasets (Section 3.1.6). Moreover, SFAIRA showcases models trained on highly varied datasets without relying on covariates like organ type or experimental method (Section 3.1.7). Ultimately, SFAIRA aims to promote model reuse and extensive profiling across a broad spectrum of standardised datasets, enhancing the efficiency and scope of research in the field.

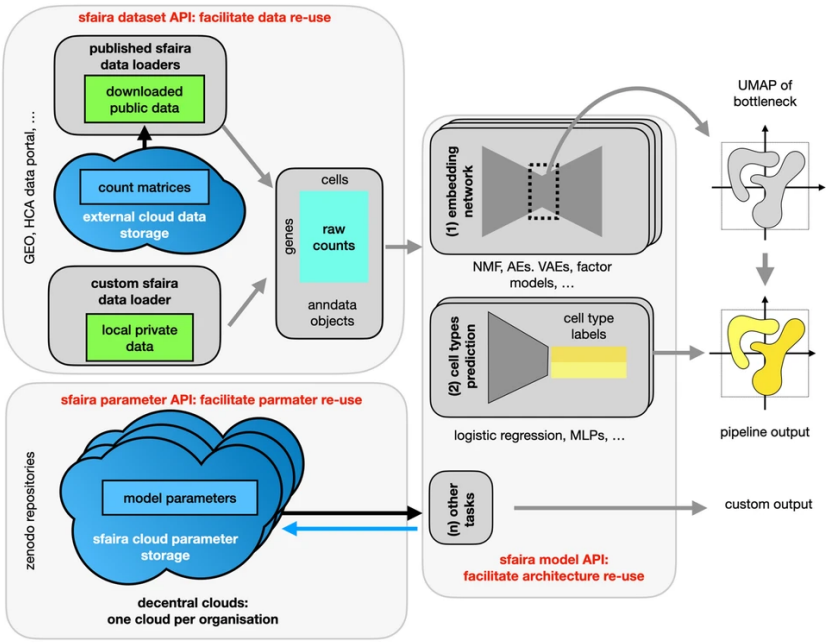

Figure 3.1.: **sfaira provides access to streamlined scRNA-seq datasets and automates key steps in exploratory data analysis.** The SFAIRA dataset and parameter API. The SFAIRA dataset API provides streamlined access to scRNA-seq data scattered across different public repositories. The SFAIRA model API can process these datasets to generate automated embedding and cell type assignment using pre-trained models stored on cloud servers via the SFAIRA parameter API. Adapted from [258] (Licensed under the Creative Commons Attribution 4.0 International license: `https://creativecommons.org/licenses/by/4.0/`).

3.1.1. A unified framework for accessing datasets, models, annotations, and model parameters

At the time of publication, SFAIRA encompassed 41 studies with 220 datasets and 8.0 million cells (Fig. 3.2) and has since been extended (see Section 4.1). Our data zoo supports advanced data queries based on metadata like organism, tissue type, and experimental protocol, incorporating standardised cell ontology labels [66] for consistent cell type annotation across datasets. Additionally, SFAIRA integrates ontologies for diseases [310], anatomy [68], cell lines [311], experimental methods [67], and more to enhance compatibility between queried datasets. Furthermore, the comprehensive metadata annotation fields enable the differentiation of multiple tissue sources, including primary tissue, tumour tissue, two-dimensional cell culture, and three-dimensional organoid models. SFAIRA provides specific annotation fields for the efficient management of organoid data, such as for the source cell line, the type of differentiation protocol and the exact protocol reference. SFAIRA supports advanced data subsetting queries across annotation hierarchies via ontologies. The feature space of any queried dataset can be automatically mapped to a provided genome assembly and annotation for easy concatenation of different datasets.

The model zoo component provides a seamless interface and diverse model implementations, simplifying model usage without navigating technical details. Provided models include supervised linear and non-linear cell type classification models and both linear and non-linear neural-network-based embedding models. Users can easily access pre-trained models and their parameters, facilitating efficient model application within PYTHON workflows. We employ a global versioning system for models and parameters, ensuring reproducibility and ease of access.

3.1.2. Scalable and streamlined data access and management with the sfaira data zoo

Addressing the challenge of managing large, heterogeneous, on-disk data collections, SFAIRA's data zoo utilises data-set-specific loader classes for efficient data interaction. These classes support easy creation, maintenance, and usage, benefitting from the rich functionality provided by pre-implemented parent classes. The zoo also introduces lazy-loaded dataset representations for subsetting data before loading into memory. It further provides options for storing data in H5AD-based backed ANNDATA [259] objects or collections for use with distributed-computing frameworks like DASK (`https://dask.org/`).

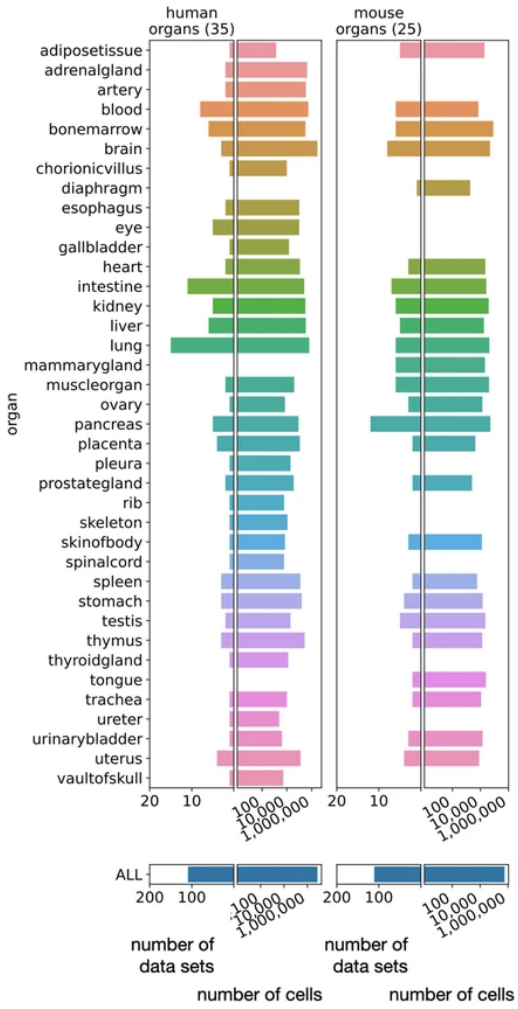

Figure 3.2.: **The sfaira data zoo encompasses a comprehensive collection of human and mouse scRNA-seq datasets.** Summary of the SFAIRA data zoo for mouse and human scRNA-seq samples at the time of publication, representing 220 datasets and 8.0 million cells. Adapted from [258] (Licensed under the Creative Commons Attribution 4.0 International license: `https://creativecommons.org/licenses/by/4.0/`).

sfaira provides streamlined access to large, cross-study datasets, allowing easy data statistics queries. One common query is gene-based, which can help put observations in context. For example, having a reference range for gene expression levels observed with scRNA-seq can aid interpretation. This can be done with sfaira via a simple query. An example is Ins1 expression across organs and cell types in mice, as shown in Fig. 3.3a. We find Ins1 expression in a range between 0 and 2500 normalised counts in mice, with all expression coming from pancreatic cell types. It is worth noting that summary metrics per cell type are often more helpful than dataset averages, which can be skewed towards frequent cell types or extrema, such as maximum expression, which can be heavily influenced by the variance of the expression distribution. We next analysed gene-gene dependencies, which can be used to establish regulatory relationships through correlation. We investigated the correlation of two cell-cycle-associated genes, Mcm5 and Pcna (Fig. 3.3b), to obtain a range for their correlation and estimate how often these genes' expression correlates across tissues.

Another query type is based on subsetting operations across cells using cell and dataset metadata, enabled by the homogenous ontology-based metadata annotation across datasets in sfaira. We also implemented relational reasoning of hierarchical metadata items based on ontologies, which is critical to obtaining meaningful subsets. We showcase a few exemplary data zoo summary statistics using metadata-based queries in Fig. 3.3c-g. Complexity plots of the total number of cells versus the number of most fine-grained cell type labels per organ generated with sfaira provide a guideline for the prioritisation of organs for further cell type discovery (Fig. 3.3c). Focusing on cell types, we queried the fraction of T cells across organs (Fig. 3.3d). Such an analysis can help characterise specific cell types across organs and datasets. Lastly, we compiled a summary of total reads per cell summary statistics and protocol summary statistics (Fig. 3.3e-g).

3.1.3. Automated single-cell data analysis with sfaira

End-to-end parametric approaches can eliminate the need for feature engineering. This has been a significant breakthrough in image-based deep learning [312]. In single-cell analyses, feature engineering refers to the initial analysis steps, such as normalisation, log-transformation, gene filtering, selecting components from PCA, and batch correction [39, 313] (Fig. 3.4a). These steps are necessary to obtain meaningful embeddings and clusterings. However, they can be a bottleneck in analysis workflows.

Pretrained embedding models can generate latent spaces amenable to downstream tasks

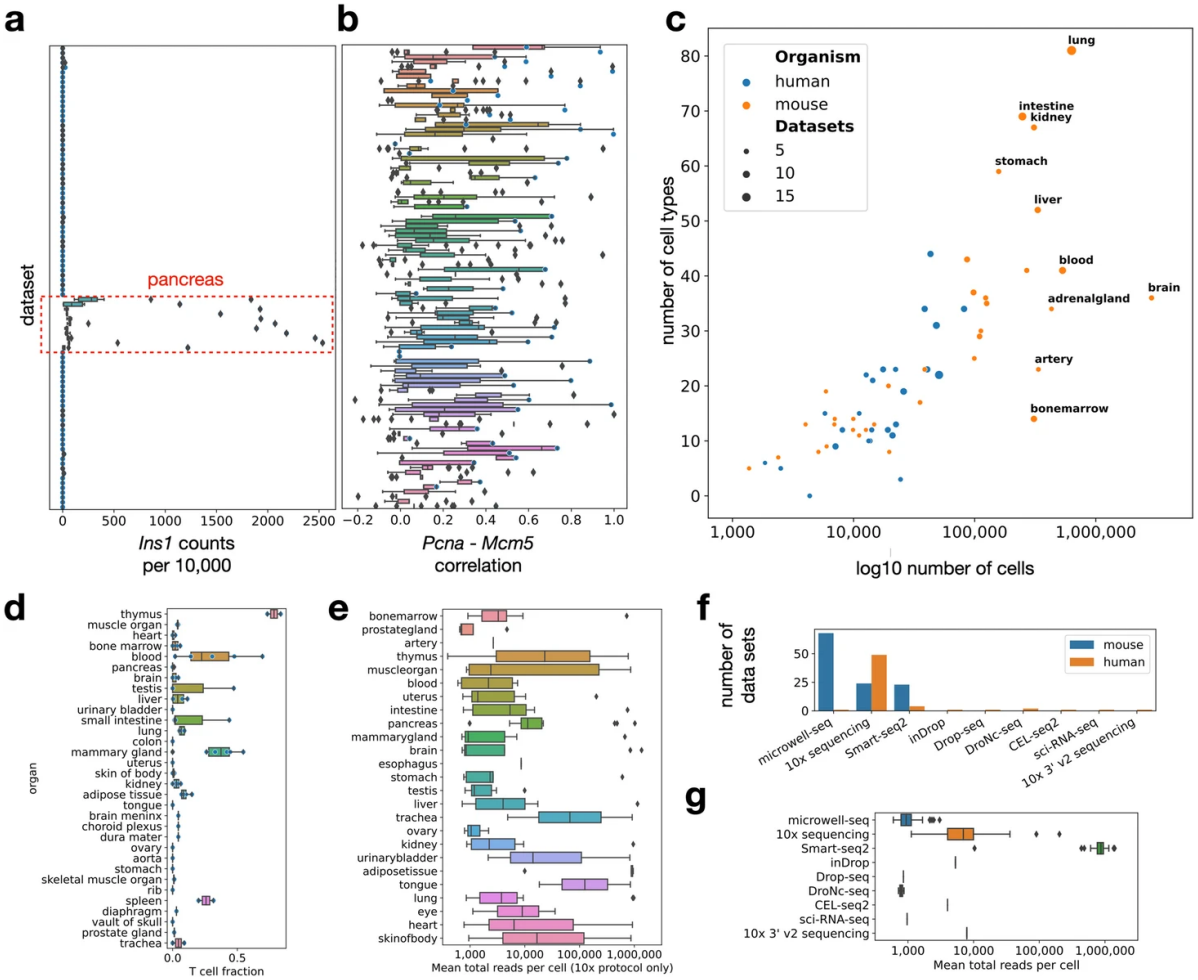

Figure 3.3.: **Contextualising data statistics with the sfaira data zoo.** **a.** Comparing gene expression patterns across organs. Mean normalised expression of Ins1 by organ and cell type across mouse datasets shows pancreas-specific expression. **b.** Distribution of Pearson correlation coefficients between the two cell-cycle-associated genes Pcna and Mcm5 across mouse datasets by organ and cell type. **c.** The number of cells in the SFAIRA data zoo per organism and organ versus the number of cell types reveals a broad range of dataset complexity. **d.** Querying cell type fractions in tissues across organisms with SFAIRA. This shows the fraction of T cells per mouse dataset, grouped by organ. **e.** Quering technical statistics of datasets with SFAIRA: mean total counts per cell in mouse and human organs for 10x Chromium datasets. **f.** Number of datasets in the data zoo across experimental protocol at the time of publication. **g.** Mean total counts per dataset by protocol (using UMIs where available, otherwise reads). Reproduced from [258] (Licensed under the Creative Commons Attribution 4.0 International license: `https://creativecommons.org/licenses/by/4.0/`).

without prior feature engineering. For instance, we processed human PBMC data in a standard preprocessing workflow [35, 39] and compared the result to a UMAP [208] of a linear model embedding. Both approaches separated annotated cell types into distinct clusters, demonstrating that both captured the biological heterogeneity of the system (Fig. 3.4b). We performed four additional zero-shot analysis examples on unseen data (Supplementary Fig. B.1) [248–251]. Quantifying the reconstruction error of the respective embedding model allows the comparison of different embeddings. In this reconstruction task, the linear model achieved a mean negative log-likelihood of 0.16.

We utilised SFAIRA's automated cell type annotation feature to label cells and investigate whether we could facilitate data interpretation with a preliminary proposal of cell type labels. We used a multi-layer perceptron model trained on different datasets to predict cell types, obtaining comparable results to the labels from the curated annotation (Fig. 3.4b). It is worth noting that with the addition of more data and improved classifier models trained on them, these initial annotations can become increasingly fine-grained. This example demonstrates that pre-trained embedding and cell type classification models can perform an automated initial analysis of scRNA-seq data, which can be further refined as required. In the next section, we elaborate on the pre-training details of such cell type classifiers and embedding models that allow these workflows to be executed on a large scale.

3.1.4. Facilitating model distribution for reproducible on-premise analyses

SFAIRA offers two types of model classes: gene expression reconstruction models (embedding models) and supervised classification models for cell type labelling (Fig. 3.1). These models are defined by their input and output spaces and architecture hyperparameters. SFAIRA also allows the integration of other model classes for additional purposes. Input feature space standardisation is made easy by coupling input gene sets to genome assemblies and functional annotation of gene sets. For instance, one can define an input feature space as the protein-coding genes in GRCh38 version 102 (Fig. 3.1). The label space of cell type classifiers is a set of cell types in the cell ontology [66], making hierarchical labels defined in the ontology available to the cost function of the model.

The implemented models cover a range of popular approaches: matrix factorisations, AEs, and VAEs for reconstruction models, as well as logistic regression models and multi-layer perceptrons for cell type classification. Third-party organisations can maintain their own public and private repositories of model weights (model zoos) on servers or in local directories

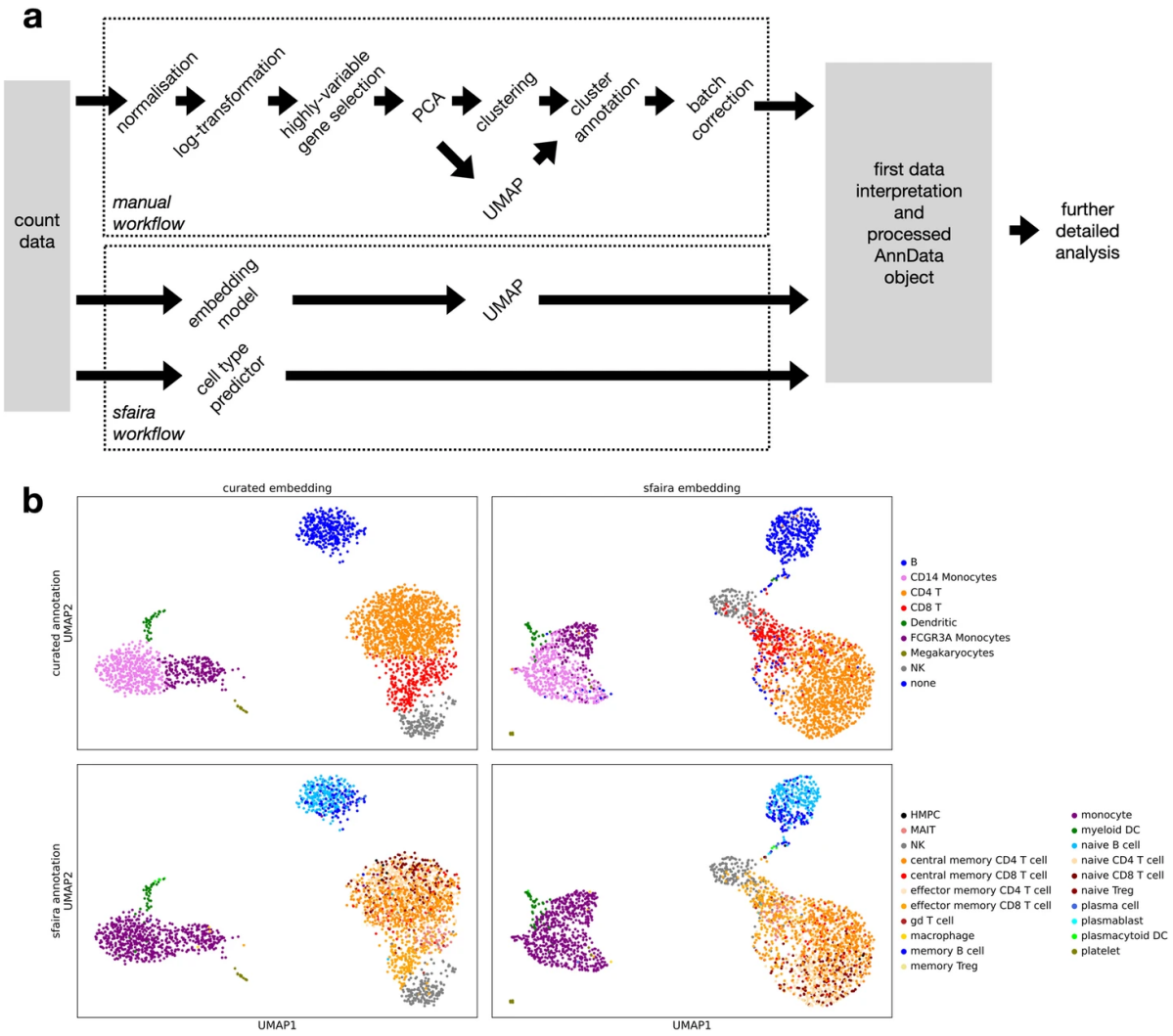

Figure 3.4.: **Sfaira drastically facilitates and partially automates exploratory analysis of scRNA-seq data.** **a.** Standard manual scRNA-seq data analysis pipeline and automated SFAIRA analysis pipeline. **b.** Comparison of analysis results from a manual feature engineering workflow with the results from the SFAIRA-automated embedding and cell type annotation on a human PBMC dataset. Shown are UMAPs representations from a PCA of a manually engineered feature space and a latent space of a linear SFAIRA embedding model. Colours show manual cell type labels and automatically predicted SFAIRA cell type labels. Reproduced from [258] (Licensed under the Creative Commons Attribution 4.0 International license: `https://creativecommons.org/licenses/by/4.0/`).

on-site, mitigating data privacy concerns. These parameter set versions are identified by the organisation that performed model training, the training data, and optimisation hyperparameters used to train the model. Versioning parameter sets by the organisation, input data and hyperparameters allows for incremental updates as new data or improved estimates become available in an ongoing grid search across optimisation hyperparameters.

SFAIRA provides an infrastructure that enables end-users to switch easily between different model types from various providers, facilitating model distribution and access. This reduces the effort required to implement and compare models, thereby improving decisions on pre-trained model usage. Additionally, decentralised storage of model weights enables the model zoo to respond quickly to new developments in the community.

3.1.5. Leveraging the cell ontology hierarchy to manage diverging annotation granularity

New biological insights, for example, from atlassing studies, can lead to amendments to existing cell type labels. This poses a significant challenge in deploying predictive models for cell type labels based on scRNA-seq data [314]. To address this issue, we link models to specific ontology versions and allow manual ontology extensions by the user to capture new developments as they emerge. Another challenge is that cell types in published studies are often annotated at varying resolutions. One study might report *leukocytes* in a given tissue, while a different study might operate at a finer resolution, distinguishing between *T cells* and *macrophages*. To solve this problem without time-consuming reannotation of all data, we propose the usage of *aggregated cross-entropy*, an adaptation of the common cross-entropy loss and an accuracy metric that can assign observations to single labels or sets of labels during training and testing (Fig. 3.5a, Section 2.1.5).

Previous research suggests that linear models can predict cell types accurately [315]. We trained three classes of models: logistic regression models, multi-layer densely connected feedforward neural networks, and a novel marker gene-centric linear model. The latter operates in a learned marker gene space, where a sigmoid mapping first transforms each gene into a binary on-off state. This approach can facilitate model interpretation and enables integration of prior knowledge via parameter priors in the marker state embedding layer. In agreement with previous findings on selected organs [316] (Fig. 3.5b,c), models generally performed equally well, with a median accuracy of 0.64 in human and 0.93 in mouse samples. The data zoo enables streamlined training of the models. It allowed us to relate

classifier performance to class frequencies (Fig. 3.5d,e) and zoom into individual classes (Supplementary Fig. B.2).

3.1.6. Serving embedding models for transfer learning

Embedding models produce a low-dimensional data representation necessary for many downstream analyses. Some frequently used models for representation learning on scRNA-seq are PCA, nonnegative matrix factorisation [317, 318], AEs [41], and VAEs [40]. Embedding models have been successfully used in transfer learning [317, 318], where public data informs the learned representations. However, encoder-decoder-based workflows in unsupervised scRNA-seq data analysis usually involve refitting the model on each new dataset for two reasons. Firstly, it is challenging to identify applicable pre-trained models in the literature. Secondly, unsuccessful transfer training of pre-trained models may result in unresolved relevant variation in the data, such as new cell states. SFAIRA provides users with a structured zoo of embedding models. It offers an extensive data library for pre-training, reducing the probability of unseen components of variation relevant to the test task during training. In this study, we benchmark our models on an extensive data collection to demonstrate that these limitations do not apply to the provided models.

We defined whole-dataset-based test splits for data from across organs to see how well the models can handle unseen confounding effects. The human and mouse lung dataset embeddings exemplify that cell types are well separated (Fig. 3.6a,b). We used four classes of embedding models to compare reconstruction errors in cross-validation splits across 35 human and 25 mouse tissues. We found that linear models perform comparably to non-linear models, with a median top-performing negative log-likelihood of 0.13 for human and 0.50 for mouse samples (Fig. 3.6c,d). Linear models also perform well for human PBMC data, with a best-achieved negative log-likelihood of 0.08, which is similar to the reconstruction error found on the held-out PBMC data (Fig. 3.4b). These models outperform baseline random projection models, showing that scRNA-seq data can be reconstructed well by pre-trained linear models (Section 2.1.6, Supplementary Fig. B.3) .

SFAIRA improves embedding analyses in three ways. Firstly, we remove the need for grid searches or feature engineering by deploying pre-trained models with optimised hyperparameters. Secondly, we facilitate future analyses given that previously annotated architecture choices, such as bottleneck dimensions, can be applied to a new study, thereby adding value beyond pure representation capabilities. Finally, we facilitate the inherently more challenging

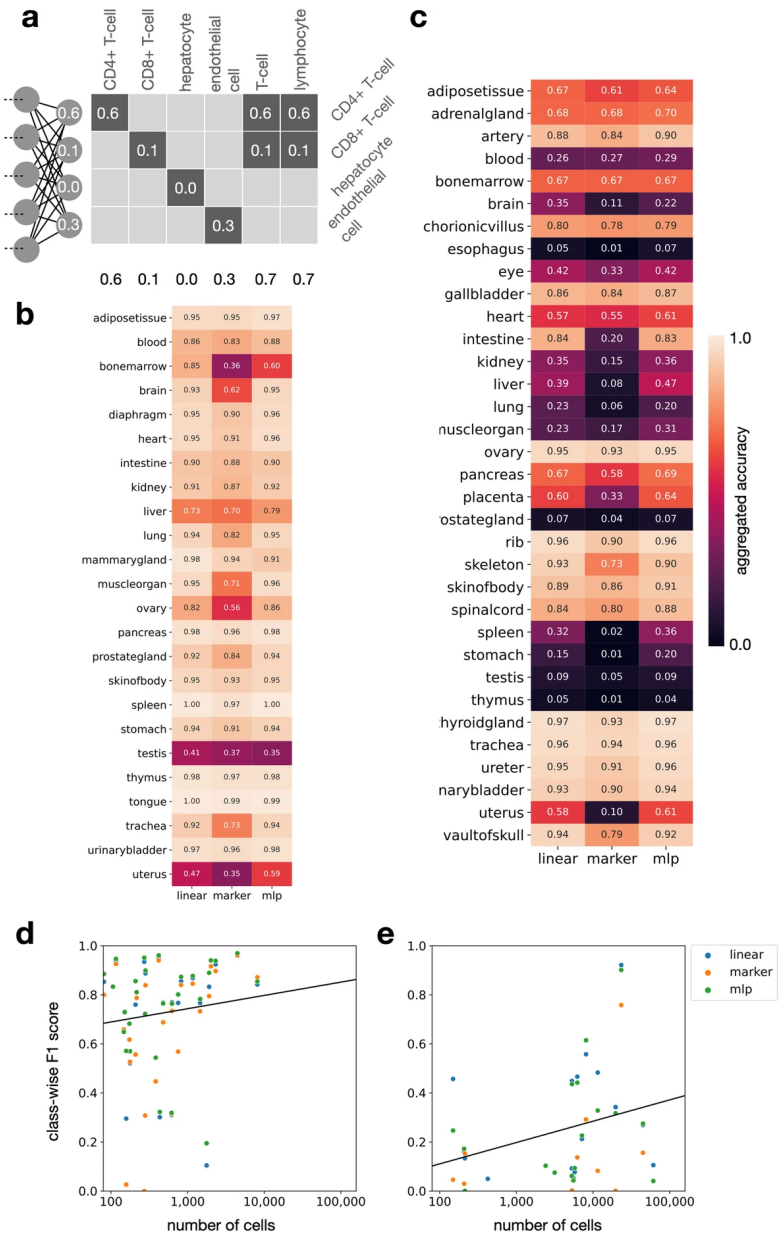

Figure 3.5.: **sfaira uses the cell ontology [66] to train models on datasets with diverging annotation resolution. a.** Example of loss aggregation for diverging cell type annotation granularity (see Section 2.1.5). The y-axis contains leaf nodes of the cell ontology, which can be combined linearly to yield the predicted probability mass of any coarser node in the ontology graph (x-axis). **b., c.** Cell type classification accuracy in mouse (b) and human (c) organs on fully held-out test-datasets. linear, linear classifier (logistic regression); marker; marker gene-based classifier; mlp, multilayer percetron. **d., e.** Class-wise F1 prediction scores by the number of cell types per class of cell type classifiers by model on lung data from mouse (d) and human (e) tissue show a correlation between the two variables. Reproduced from [258] (Licensed under the Creative Commons Attribution 4.0 International license: `https://creativecommons.org/licenses/by/4.0/`).

training of highly interpretable models by providing streamlined access to diverse datasets. The presented embedding models exemplify candidates compatible with a model zoo, but further pre-trained model classes could be used in the single-cell context [319].

3.1.7. Diverse training data induces model regularisation

Data integration is a delicate balance between removing technical effects resulting in inter-sample variance, and preserving biologically meaningful variance [39]. Instead of explicitly removing inter-sample variance, we focus on embedding models that can detect axes of biologically meaningful variation in an unseen dataset, a technique known as zero-shot learning [320]. One of the challenges in this approach is distinguishing between models that capture only the relevant axes of variation and models that simply capture all variation in the data, known as overfitting. Regularisation methods such as *L1* or *L2* parameter constraints, dropout mechanisms [41], and dimension bottlenecks in latent representations can help alleviate overfitting. While these methods can limit the degrees of freedom in over-parameterised models, they generally do not improve model interpretability.

An alternative approach is using a highly diverse training dataset while preserving the model capacity to encourage learning higher data abstractions instead of overfitting the entire training domain. This approach contrasts with using variance-explaining covariates in conditional embedding models, commonly used in data integration studies, to remove domain differences through a projection mechanism [42]. SFAIRA provides an extensive streamlined data zoo, enabling the successful training of embedding models on large scRNA-seq datasets spanning different studies and entire organisms (Fig. 3.6e,f). In summary, SFAIRA facilitates model regularisation through highly diverse training datasets to extend data reuse from projection-based data integration to pre-trained higher-abstraction embedding models.

3.1.8. Gradient maps for interpreting non-linear embedding models

Many scRNA-seq embedding models are based on PCA for its interpretability since it allows a straightforward interpretation of latent dimensions as linear combinations of input features. Similarly, gradient maps from bottleneck activations to input features enable local interpretations in non-linear embeddings of encoder-decoder networks. Gradient maps promise to offer a higher-level view of GRNs by correlating bottleneck dimensions to molecular pathways or similar complex regulatory elements. Our study found that

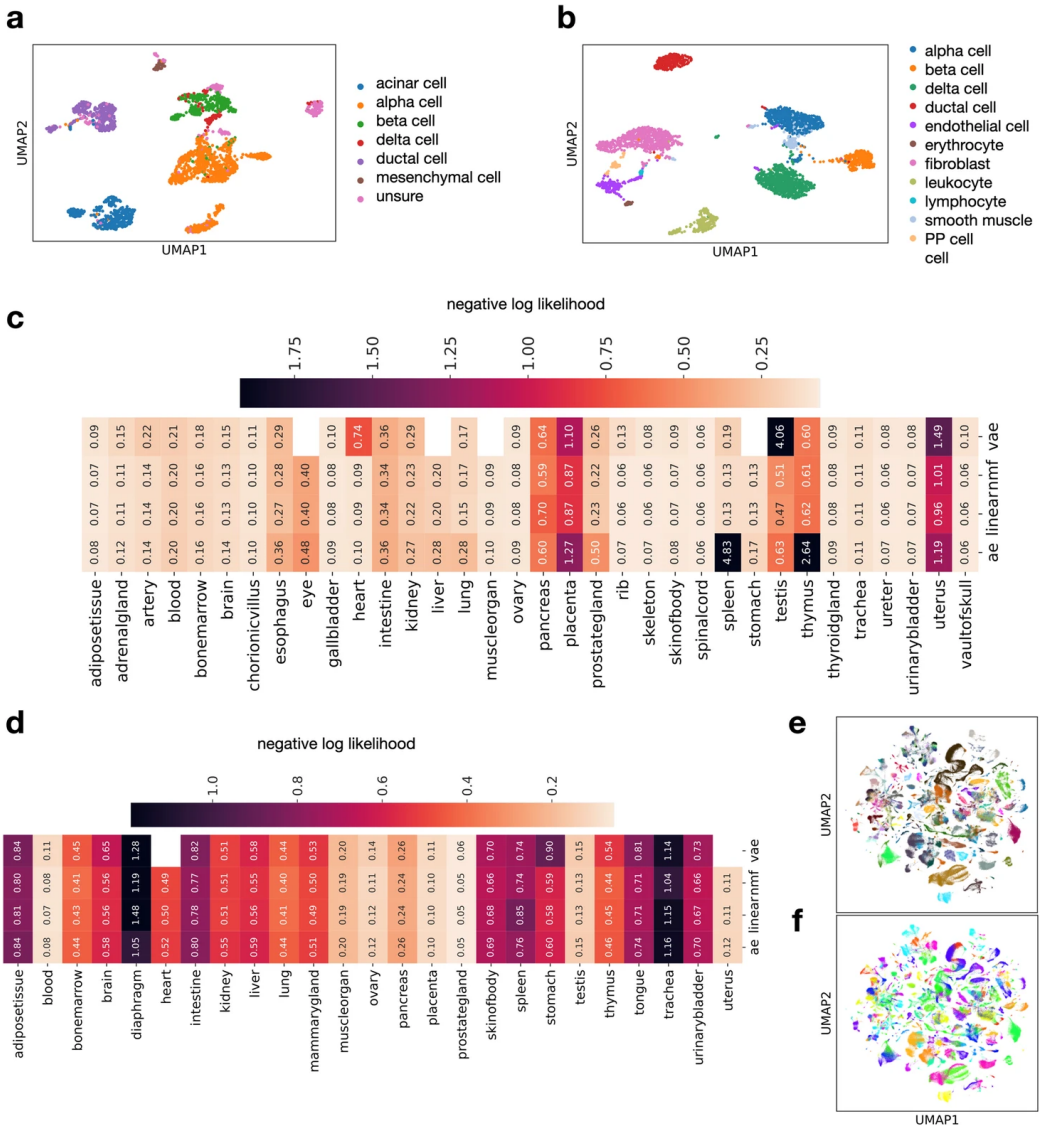

Figure 3.6.: **Training streamlined embedding models across tissues and on whole atlases with sfaira. a., b.** sfaira's pre-trained embedding models can meaningfully embed unseen data. UMAP of the latent space of the best-performing embedding model for pancreas data from human (a) and mouse (b). The superimposed colours correspond to the original, author-provided cell type labels. **c., d.** Reconstruction loss comparison across organs and organisms. We use the negative binomial likelihood as a metric on reconstructed held-out test data with different embedding models on human (c) and mouse (d) data. linear, a PCA-like linear model; nmf, non-negative matrix factorisation; ae, AE; vae, VAE. **e., f.** sfaira enables the training of embedding models on extensive datasets. UMAP of the latent space of an embedding model trained on all mouse data in the sfaira data zoo at the time of publication with the dataset (e) and cell type (f) labels superimposed. Reproduced from [258] (Licensed under the Creative Commons Attribution 4.0 International license: `https://creativecommons.org/licenses/by/4.0/`).

gradient maps of the embedding space to the model input grouped similar cell types within a hierarchical clustering of the gradient correlation matrix, recapitulating cellular ontology relationships in two sample datasets (Fig. 3.7a, Supplementary Fig. B.4a).

We also found that linear models and AEs are comparable in the number of features highlighted by these gradient-based mechanisms per cell type (Fig. 3.7a, Supplementary Fig. B.4a) and have a similarly shaped marginal distribution of normalised gradients (Fig. 3.7b, Supplementary Fig. B.4b). Additionally, models trained on smaller datasets tend to collapse to only using a fraction of the gene space and representing cells based on feature correlations in this space. Growing datasets require learning more complex representations, and any such model collapse can be readily identified with gradient-based approaches.

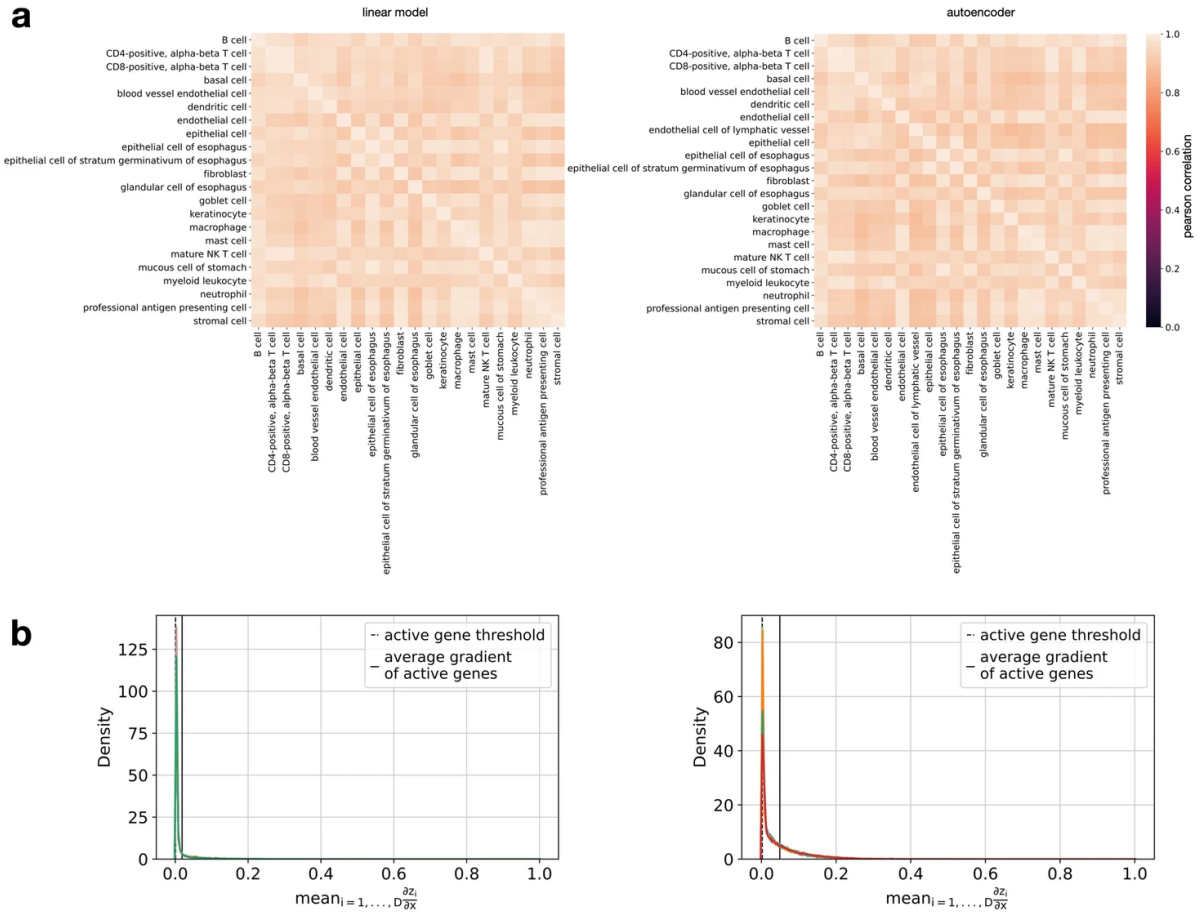

Figure 3.7.: **Gradient-based model interpretation and data regularisation of non-conditional embedding models: linear and AE embedding models for human oesophagus.** **a.** Correlation of cell type-wise aggregated gradients of the latent space with respect to the input features. **b.** Distribution of feature-wise aggregated gradients of the latent space with respect to the input features. Reproduced from [258] (Licensed under the Creative Commons Attribution 4.0 International license: `https://creativecommons.org/licenses/by/4.0/`).

3.2. An integrated transcriptomic cell atlas of human neural organoids

This result section corresponds to, and is in part identical to, the one presented in the following publication:

He, Z.*, **Dony, L.***, Fleck, J. S.*, Szałata, A., Li, K. X., Sliškovic, I., Lin, H. -C., Santel, M., Atamian, A., Quadrato, G., Sun, J., Pasca, S. P., Camp, J. G., Theis, F. & Treutlein, B. An integrated transcriptomic cell atlas of human neural organoids. *In revision.* Preprint [204] doi: 10.1101/2023.10.05.561097 (2023)

"*" denotes an equal contribution.

The HNOCA Supplementary Tables 1-6 referenced in this section are available online: `https://www.biorxiv.org/content/10.1101/2023.10.05.561097v1.supplementary-material`. The code used in the analysis of the results presented in this section is available on GitHub: `https://github.com/theislab/neural_organoid_atlas`.

With respect to the aims of this thesis, this section addresses Challenge 2, as outlined in Section 1.4: The curation of large-scale transcriptomic cell atlases can uncover the mechanisms underlying development and disease. The human brain is one of the most complex organs in the human body and is critical to our overall health and well-being. Neural organoids have emerged as powerful *in vitro* models of human neurodevelopment. Combining a large number of existing neural organoid scRNA-seq datasets into a comprehensive single-cell atlas would uncover underrepresented sections of the developing human brain in organoid models. It would further provide insight into systematic transcriptomic differences between organoids and primary fetal brain and pave the way towards more faithful organoid model systems. It would also enable the rapid and effective contextualisation of newly generated neural organoid datasets.

My contributions to the results presented here are as follows: Together with Zhisong He and Jonas S. Fleck, I collected and retrieved the scRNA-seq data included in HNOCA with suggestions from Sergiu P. Pasca, J. Grayson Camp and Barbara Treutlein. Together with Zhisong He, I performed the DE and transcriptomic comparison analysis with support from Irena Sliškovic and Katelyn X. Li. Together with Katelyn X. Li, Irena Sliškovic and Artur Szałata, I curated the HNOCA data and harmonised the associated metadata. I performed HNOCA data preprocessing and integration with support from Katelyn X. Li and Irena

Sliškovic. Together with Zhisong He, Jonas S. Fleck and Artur Szałata, I developed the HNOCA data processing pipeline. Together with Katelyn X. Li, I performed the benchmark of integration methods. I provided input for the cell type hierarchy curation by Zhisong He and Jonas S. Fleck. I coordinated the work of Artur Szałata, Katelyn X. Li, and Irena Sliškovic by defining tasks and providing input where needed.

This study harmonises 36 scRNA-seq datasets from various human neural organoid protocols to create the HNOCA (Section 3.2.1). We establish an advanced computational pipeline that enables a thorough quantitative comparison of the organoid atlas to a recently published reference dataset of the developing human brain [46]. By combining the two resources, we reannotate neural and non-neural populations in organoids and estimate the capacity of different protocols to generate neural cells from various brain regions. We thereby identify primary cell populations under-represented in neural organoids (Section 3.2.2). We investigate systematic transcriptional deviations from primary tissue in neural organoids and identify a shared signature of cell stress specific to organoid-derived neurons (Section 3.2.3). Lastly, we highlight novel cell states in a recently published neural organoid morphogen screen [321], using the HNOCA as a reference and representation of the current neural organoid cell landscape (Section 3.2.4). Altogether, this study provides a rich atlas resource for assessing the fidelity of neural organoid cell states, contextualising perturbed and diseased cellular states, and guiding future protocol development.

3.2.1. Building the Human Neural Organoid Cell Atlas

We built a transcriptomic human neural organoid atlas by collecting scRNA-seq data and harmonising detailed technical and biological metadata from 33 published [96, 99, 101, 103, 104, 116–125, 127, 133, 142, 193, 252–257] and three unpublished datasets (HNOCA Supplementary Table 1). We applied consistent preprocessing and QC to create the 1.77 million cell HNOCA from these datasets (Fig. 3.8a). The HNOCA includes cell types and states generated with 26 distinct neural organoid differentiation protocols (three unguided and 23 guided), covering a time range from seven to 450 days (Fig. 3.8b). We implemented a three-step integration pipeline to remove batch effects:

1. We anchored our analysis in primary human fetal brain data by mapping the HNOCA to a single-cell transcriptomic reference dataset of the developing human brain [46] using the RSS [127] method.

2. To enable a semi-supervised (label-aware) integration approach, we created an initial marker-based hierarchical cell type annotation using our newly developed annotation algorithm SNAPSEED (see Section 2.2.3).

3. We used SCPOLI [82] for label-aware integration based on the SNAPSEED annotations created above.

Evaluating a range of unsupervised and semi-supervised data integration approaches using the SCIB pipeline [78], we identified SCPOLI as the top-performing method for our data (Supplementary Fig. B.5). Following reclustering on the integrated representation, we refined the cell type annotations based on the expression of canonical markers, organoid sample age, as well as the auto-generated cell type labels. The integrated embedding allowed us to identify three neuronal differentiation trajectories in UMAP [208] space linked to the dorsal telencephalon, ventral telencephalon and non-telencephalic regions. We also identified glial trajectories, leading from progenitor cells to astrocytes and OPCs (Fig. 3.8c-e; Supplementary Fig. B.6). Cells from both unguided and guided protocols covered all differentiation trajectories (Fig. 3.8f).

We computed a real-time-informed pseudo time for the dorsal telencephalic neuronal trajectory in the HNOCA using neural optimal transport [264] to understand the relevant dynamics and transitions of cell states and types (Fig. 3.8h). This allowed us to observe consistent pseudo-temporal expression profiles of genes such as the NPC marker SOX2, as well as neuronal markers BCL11B (CTIP2) for deeper-layer neurons, and SATB2 for upper-layer neurons (Fig. 3.8i). Sub-clustering of the non-telencephalic neurons to better resolve their profound heterogeneity exposed numerous neuronal populations expressing distinct marker genes (Fig. 3.8j-k).

3.2.2. Mapping the Human Neural Organoid Cell Atlas to a primary human fetal brain reference

A comprehensive reference atlas of human neural cell types and states is necessary to refine our annotations further and, in particular, accurately annotate the diverse non-telencephalic neuronal populations. Matching the early developmental timepoints found in organoids, we used a recently published scRNA-seq reference dataset of first-trimester developing human brain [46] (Fig. 3.9a) as a reference for comparison with the HNOCA. Using SCARCHES [42], we projected the HNOCA onto the SCANVI-integrated [81] primary fetal reference data. By

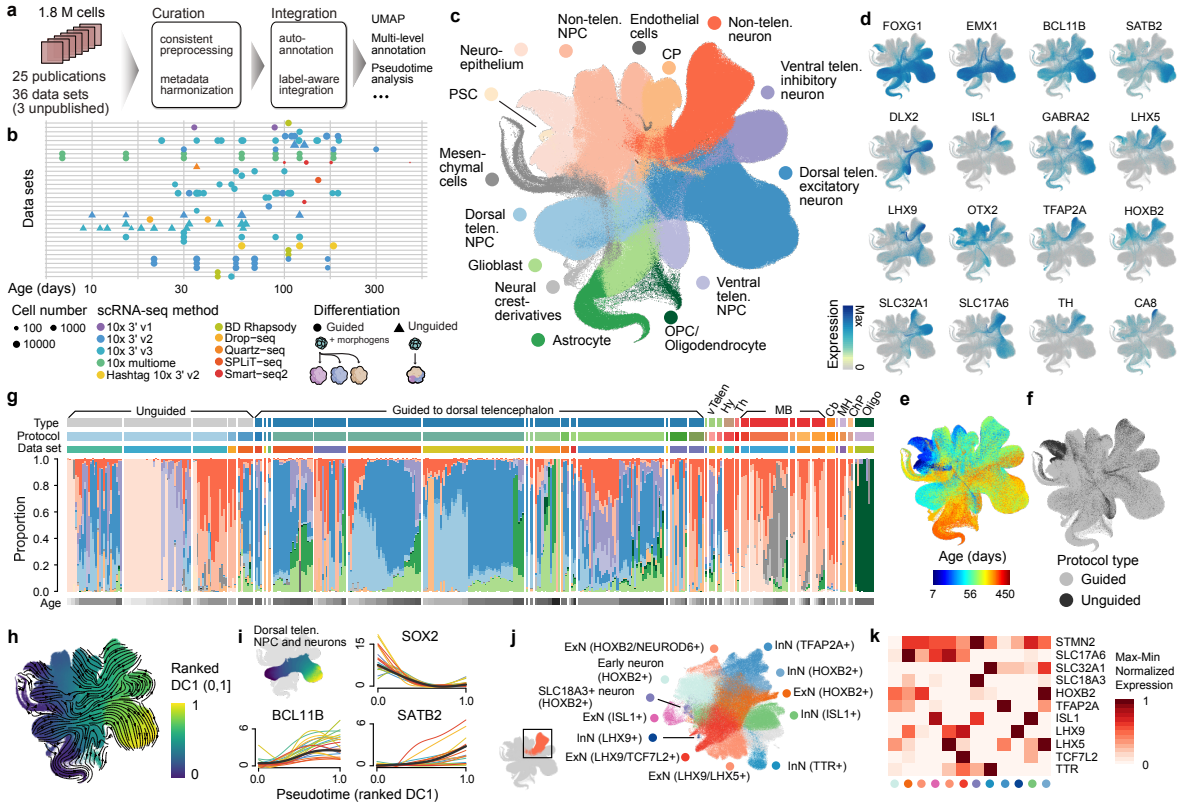

Figure 3.8.: **HNOCA overview. a.** The HNOCA construction pipeline. **b.** HNOCA sample metadata. **c., d., e., f.** UMAP of the integrated HNOCA, coloured by level-2 cell type annotations (PSC, pluripotent stem cell; NPC; CP, ChP; OPC, OPC; telen, telencephalon)(c), marker gene expression (d), sample age (e), and organoid differentiation protocol types (f). **g.** HNOCA cell type (level-2) proportions. Every stacked bar represents one biological replicates, grouped by dataset and ordered by increasing sample ages within groups. The bars on the top show the 36 datasets, the organoid differentiation protocols and protocol types (vTelen, ventral telencephalon; Hy, hypothalamus; Th, thalamus; MB, midbrain; Cb, cerebellum; MH, medulla; ChP, ChP; Oligo, oligodendrocyte). The bottom bars show the sample age. **h.** HNOCA UMAP coloured by top-ranked diffusion component (DC1) on the real-time-informed transition matrix between cells. The stream arrows visualise the inferred developmental trajectories. **i.** Expression profiles along the cortical pseudo time of SOX2 (RG), BCL11B (deeper layer cortical excitatory neurons) and SATB2 (upper layer cortical excitatory neurons). **j.** UMAP of non-telencephalic neurons, coloured and labelled by cellular subtypes. **k.** Heatmap showing the relative expression of selected genes across different non-telencephalic neuron clusters. The colours of the dots at the bottom represent cluster identities, as shown in (j). Reproduced from [204] (Licensed under the Creative Commons Attribution-NonCommercial 4.0 International license: `https://creativecommons.org/licenses/by-nc/4.0/`).

doing so, we could establish a shared latent space and create a bipartite weighted kNN graph between cells in the HNOCA and the primary reference atlas. Using the graph, we transferred the *CellClass* and *Subregion* labels, as well as the NTT information of neuroblasts and neurons, to the HNOCA. The transferred labels were consistent with our assigned labels (Supplementary Fig. B.7) and allowed us to refine the regional annotation of non-telencephalic NPCs and neurons in the HNOCA, as well as the NTT annotation of the non-telencephalic neurons (Fig. 3.9b). Altogether, these annotations comprise the final hierarchical cell type annotation of the HNOCA (Supplementary Fig. B.7).

We assessed the ability of the different neural organoid protocols to generate neural cells from various brain regions using the final regional annotation of the HNOCA (Fig. 3.9c; (Supplementary Fig. B.7, B.8); HNOCA Supplementary Table 2). The results showed that the unguided neural organoid protocols could generate cells from all brain regions with varying proportions across datasets, indicating high variability between batches and lines. In contrast, guided organoid protocols predominantly generated cells of the targeted brain region, highlighting morphogen guidance as an efficient tool to generate cells from specific brain regions. Additionally, datasets of organoids guided to a specific region often showed enrichment of cells from neighbouring brain regions on the neural axis. For instance, several midbrain organoid protocols also produced higher proportions of diencephalic and rhombencephalic neurons, indicating less precise morphogen guidance for these brain regions.

We evaluated the effectiveness of the different organoid protocols in covering the primary neural cell type space by computing an organoid presence score for every primary cell type (Section 2.2.11). We normalised the scores per organoid dataset (Supplementary Fig. B.9; HNOCA Supplementary Table 3) and obtained the maximum presence score for each primary reference cell type (Fig. 3.9d), quantifying how well it is represented in at least one HNOCA dataset. Our analysis confirmed the absence of erythrocytes, immune cells, and vascular endothelial cells in the HNOCA (Fig. 3.9e). These cell types are derived from non-ectodermal germ layers during development, while organoids are generally exclusively ectoderm-derived. We observed varying proportions of different brain regions represented across the neural cell populations of the organoid datasets, including RG, IPs, and neurons. More than half of the datasets were generated using protocols guiding towards the telencephalon. Accordingly, telencephalic cell types were best represented in the HNOCA, while cell types of the thalamus, midbrain, and cerebellum were least represented. This included thalamic reticular nucleus (TRN) GABAergic neurons, dorsal midbrain m1-derived GABAergic neurons and m1/m2-derived glutamatergic neurons, and cerebellar Purkinje cells (Fig. 3.9f,g). Importantly, even though these cell types are less abundant in the HNOCA organoid datasets than in

the primary reference, organoid protocols can generate these cell types when using the appropriate morphogens. For example, Purkinje cells can be generated by cerebellum and midbrain organoid protocols.

3.2.3. Comparing transcriptomes across organoid and fetal neuronal cells

Beyond cellular composition, we aimed to compare the gene expression patterns of primary brain tissue and organoids generated through various differentiation protocols. Using pseudobulk replicates, we identified DEGs for each neural cell type in the HNOCA by comparing them to their matching primary reference cell type [46] (Fig. 3.10a; HNOCA Supplementary Table 4). We found that for most neural cell types, more than a third (mean = 39.8 %, sd = 11.5 %) of DEGs were common across at least half of the protocols (*protocol-common DEGs*), indicating that a considerable fraction of the transcriptomic differences between organoid and primary cells was independent of the organoid differentiation protocol (Fig. 3.10b). We validated our findings using an additional primary human cortical scRNA-seq dataset [48] to ensure that the protocol-common DEGs were not an artefact of using only a single reference dataset. We found a substantial overlap with the previously identified protocol-common DEGs (Supplementary Fig. B.10; HNOCA Supplementary Table 5). We further identified a set of ubiquitous differentially expressed genes (uDEGs) that were differentially expressed in at least half the differentiation protocols across at least 14 out of the 16 regional neural cell types (Fig. 3.10c). These uDEGs showed a high level of consistency in log2FCs across neuron types and protocols, with a correlation above 0.8 for over half of the combinations, highlighting their collective behaviour (Fig. 3.10d). Out of all 994 uDEGs, 244 genes had a negative log2FC throughout, and 622 genes consistently had a positive log2FC, leaving only 128 genes (14 %) regulated in an inconsistent direction across subtypes or protocols (Fig. 3.10e).

We next aimed to identify the biological pathways associated with these three groups of uDEGs in our study. We performed an enrichment analysis of the Gene Ontology Biological Pathways [287, 288] set to achieve this. Our analysis revealed that the downregulated uDEGs were linked to neurodevelopmental processes such as *neuron cell-cell adhesion* and *synapse organisation*. On the other hand, the upregulated uDEGs were enriched in multiple metabolism-associated terms, such as *mitochondrial ATP synthesis coupled electron transport* and *canonical glycolysis* (Fig. 3.10f). We confirmed these enriched terms using our validation approach across multiple primary cortex reference datasets [46, 48]. We observed few to no enriched terms in the inconsistently regulated uDEGs (Supplementary Fig. B.10).

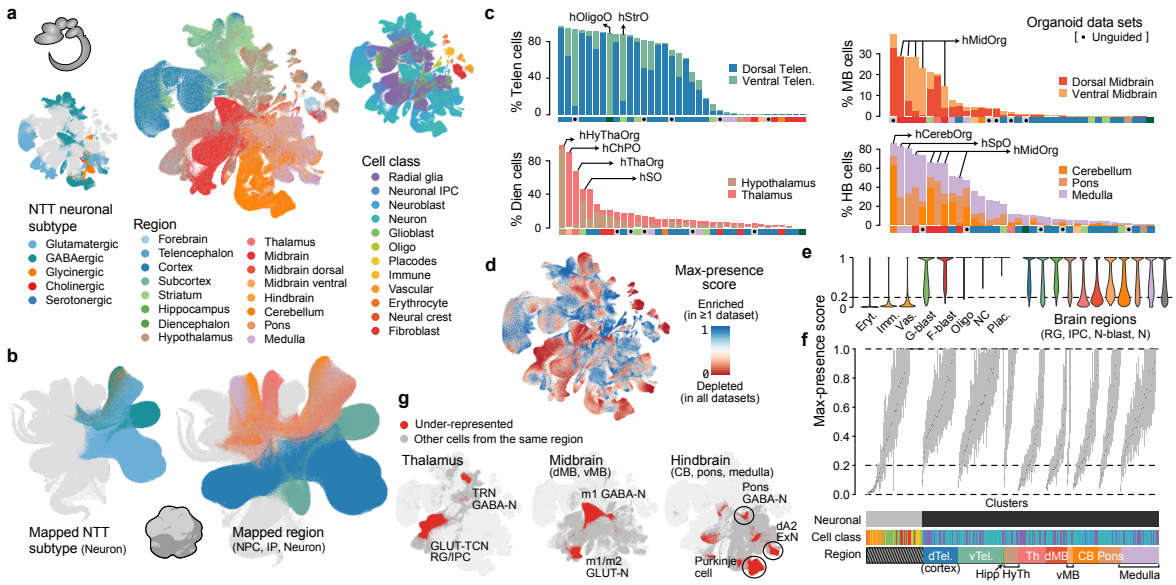

Figure 3.9.: **Refining cell type annotations and identifying underrepresented cell types by reference mapping to primary fetal brain tissue.** **a.** UMAP of the first-trimester human developing brain cell atlas [46], coloured by NTT subtypes (left), regional labels (middle), and annotated cell classes (right). **b.** UMAP of HNOCA, coloured by the mapped NTT subtypes of neurons (left) and mapped regional labels of NPCs, IPs, and neurons. **c.** Fraction of neural cells (NPCs, IPs, and neurons) from different brain regions, including telencephalon (dorsal and ventral), diencephalon (hypothalamus and thalamus), midbrain (dorsal and ventral), and hindbrain (cerebellum, pons and medulla), in different datasets. The datasets (x-axes) are ordered by the total fraction of the respective brain region in the dataset (bar height). Dots below the bars mark datasets based on unguided differentiation protocols, and selected guided protocols are labelled by name. The colours below each bar show the organoid protocol type (unguided and guided to different regions). **d.** UMAP of the first-trimester human developing brain cell atlas coloured by maximum HNOCA presence score. A high max presence score suggests enrichment of the corresponding primary cell state in at least one HNOCA dataset, and a low score suggests under-representation of the cell state in all HNOCA datasets. **e.** Distribution of max presence scores of different cell classes and brain regions in the first-trimester human developing brain reference atlas. **f.** Box plots showing the distribution of max presence scores across different primary reference cell clusters. The annotation below marks the cell class and the consensus region information of the primary reference cell clusters. Tel., telencephalon; Hippo, hippocampus; HyTh, hypothalamus; Th, thalamus; MB, midbrain; d, dorsal; v, ventral. **g.** UMAP of the first-trimester human developing brain atlas showing selected primary neural cell types and states in thalamus (left), midbrain (middle) and hindbrain (right) that are under-represented in HNOCA (in red). Reproduced from [204] (Licensed under the Creative Commons Attribution-NonCommercial 4.0 International license: `https://creativecommons.org/licenses/by-nc/4.0/`).

It has been previously reported that the limitations of current culture media and impeded nutrient supply can lead to metabolic changes and activation of energy-associated pathways in neural organoids [103, 104]. Upon scoring the two most significantly enriched gene sets in upregulated uDEGs across HNOCA and the primary reference atlas [46], we identified the *canonical glycolysis* term as more specific to organoid cells and as showing a better separation between organoid and primary cells (Fig. 3.10g). As such metabolic changes are considered a cell culture-related phenomenon [104], we used the *canonical glycolysis* score as a universal proxy for these stress effects in organoids (Supplementary Fig. B.11).

We wanted to investigate the specificity of increased glycolysis in organoids to different cell types. Investigating the *Kanton et al.* [127] and *Braun et al.* [46] datasets as representative examples of organoids and primary data, we found that glycolysis scores were distributed evenly across all neural cell types but with an overall increase in organoid cells (Supplementary Fig. B.11). We focused on dorsal telencephalic neurons – generated by most differentiation protocols – to compare the distribution of glycolysis scores across these protocols. We discovered that certain protocol features were correlated with metabolic cell stress, including the use of maturation media, slicing/cutting of organoids, and, to a lesser extent, shaking/spinning of organoids, all reducing metabolic stress (Fig. 3.10h). We noticed a negative correlation between the average glycolysis scores and transcriptomic similarities of organoid and primary reference cell types across differentiation protocols [103, 104]. Interestingly, the correlation was much lower for variable TFs than for the entire gene space, indicating that these changes in the metabolism of organoids only narrowly impact the core molecular identity of neuronal cell types defined by TF activity (Supplementary Fig. B.11).

To evaluate the transcriptomic similarity between organoid and primary reference cell types beyond metabolic effects, we analysed the expression of 366 variable TFs. We correlated TF expression between matching primary and organoid neuronal subtypes. Our results showed that guided and unguided protocols for organoid differentiation generated neuronal cell types of similar fidelity, while transcriptomic differences across different brain regions were more pronounced. For instance, organoid neurons from the dorsal part of several brain regions, such as the dorsal telencephalon, dorsal midbrain, and cerebellum, demonstrated higher similarity to their primary counterparts across organoid datasets than cell types derived from the ventral part of most brain regions, such as the hypothalamus, ventral midbrain, and pons (Fig. 3.10i).

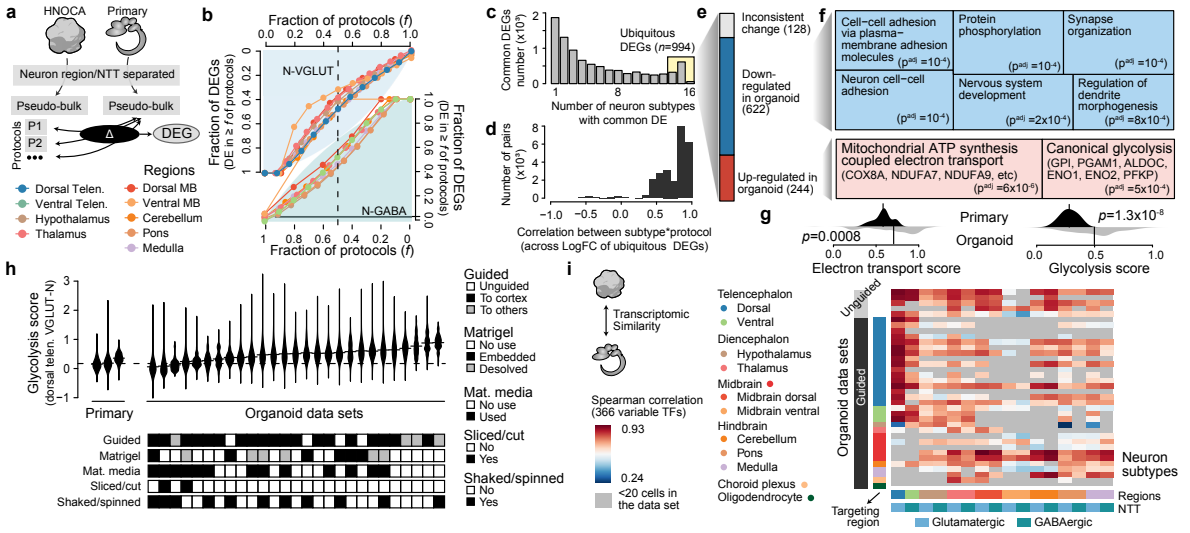

Figure 3.10.: **Metabolic cell stress is a universal transcriptomic differentiator between organoids and fetal brain.** **a.** DE analysis approach. **b.** Fraction of DEGs per subtype-specific protocol. The top-left half corresponds to glutamatergic neurons, while the bottom-right half corresponds to GABAergic neurons. The different colours indicate brain regions. The x-axis shows the fraction of protocols, denoted as f. The y-axis shows the fraction of DEGs in at least f of the respective subtype-specific protocols. The vertical dashed lines denote $f = 0.5$. **c.** Number of protocol-common DEGs, per number of neural cell types that share a DEG. If found in at least half of the respective subtype-specific protocols, we define a DEG as protocol-common for a neural subtype. The 994 genes fulfilling this criterion in at least 14 of 16 neural cell types are defined as uDEGs. **d.** Distribution of correlation of uDEG log2FCs between different neural subtypes and protocols. **e.** UDEGs are classified into three categories: inconsistent log2FC (different subtypes show different DE direction), downregulated, and upregulated in organoids. **f.** Gene ontology enrichment analysis of downregulated (upper, blue) and upregulated (lower, red) uDEGs. The areas correspond to log-transformed adjusted p-values. **g.** Distribution of gene set scores (left: *mitochondrial ATP synthesis coupled electron transport*, right: *canonical glycolysis*) in primary neural cell types (upper, dark), and organoid counterparts (lower, light). P-values show the significance of the two-sided Wilcoxon test. Both scores are calculated with the *score_genes()* function in SCANPY. **h.** Glycolysis score, as a proxy of cell stress, of dorsal telencephalic excitatory neurons (dTelen VGLUT-N), split by the three primary developing human brains and 27 organoid datasets with at least 20 dTelen VGLUT-N. Datasets are ordered by their glycolysis score medians. The lower panel shows selected features of differentiation protocols potentially affecting cell stress. **i.** Spearman correlation between gene expression profiles of neural cell types in HNOCA and those in the human developing brain atlas [46], across variable TFs. Reproduced from [204] (Licensed under the Creative Commons Attribution-NonCommercial 4.0 International license: `https://creativecommons.org/licenses/by-nc/4.0/`).

3.2.4. Evaluating new neural organoid protocols using the Human Neural Organoid Cell Atlas

The HNOCA and our analytical framework to compare cell type composition and transcriptomic fidelity between organoids and a primary reference can also be used to contextualise new neural organoid datasets. For example, we used this method to examine the scRNA-seq data of a recently published multiplexed neural organoid morphogen screen [321]. Using scArches [42], we projected the data onto the same latent spaces as the HNOCA and the primary reference [46] (Fig. 3.11a, Supplementary Fig. B.12; HNOCA Supplementary Table 6). As before, this allowed us to transfer regional labels from the primary reference to the query cells. We found the predicted regional labels to be highly consistent with the original annotation while providing higher resolution within each broader brain region (Fig. 3.11b). This finer annotation allowed for a more detailed assessment of the effects of different morphogens on neuron generation in different brain regions (Fig. 3.11c). We calculated presence scores for each primary reference cell in each screen condition and compared them to the presence scores obtained for the HNOCA using hierarchical clustering on average presence scores of primary reference atlas clusters (Fig. 3.11d). We observed distinct presence score patterns in many screen conditions, which suggests that organoids generated under these conditions have regional cell type compositions that differ from those in the HNOCA datasets.

We summarised the max presence scores for the entire morphogen screen dataset to evaluate which primary reference cell types showed an increased presence in the screening data. We compared them to the respective scores calculated for the HNOCA data (Fig. 3.11e,f) and identified multiple reference cell clusters that displayed a substantial abundance increase under certain screen conditions (Fig. 3.11g). The most affected cell types included LHX6/ACKR3/MPPED1 triple-positive GABAergic neurons in the ventral telencephalon, the dopaminergic neurons in the ventral midbrain, and Purkinje cells in the cerebellum. Overall, this projection of morphogen screen query data to the HNOCA and the primary reference atlas enabled a refined regional annotation of the new data. It further allowed for an efficient evaluation of the capacity of novel organoid differentiation protocols to generate neuronal cell types that were previously underrepresented or absent in neural organoids.

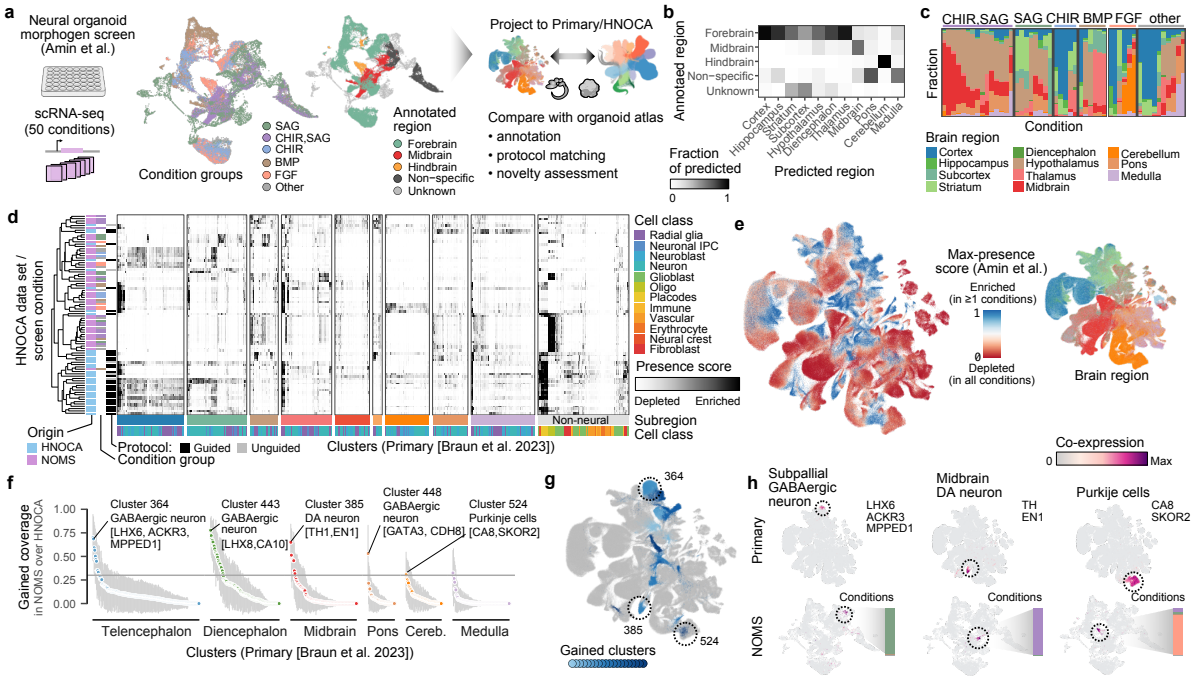

Figure 3.11.: **Annotating and evaluating neural organoid morphogen screen scRNA-seq data by mapping to HNOCA and the human developing brain atlas [46].** **a.** Projection approach for the neural organoid morphogen screen [321] scRNA-seq data to the HNOCA, using the human developing brain atlas [46] as the primary reference. The UMAPs show screen condition groups (left) and the regional annotation of the screen data (right). **b.** Similarity of the regional screen data annotation (rows) and the scArches-transferred regional labels from the primary reference. **c.** Regional composition of cells in the screen dataset from projection to the primary reference. Every stacked bar represents one screened condition. **d.** Clustering of average primary presence scores of HNOCA together with different conditions in the screen datasets. The dendrogram on the left shows the hierarchical clustering result. The heatmap shows average presence scores per cluster in the primary reference (columns), given data of each protocol involved in HNOCA or the screen condition (rows). **e.** UMAP of the primary reference data, coloured by the dissected regions (left) and the maximum presence scores across the screen conditions (right). **f.** Increase in cell cluster coverage from the screen conditions relative to HNOCA datasets, defined as the difference of average maximum presence scores per cluster of the primary reference with negative values trimmed to zero, between the screen dataset to HNOCA. The grey horizontal line shows the threshold (0.3) to define gained clusters in screen data. **g.** UMAP of the primary reference, with gained clusters highlighted in shades of blue. Dashed circles highlight three clusters with the highest coverage gain in the telencephalon, midbrain and hindbrain. **h.** Co-expression scores of cluster marker genes of the three clusters highlighted in g, in the primary reference (upper) and screen dataset (lower). Reproduced from [204] (Licensed under the Creative Commons Attribution-NonCommercial 4.0 International license: `https://creativecommons.org/licenses/by-nc/4.0/`).

3.3. Chronic exposure to glucocorticoids amplifies inhibitory neuron cell fate during human neurodevelopment in organoids

This result section corresponds to, and is in part identical to, the one presented in the following publication:

Dony, L., Krontira, A. C., Kaspar, L., Ahmad, R., Demirel, I. S., Grochowicz, M., Schäfer, T., Begum, F., Sportelli, V., Raimundo, C., Koedel, M., Labeur, M., Cappello, S., Theis, F. J., Cruceanu, C.*, Binder, E. B.* Chronic exposure to glucocorticoids amplifies inhibitory neuron cell fate during human neurodevelopment in organoids. *In review.* Preprint [205] doi: 10.1101/2024.01.21.576532 (2024)

"*" denotes an equal contribution.

The GC Supplementary Tables 1-9 referenced in this section are available online: `https://www.biorxiv.org/content/10.1101/2024.01.21.576532v1.supplementary-material`. The processed count matrices in H5AD format (scRNA-seq and scATAC-seq) with associated metadata, as well as all the code used in the analysis of the results presented in this section is available from ZENODO: `https://doi.org/10.5281/zenodo.10391945`. Raw scRNA-seq data (including filtered count matrices) are available from the Gene Expression Omnibus repository: GSE252522 (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE252522`): scRNA-seq *Veh*, *Chr*, *Chr-Veh*, and *Chr-Acu* conditions. GSE189534 (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE189534`): scRNA-seq *Veh-Veh* and *Veh-Acu* conditions; Line FOK4 = Line2; Line 409b2 = Line3; Veh-Veh = Veh, Veh-Acu = Dex. Raw scATAC-seq data (including filtered count matrices) are available from the Gene Expression Omnibus repository: GSE252523 (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE252523`).

Concerning the aims of this thesis, this section addresses Challenge 3, as outlined in Section 1.4: Improved *in vitro* models of human brain development in health and disease are the foundation for developing future treatment strategies. Novel model systems such as neural organoids have enabled promising advances in understanding the genetic contribution to the onset and development of psychiatric diseases in a human context. While it is widely accepted that environmental factors also strongly contribute to these pathological processes, the underlying mechanisms have been investigated to a significantly smaller extent.

My contributions to the results presented here are as follows: Together with Elisabeth B. Binder and Cristiana Cruceanu, I defined the main questions and analysis steps of this work.

I analysed all scRNA-seq and scATAC-seq data. I interpreted and visualised the results with input from the other authors.

This work studies the effects of an environmental challenge during neurodevelopment on gene expression and cell fate specification. We used the sGC dexamethasone as an example of a common environmental challenge during *in utero* brain development and a known risk factor for the development of mental illnesses. We quantified the transcriptional response of chronic GC exposure in neural organoids with scRNA-seq directly after the administration and following a wash-out period. We found that GC exposure does not induce significant metabolic stress in organoids, given appropriate QC filtering. Through mapping to HNOCA, we discovered that the neuronal lineage of our unguided organoid model covered exclusively the non-telencephalic areas of the brain (Section 3.3.1). We further identified many critical neurodevelopmental genes with altered transcription following chronic GC exposure (Section 3.3.2) and found that it leads to priming of the inhibitory neuron lineage (Section 3.3.3). We observed an increased number of inhibitory neurons following GC exposure both in the scRNA-seq data and in immunohistochemical stainings (Section 3.3.4). We identified PBX3 as an example of an important lineage driving TF following GC exposure (Section 3.3.5) and confirmed its association with the inhibitory neuron lineage by collecting scATAC-seq data and constructing multimodal GRNs (Section 3.3.6). Overall, we demonstrate a robust effect of GC exposure on neurodevelopment in organoids in the form of altered expression of key neurodevelopmental genes and priming of the inhibitory neuron lineage.

3.3.1. Cell viability is not significantly affected by chronic glucocorticoid exposure in neural organoids

We designed a chronic GC exposure paradigm in neural organoids to study the effects of GCs on neurodevelopment. We constructed scRNA-seq libraries from 70-day-old organoids exposed to GCs for ten days (*Chr* condition) and the respective vehicle control organoids (*Veh* condition). These samples allowed us to quantify the instantaneous transcriptional effects of chronic GC exposure. Additionally, we constructed scRNA-seq libraries from the *Veh* and *Chr* condition organoids after an additional 20 days under standard culture conditions at day 90 (*Veh-Veh* and *Chr-Veh* conditions, respectively) to identify any lasting effects. Lastly, we wanted to compare the transcriptional effects of chronic and acute GC exposure. We obtained additional samples at day 90, with added 12-hour acute exposure before collection (*Veh-Acu* and *Chr-Acu* conditions). We replicated all analyses in organoids from 2 human iPSC lines, Line 409b2 and Line FOK4, with 4 replicates in each of the 6 conditions (Fig. 3.12).

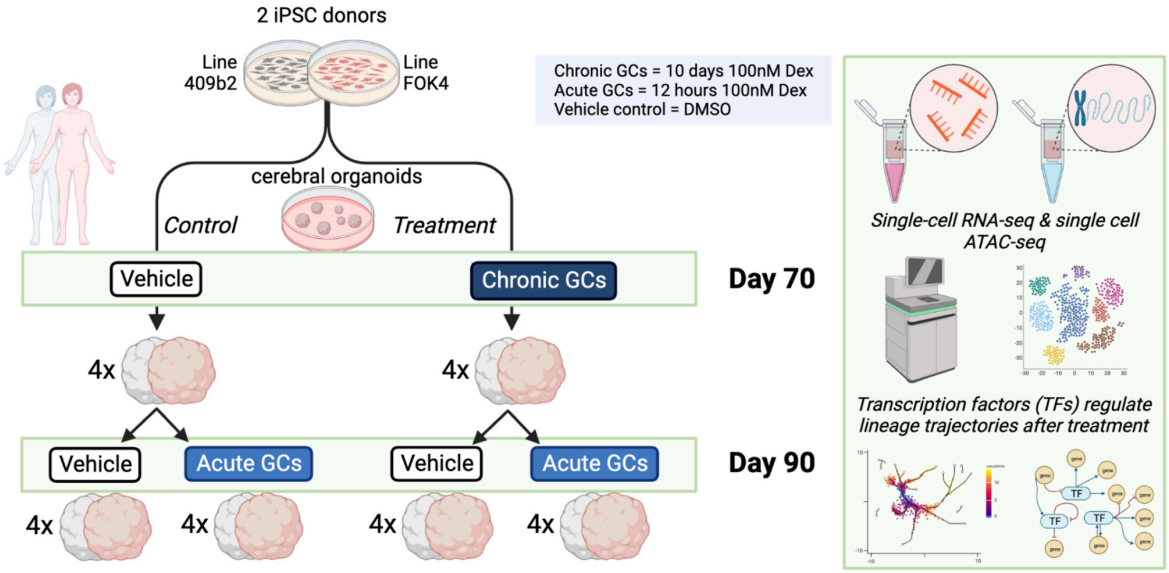

Figure 3.12.: **Experimental setup and design schematic.** We collected six treatment conditions duplicated in organoids from two cell lines. We replicated each of these 12 conditions in four independent samples. We collected two treatment conditions at day 70: *Veh* (exposed to the treatment vehicle dimethyl sulfoxide for ten days starting from day 60) and *Chr* (exposed to the GC dexamethasone for ten days starting from day 60). We derived the four additional treatment conditions collected at day 90 from the day 70 conditions with sustained culturing in regular media conditions for a further 20 days (wash-out period). The two conditions derived from the *Veh* condition were *Veh-Veh* and *Veh-Acu*, with an additional 12-hour acute GC exposure applied. Analogously, the two day-90 conditions derived from the *Chr* condition were *Chr-Veh* and *Chr-Acu*, with an additional 12-hour acute GC exposure applied. Figure created with BioRender.com and reproduced from [205] (Licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license: `https://creativecommons.org/licenses/by-nc-nd/4.0/`).

We analysed the data from the two cell lines separately, allowing us to focus on converging trends across lines. Using known marker genes of developing neural cell types, we identified eight out of the nine cell clusters in each of the lines (Fig. 3.13a; Supplementary Fig. B.13a):

- RG (GPM6B, SOX2)
- NPCs expressing cell cycle markers (TOP2A, MKI67)
- IPs (EOMES, PAX6)
- Excitatory Neurons (SLC17A6, STMN2)
- Inhibitory Neurons (GAD1, GAD2, SLC32A1, STMN2)
- A population of unspecified neurons expressing the G-protein regulator gene RGS5 (RGS5 Neurons) (RGS5, LINC00682, STMN2)
- Immature choroid plexus (ImmChP) cells (RSPO3, TPBG)
- Mature cells of the ChP (TTR, HTR2C, CLIC6).

We encountered difficulties identifying the ninth cluster (*Unknown*) in both datasets as it did not present any clear marker gene signature. Many cells in this cluster were flagged as metabolically impaired by scoring previously suggested pathways specific to non-viable organoid cells [104] (Fig. 3.13b). We identified 12 % of the cells in both datasets as non-viable (4304 cells in Line 409b2 data and 3559 cells in Line FOK4 data), with the *Unknown* cluster having the highest fraction of non-viable cells (74 % of cells in Line 409b2 and 41 % in Line FOK4), followed by the RGS5 Neurons cluster (54 % of cells in Line 409b2 and 15 % in Line FOK4). Line 409b2 also had a considerable fraction of IPs and ImmChP cells identified as non-viable cells (26 % and 20 %, respectively). Only small fractions of approximately 10 % or less were identified for all other cell types (Supplementary Fig. B.13b). We removed all cells identified as non-viable from the datasets, along with the entire *Unknown* cluster, as it was predominantly composed of non-viable cells and lacked clear expression of marker genes. After filtering out non-viable cells, we compared the viability scores between GC-exposed and control samples for different conditions and found no significant difference. This finding suggests that DE analyses between these conditions are unlikely to be confounded by differences in cell viability. Furthermore, more apparent differentiation trajectories emerged in low-dimensional space after removing non-viable cells (Fig. 3.13d, top).

In line with our previous research [284], we found that GC treatment did not cause significant variation in the low-dimensional embedding so that cells were mainly separated based on their cell identity (Fig. 3.13d, middle). Moreover, the GR gene NR3C1 expression was relatively uniform across most cell types and decreased with maturation, as previously reported [284] (Fig. 3.13d, bottom). To determine the regional identity of inhibitory and excitatory neurons in our dataset, we projected them onto the HNOCA [204]. Our neuron subtypes mapped to the inhibitory and excitatory non-telencephalic neuronal clusters of the HNOCA, while our RG primarily projected to non-telencephalic NPCs. Additional lineages included early cells of the glial lineage in organoids derived from Line 409b2 and a more pronounced lineage of the ChP in organoids derived from Line FOK4. Mapping to the HNOCA further confirmed matching neuronal cell identities in organoids from both cell lines (Supplementary Fig. B.13c).

3.3.2. Chronic glucocorticoid exposure affects the transcription of several key neurodevelopmental genes

We aimed to analyse the gene-regulatory and transcriptomic responses of our different GC treatment regimens. To minimise false-positive DE hits, we focused on converging results between the organoids derived from the two cell lines. We only considered a gene as a DEG if it was significantly regulated at a FDR smaller than 0.1 with a consistent direction of expression log2FC in organoids derived from both cell lines (Supplementary Fig. B.14a). This approach reduced our list of identified DEGs while providing potentially more robust results.

We initially focused on DEGs between the organoid conditions collected at day 70 by comparing *Veh* and *Chr* conditions separately for each identified cell type (Fig. 3.14a,b; GC Supplementary Table 1). This approach identified 803 consensus DEGs, with the most DEGs (n = 462) emerging in RG and the least in the ChP, with only three DEGs (Fig. 3.14c). However, certain cell types, specifically IPs and RGS5 Neurons, did not yield significant DEGs, likely due to the smaller cell numbers in these groups decreasing detection rates or higher variability of cells across the maturation gradients within these clusters.

Several genes important for neurodevelopment were among the most responsive DEGs by log2FC: NNAT, a gene associated with early neurodevelopment and ion channel control [322] (mean log2FC RG = 0.56); MAB21L1, associated with cerebellum development [323] (mean log2FC Excitatory Neurons = 0.31); NFIB, a TF known to be essential in brain development [324] (mean log2FC Inhibitory Neurons = -0.50); the transcriptional regulators ID3 (mean

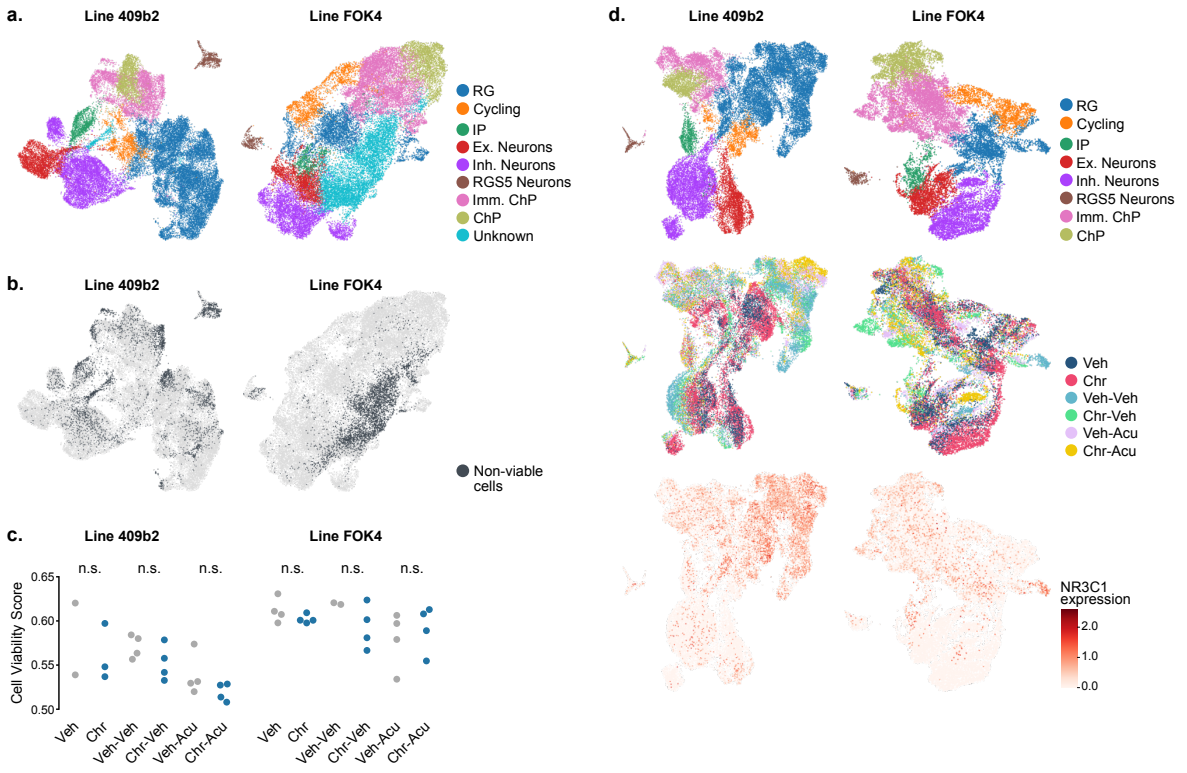

Figure 3.13.: **Chronic GC exposure in neural organoids does not induce significant metabolic stress in cells. a., b.** UMAP embeddings of Line 409b2 and Line FOK4 data coloured by cell type (a) and non-viable cell label (b). **c.** Distributions of mean viability scores between control (grey) and treated (blue) samples following non-viable cell removal. Each dot represents a sample from the indicated treatment condition. **d.** (top) UMAP embeddings of Line 409b2 and Line FOK4 data following non-viable cell removal. Coloured by cell type (top), treatment condition (middle) and NR3C1 (GR) expression (bottom). Reproduced from [205] (Licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license: `https://creativecommons.org/licenses/by-nc-nd/4.0/`).

log2FC in ImmChP = 0.46) and ID2 (mean log2FC in ImmChP and RG = 0.41 and 0.52 respectively). In addition to NFIB, we identified several differentially expressed TFs closely linked to neurodevelopment, such as SOX2, HDAC2, TCF7L2, PBX3, and YBX1. One of these TFs, YBX1, was differentially expressed in five cell types. Additional DEGs shared across five clusters included DLK1 (a regulator of hippocampal neurogenesis [325]), PCSK1N (associated with the neuroendocrine system [326]), the collagen gene COL3A1 (involved in neuronal migration [327]), and MARCKSL1 (associated with neural tube defects and regeneration [328]) (Fig. 3.14c).

We used the Gene Ontology Biological Process (GO-BP) pathways to probe our DEGs for functional enrichment. We identified neurodevelopment-associated terms in the clusters with the largest number of consensus DEGs, including: *axonogenesis*, *negative regulation of nervous system development* in RG; *axon development*, *substantia nigra development* in Excitatory Neurons; *cranial nerve development*, *regulation of neuron differentiation* in Inhibitory Neurons (GC Supplementary Table 2). However, the most significantly enriched terms did not converge on distinct neuronal pathways. Instead, they were associated with more general terms like cell-cycle regulation and intracellular transport in Excitatory Neurons and regulation of gene expression in RG and Inhibitory Neurons (Fig. 3.14d). Our list of identified DEGs also contained five of the 36 high-risk ASD genes related to transcriptional regulation previously evaluated in a knockout screen for their effect on cell fate determination [143]: ASH1L in RG, MYT1L and KMT2A in Excitatory Neurons, and FOXP1 and BCL11A in Inhibitory Neurons.

We evaluated the overlap of the immediate DE effect of chronic GC exposure with the lasting DE effect after the wash-out period at day 90 (GC Supplementary Table 3). Around half as many DEGs were regulated after wash-out compared to immediately following exposure (393 vs 803 consensus DEGs), and 13 % of the immediate transcriptomic effect (96 DEGs) was shared across both comparisons. The DEGs exclusively regulated in one of the two comparisons either corresponded to pathways only involved in a short-lived response or had already been translated to downstream transcriptional effects after the wash-out period. This would explain the 297 DEGs exclusively responding after the wash-out and could result from the relatively high fraction of TFs in the immediately regulated DEGs. Over half of the genes (61 %) regulated after chronic exposure and wash-out in 90-day-old organoids were also among the 2036 DEGs regulated following the 12-hour acute stimulation at day 90 (GC Supplementary Table 4). Additionally, the directionality of the DE effect was aligned for over 90 % of the DEGs shared between the comparisons.

We observed that more than half of the lasting DEGs after chronic exposure to GCs are

also part of the response seen after acute exposure to GCs in organoids at day 90. This implies that the lasting effects are still closely linked to direct GC effects even after a 20-day wash-out period. However, the overlap between the chronic and acute effects decreases to 40 % when considering the immediate day 70 DEGs (Fig. 3.14e), likely because of differences in neurodevelopmental age. The cells within each cell type were, in fact, separated based on the age of the organoid in the low-dimensional embedding, highlighting their transcriptomic differences (Supplementary Fig. B.14c). Furthermore, we found that 12 % of the genes that respond to acute GC exposure at day 90 had a significantly altered response based on the organoid's history of chronic treatment, which further underscores the long-term effect of chronic GC exposure (GC Supplementary Table 5).

3.3.3. Trajectory analyses reveal priming of the inhibitory neuron lineage following chronic glucocorticoid exposure

Moving beyond our DE analysis, which relies on discrete cell clusters, we were interested in understanding the effects of chronic GC exposure on the continuous differentiation trajectories in the evolving organoid system. We defined three lineage endpoints in our data: excitatory neurons, inhibitory neurons, and ChP and computed the associated lineage probabilities for every cell (Fig. 3.15a). The respective lineage probability increased steadily along the differentiation trajectory for each lineage endpoint, starting from RG and reaching their maximum at the endpoint (Fig. 3.15b). We excluded the ChP endpoint from the following analysis, given our specific interest in neuronal lineage determination and the absence of a sizeable number of DEGs in the ChP and ImmChP clusters. We validated our computed lineage probabilities in an independent public scRNA-seq dataset from 70-day-old organoids derived from 6 additional iPSC donor backgrounds [127]. We reprocessed the validation data (Fig. 3.15c) and applied the same lineage inference approach we used before. We found matching lineage trends towards inhibitory and excitatory neurons along the pseudo-temporal axis (Fig. 3.15d).

We used the concept of a driver gene introduced in [63] to link our observed DE effects to the computed lineage trajectories. For each gene in the transcriptome, we computed the correlation of its expression with the lineage probability for each endpoint across all cells in each of our two datasets and the validation data (8 total iPSC donors) (GC Supplementary Table 6). We then checked the alignment of the DE effect and driver gene direction for each DEG that was also one of the top 500 significantly associated genes for each neuronal lineage. Given an overlap of 50 % expected by chance, both neuronal lineages experienced a possible

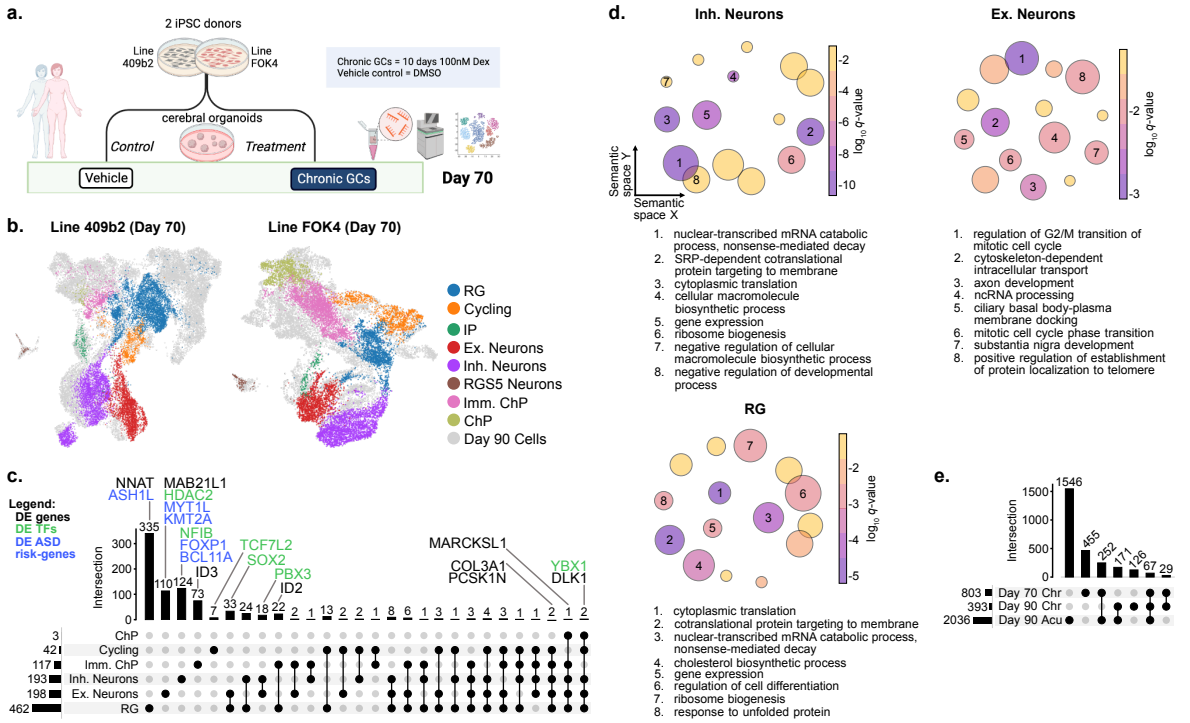

Figure 3.14.: **Key neurodevelopmental genes are differentially expressed following GC exposure. a.** Schematic of experimental design for day-70 samples only. Created with BioRender.com. **b.** UMAP embedding for day 70 data of Line 409b2 and Line FOK4 coloured by cell type. Cells from D90 samples are shown in grey. **c.** Upset plot showing consensus DE results by cell type and the number of unique and shared consensus DEGs. Selected genes are labelled, ASD risk genes are shown in blue, and further TFs are shown in green. **d.** Grouped semantic space representation of the gene ontology biological process enrichment results for the three cell types with the most detected DEGs. Circle size corresponds to the number of terms in the cluster; colour corresponds to the $\log_{10}$(q-value) of the enrichment of the representative term for each cluster. The integers within the circles enumerate the eight most significant clusters, and their representative term is written out in the legend below each plot. **e.** Upset plot showing DE results aggregated across all cell types for the effect of chronic GC exposure directly following the treatment (Day 70 Chr), the effect of chronic GC exposure after a 20-day wash-out period (Day 90 Chr), and the effect of acute GC exposure in 90-day old organoids (Day 90 Acu). Reproduced from [205] (Licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license: `https://creativecommons.org/licenses/by-nc-nd/4.0/`).

acceleration in differentiation following GC exposure, as more than half of the DEGs showed a DE effect aligned with driving the lineage. Interestingly, there was a significantly larger alignment of the DE effect in the inhibitory neuron lineage compared to the excitatory neuron lineage across the three datasets ($p = 0.04$), suggesting a somewhat selective priming of the inhibitory neuron lineage through GC exposure (Fig. 3.15e). Supporting this finding, we observed a significant positive correlation between the driver gene strength for the inhibitory neuron lineage in the validation dataset and the direction of the DE effect in the inhibitory cell clusters of our two organoid datasets ($r = 0.25$, $p = 0.002$) (Fig. 3.15f). Interestingly, BCL11A and FOXP1, two genes with the strongest correlation with inhibitory neuron lineage driver genes and the strongest upregulation following GC exposure, are associated with high risk for ASD in loss-of-function experiments [143]. Therefore, in the case of these genes, opposing regulatory effects of genetic and environmental disease risk lead to a convergent phenotype.

3.3.4. Chronic glucocorticoid exposure leads to an increased presence of inhibitory neurons in neural organoids

Building on the observed lineage priming of inhibitory neurons, we aimed to determine if exposing organoids to GCs resulted in a measurable shift in cell type identity. We used GAD1 as a specific inhibitory neuron marker across brain regions to identify cells of the inhibitory neuron lineage (Fig. 3.16a). Our two unguided organoid datasets showed an increased proportion of GAD1-positive cells at the RNA level after exposure to GCs: In Line 409b2, we observed a 1.7-fold increase (from 5.2 % to 8.7 %) and a 1.2-fold increase in Line FOK4 (from 11.6 % to 13.8 %) (Fig. 3.16b). We computed the inhibitory-to-excitatory neuron ratio to verify that the increase in GAD1-positive cells was not solely caused by an increased number of neurons present in GC-exposed organoids. We used GAD1 and the solute carrier SLC17A6 as markers for the two groups. The ratio consistently increased after GC exposure in both datasets. In Line 409b2, the ratio increased from 0.52 (*Veh*) to 0.86 (*Chr*), while in Line FOK4, it increased from 2.34 (*Veh*) to 2.94 (*Chr*).

We replicated our GC exposure paradigm in a new experiment using regionalised ventral organoids of a CRISPR/Cas9 edited version of Line 409b2 that expressed GFP-tagged GAD1 protein. By guiding differentiation with ventralisation factors, we were able to obtain an abundant representation of inhibitory neurons [281] (Supplementary Fig. B.15), allowing us to validate whether chronic GC exposure would also induce an increased inhibitory neuron abundance in a more natural environment for inhibitory neuron development. Using

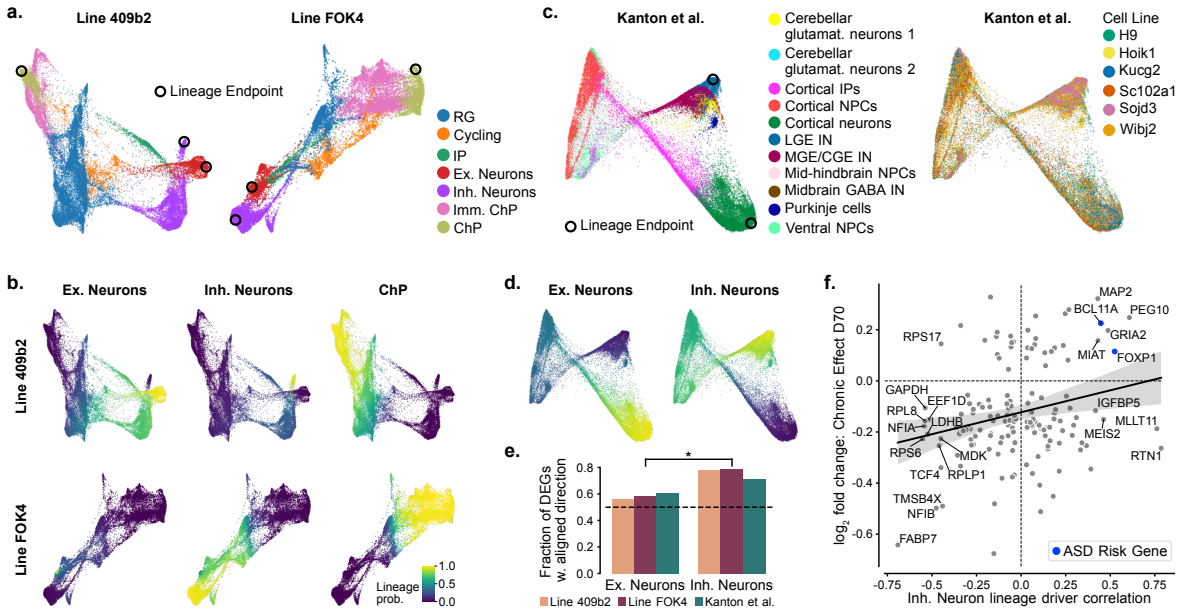

Figure 3.15.: **TF-induced priming of the inhibitory neuron lineage in neural organoids.** **a.** Force-directed graph embedding of organoid data from Line 409b2 and Line FOK4 (with the RGS5 Neuron cluster removed) coloured by cell type. Lineage endpoints are labelled with black circles. **b.** Lineage probabilities per cell for the three lineage endpoints in Line 409b2 and Line FOK4 data. **c.** Force-directed graph layout of validation data (published 70-day-old organoid data derived from six additional cell lines) [97]. Coloured by cell type (left) and cell line (right). Lineage endpoints are labelled with black circles. CGE, caudal ganglionic eminence; MGE, medial ganglionic eminence; LGE, lateral ganglionic eminence; IN, interneuron; glutamat., glutamatergic; IPs, intermediate progenitors; NPCs, neural progenitor cells. **d.** Force-directed graph layout of the validation data from six cell lines, coloured by lineage probabilities for the two neuronal lineage endpoints. **e.** Fraction of consensus DEGs with aligned log2FC direction and driver gene direction (based on the top 500 significant driver genes, recomputed for each of the three datasets). The dashed line indicates the fraction expected by chance (0.5). **f.** Magnitude of driver gene correlation with the inhibitory neuron lineage in the validation data [321] vs log2FC of consensus DE effect measured in our two cell lines. Genes with the highest lineage correlation are labelled by name, and genes associated with high risk for ASD [143] are marked in blue. Reproduced from [205] (Licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license: `https://creativecommons.org/licenses/by-nc-nd/4.0/`).

fluorescent microscopy and immunohistochemistry stainings, we confirmed an increased level of GAD1 protein after GC exposure (Fig. 3.16c). Furthermore, chronic GC exposure led to a significant 2.27-fold increase in GAD1-positive cells (p = 0.048, Fig. 3.16d, GC Supplementary Table 7). We reproduced this finding in an additional staining experiment showing a fold-change of 2.11 (p = 0.0043) (GC Supplementary Table 7).

We noted a sustained increase in the abundance of GAD1-positive cells at the RNA level in GC-exposed organoids at day 90 following wash-out (Fig. 3.16b). In Line 409b2, the percentage of GAD1-positive cells increased by 1.2-fold, from 3.4 % to 3.9 %. In Line FOK4, the increase was 1.5-fold, from 13.5 % to 20.1 %. However, we observed inconsistent results in the inhibitory-to-excitatory neuron ratio in our two datasets using GAD1 and SLC17A6 as markers at day 90. Specifically, in Line 409b2, the ratio decreased from 1.54 (*Veh*) to 1.15 (*Chr*), whereas in Line FOK4, it increased from 5.29 (*Veh*) to 6.30 (*Chr*).

Overall, our organoid system showed evidence of priming of the inhibitory neuron lineage following exposure to GC, resulting in an increased number of GAD1-positive cells at the RNA and protein level in separate experiments in both unguided and ventralised organoids.

3.3.5. Increased PBX3 expression drives inhibitory neuron priming following chronic glucocorticoid exposure

To understand the mechanisms behind the priming of the inhibitory neuron lineage, we determined the lineage-driving TFs with the most considerable expression changes after GC exposure in our datasets. We found 15 TFs that had an aligned direction of expression log2FC (in inhibitory neuron consensus DEGs) and driver correlation (with the inhibitory neuron lineage in the validation data) (Fig. 3.17a). We postulated that a TF actively mediating the lineage priming would be differentially expressed after GC exposure and have a significant number of its target genes regulated. We identified 18 TFs that fulfilled these criteria across the DEGs in all cell types: BCL6, EGR1, ID2, ID3, ID4, NEUROD1, NEUROD2, NFE2L2, NFIA, NFIB, NR2F1, NR2F2, NRG1, PBX3, SALL2, SOX2, TFAP2A, and YBX1 (GC Supplementary Table 8).

Intersecting the two TF subsets above revealed the hox-gene PBX3 as a positive inhibitory neuron lineage driver in the validation data that is upregulated by GCs. Conversely, NFIA, NFIB, EGR1, and YBX1 were downregulated by GCs and negatively correlated with the inhibitory neuron lineage in the validation data. PBX3 was the only TF consistently among

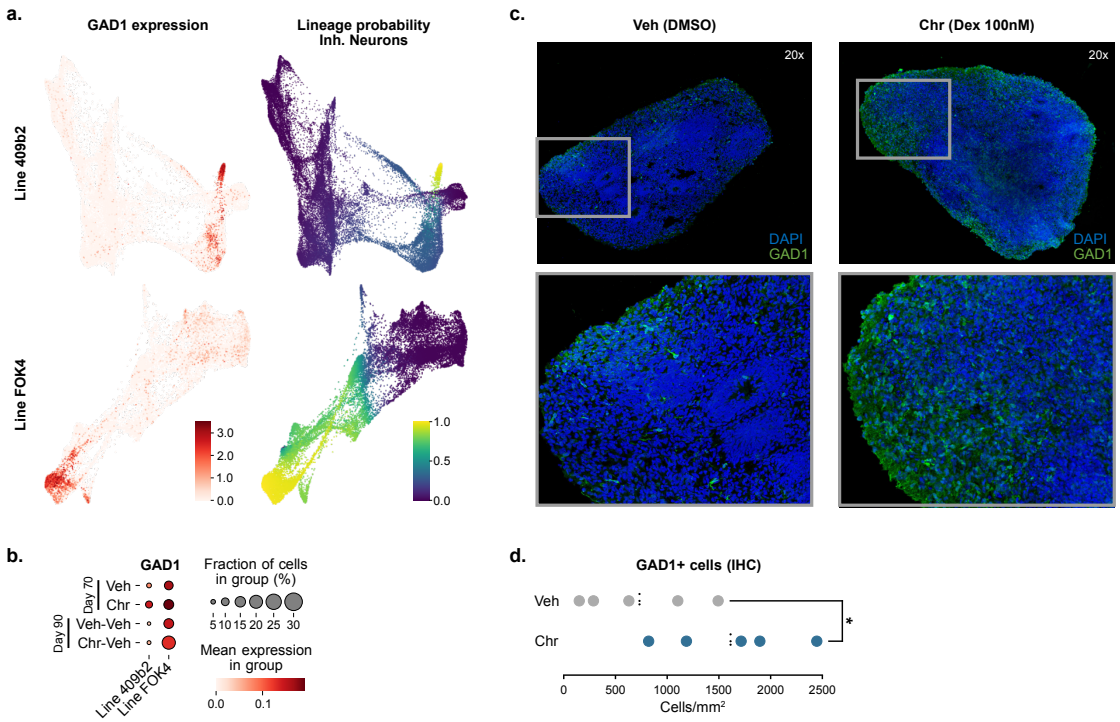

Figure 3.16.: **Increased abundance of inhibitory neurons in GC-exposed organoids. a.** Force-directed graph layout. Coloured by expression of the inhibitory neuron marker GAD1 (left) and absorption probability per cell for the inhibitory neuron lineage (right). **b.** Fraction of GAD1-positive cells and mean GAD1 expression across all cells in *Veh*, *Chr*, *Veh-Veh*, and *Chr-Veh* conditions for both cell lines. **c.** Representative images of whole slice ventralised organoids at day 70 in culture, after ten days of chronic treatment with GCs (100nM dexamethasone; *Chr* condition - right) and control (*Veh* condition - left). Images were acquired at 20x magnification, showing DAPI (blue) and GAD1 (green) stainings. Lower panel: zoomed-in inserts. DMSO, dimethyl sulfoxide; Dex, dexamethasone. **d.** Density of GAD1-positive cells across entire organoid tissue slices (n = 5 per condition). The mean density for each condition is indicated as a dotted black line. IHC, immunohistochemistry. Reproduced from [205] (Licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license: `https://creativecommons.org/licenses/by-nc-nd/4.0/`).

the top ten per cent of driver genes for the inhibitory lineage in all three datasets (top 1.4 % - 6.5 %, depending on the dataset). At the same time, it was not among the top ten per cent of drivers for the excitatory neuron lineage in any of the datasets (Fig. 3.17b, Supplementary Fig. B.16a). While all five identified TFs might play a role in the lineage priming, PBX3 showed the most evident effect and, therefore, remained the focus of our further analysis as an example of a GC-responsive lineage-priming TF. PBX3 was expressed across all cell types, with the highest expression levels in inhibitory neurons (Fig. 3.17c, Supplementary Fig. B.16b). PBX3 is associated with hindbrain-related processes such as breathing, locomotion, and sensation [329].

We visualised the expression of PBX3 along the pseudo-temporal lineage trajectories. We found it to increase sharply toward the inhibitory neuron endpoint, backing its relationship with this lineage (Fig. 3.17d, Supplementary Fig. B.16b). This was in line with our observations from a recent atlas of first-trimester human brain development [46], which showed that PBX3 expression is developmentally regulated, with overall higher expression in GABA-ergic compared to glutamatergic neurons and higher in non-telencephalic compared to telencephalic brain regions, with the highest expression in the cerebellum (Supplementary Fig. B.16c). These findings suggest that PBX3 plays a consistent role in the development of fetal brain and neural organoids.

We observed an increased fraction of PBX3-positive cells following GC exposure at the transcriptional level in both cell lines: from 21 % to 26 % in Line 409b2 and 22 % to 38 % in Line FOK4. Moreover, we also noticed an increase in the fraction of double-positive cells for PBX3 and GAD1: from 1.7 % to 3.9 % in Line 409b2 and from 4.5 % to 8.5 % in Line FOK4. Across double-positive cells from both datasets, we observed a significant positive correlation between the expression levels of PBX3 and GAD1 (Line 409b2: r = 0.43, p = 4.1e-16; Line FOK4: r = 0.36, p = 2.8e-15) (Supplementary Fig. B.16d).

We validated our findings as before in ventrally-guided organoids at day 70, using GFP-tagged GAD1 and immunohistochemistry stainings of PBX3 to visualise protein abundances. Our observations were consistent with the scRNA-seq data, showing that the PBX3 protein was expressed mainly in more mature neurons that had already migrated to the outer ventricular zone (Fig. 3.17e). We quantified the number of PBX3-positive cells across the entire organoid slices and found a 1.73-fold increase in chronically GC-exposed organoids compared to controls (p = 0.022; Fig. 3.17f, GC Supplementary Table 7). We also counted cells double-positive for PBX3 and GAD1 and found a 3.35-fold increase following GC exposure(p = 0.0041; Fig. 3.17g, GC Supplementary Table 7). Altogether, the results presented in this section

suggest that GC-induced overexpression of PBX3 plays a role in priming the inhibitory neuron lineage.

3.3.6. Multimodal gene regulatory networks support the role of PBX3 in mediating lineage priming

To investigate the role of PBX3 and the regulatory mechanisms underlying lineage priming, we obtained scATAC-seq data for 90-day-old organoids from Line 409b2 using the same treatment paradigm as before. Analysing the gene expression data of the Line 409b2 organoids alone showed similar trends as observed in our joint analysis earlier: the magnitude of the DE effect in inhibitory neurons was still significantly positively correlated with the driver gene strength for the inhibitory neuron lineage with 70-day-old organoids of Line 409b2 analysed individually (*Veh* vs *Chr*) (r = 0.077, p = 0.046). We found that after the 20-day wash-out period lasting until day 90 (*Veh-Veh* vs *Chr-Veh*), this correlation increased in Line 409b2 organoids (r = 0.15, p = 4.7e-3), suggesting a long-term effect of the treatment on lineage determination (Fig. 3.18a).

We integrated the scATAC-seq data with the matching scRNA-seq samples to infer multimodal GRNs for treatment and vehicle organoids to uncover the gene regulatory mechanisms underlying inhibitory neuron lineage priming (Fig. 3.18b, GC Supplementary Table 9). We used the expression of TFs and their target genes, accessibility of TF binding sites, and prior biological information, such as conserved regions of the genome, to infer these GRNs with the Pando tool [254]. Centring the vehicle organoid GRN on PBX3 allowed us to visualise its baseline downstream regulatory interactions. We found that over a third (35 %) of all genes in this TF-centred GRN were differentially expressed in at least one of the three GC treatment conditions (*Veh* vs *Chr*, *Veh-Veh* vs *Chr-Veh*, and *Veh-Veh* vs *Veh-Acu*) (Fig. 3.18c), highlighting the role of PBX3 in mediating the transcriptional response to GC exposure. In the PBX3-centered GRN of GC-exposed organoids, 36 % of direct downstream targets were gained compared to the vehicle condition. Additionally, 30 % of PBX3 downstream targets were among the top 500 drivers of inhibitory neuron lineage (Fig. 3.18d). This observation supports our earlier finding that large portions of the neural organoid GRN are responsive to GCs and linked to inhibitory neuron lineage priming.

We also compared the relationship of PBX3 with the top 500 inhibitory drivers between the vehicle and GC-exposed GRNs. We observed a considerable relative increase of 32 % (from 0.54 to 0.71) in the fraction of inhibitory neuron driver genes in the direct PBX3 targets

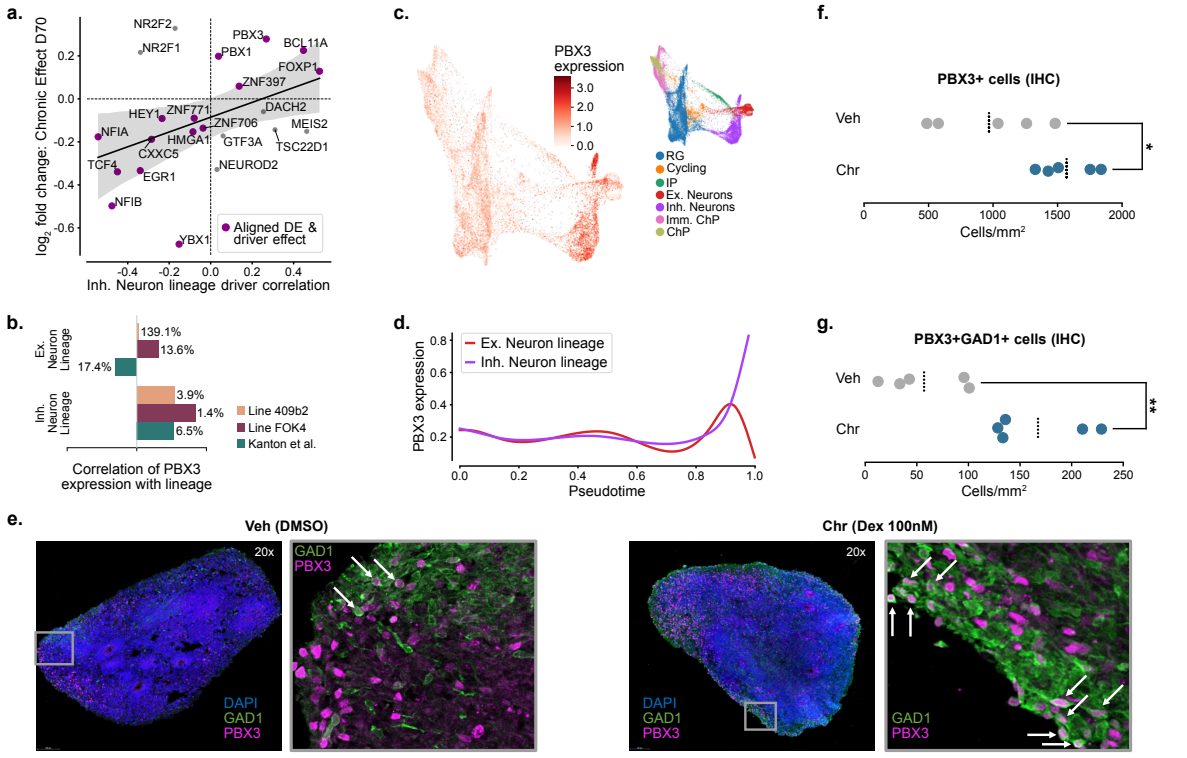

Figure 3.17.: **Inhibitory lineage priming is aligned with GC-induced PBX3 regulation.** **a.** Magnitude of driver gene correlation with the inhibitory neuron lineage in the validation data [97] vs log2FC of GC day-70 DE effect measured in our two cell lines. Showing only TFs. Genes with an aligned direction of log2FC and lineage correlation are marked in purple. **b.** Correlation of PBX3 expression levels with lineage probability across the excitatory and inhibitory neuronal lineages in all three datasets. The percentile of PBX3 among all significant driver genes ranked by driver strength is written next to each bar. **c.** Expression of PBX3 on a force-directed graph embedding of Line 409b2 data (left) with cell type reference (right). **d.** Pseudo-time-resolved expression of PBX3 for each of the two neuronal lineage endpoints in Line 409b2. **e.** Co-expression of GAD1-positive cells (green) with PBX3-positive cells (magenta) in 70-day-old ventralised organoids of Line 409b2 in *Veh* and *Chr* conditions. 63x magnification zoom-in images are shown to the right of the 20x whole-slice images. White arrows mark examples of double-positive cells. DMSO, dimethyl sulfoxide; Dex, dexamethasone. **f.** Distribution of PBX3-positive cell density across entire organoid tissue slices (n = 5 per condition). The mean density for each condition is indicated as a dotted black line. IHC, immunohistochemistry. **g.** Distribution of PBX3-GAD1-double-positive cell density across entire organoid tissue slices (n = 5 per condition). The mean density for each condition is indicated as a dotted black line. Reproduced from [205] (Licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license: `https://creativecommons.org/licenses/by-nc-nd/4.0/`).

following GC exposure (Fig. 3.18d). Notably, some of the direct and second-order PBX3 target genes in the GRN of GC-exposed organoids showed a high correlation with the inhibitory neuron lineage and a robust expression change after exposure to GCs. Examples included CELF4, a gene associated with synaptic development [330], depression-like behaviour in mice [331], and ASD [332], as well as SYNPR, a common inhibitory neuron marker gene [333, 334] (Fig. 3.18a). Overall, introducing a second data modality into the analysis underscored a possible involvement of PBX3 in GC-mediated inhibitory neuron lineage priming.

This concludes the results section of this thesis. The results of the three interconnected contributions introduced in this section are each summarised again at the beginning of the relevant discussion sections on the following pages.

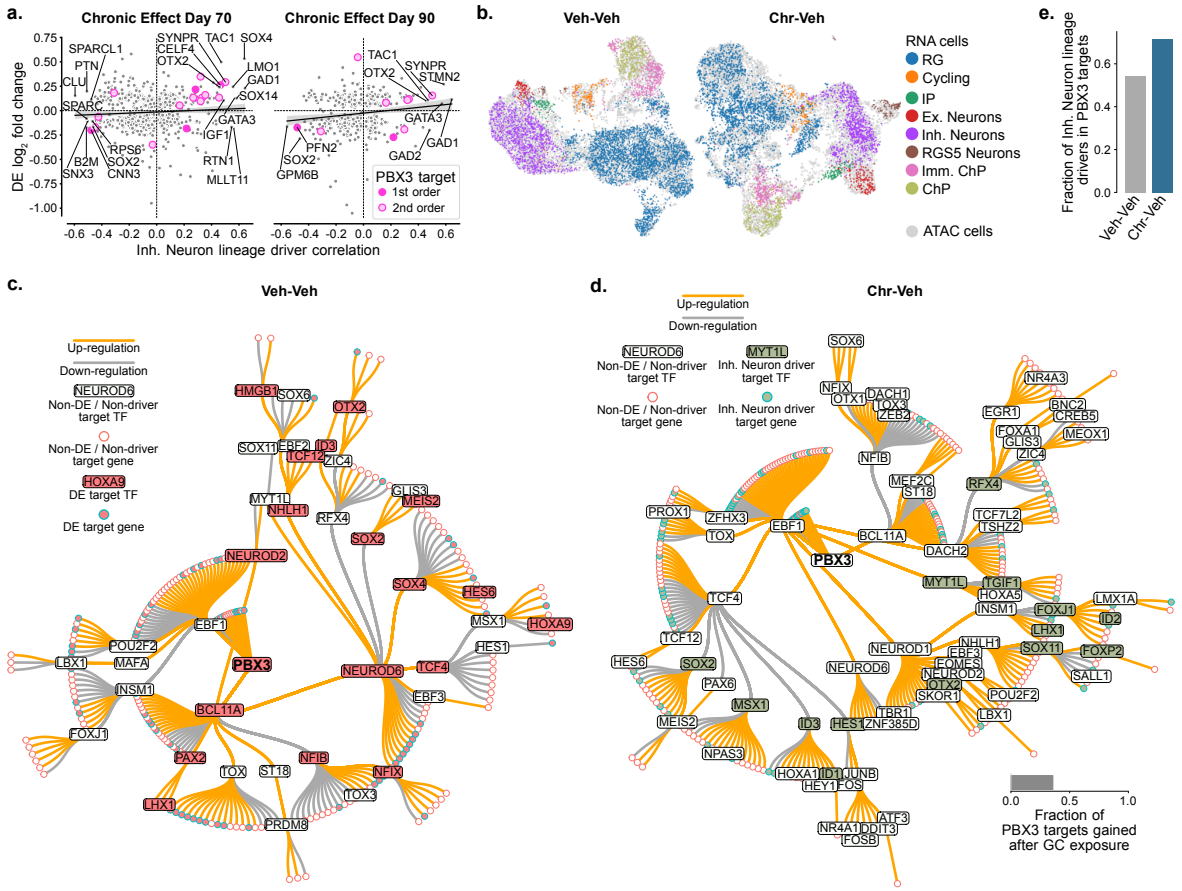

Figure 3.18.: **Multi-modal GRNs associate PBX3 with inhibitory neuron priming in organoids from Line 409b2. a.** Magnitude of driver gene correlation with inhibitory neuron lineage probability vs log2FC of Line 409b2. First and second-order PBX3 target genes in the inferred chronic GRN are labelled in pink. Left: Directly following treatment (70 days in culture). Right: after 20 days of wash-out (90 days in culture). Genes with an absolute lineage correlation greater than 0.45 are labelled by name. **b.** UMAPs of integrated scRNA-seq and scATAC-seq data of Line 409b2 at 90 days in culture. ScRNA-seq data is coloured by cell type, and scATAC-seq data is shown in grey. Vehicle organoid data (left) and GC-exposed organoid data (right). **c.** GRN centred around PBX3 in vehicle organoids with consensus DEGs from any of the three DE comparisons: D70 Chr, D90 Chr, D90 Acu) coloured red and TFs labelled by name. **d.** GRN centred around PBX3 in treated organoids with top 500 inhibitory neuron driver genes coloured green, TFs labelled by name. The bar chart shows the fraction of newly gained direct PBX3 downstream targets in the *Chr* condition. **e.** Fraction of inhibitory neuron drivers in direct TF downstream targets for control (*Veh-Veh*) and GC exposed organoids (*Chr-Veh*) of Line 409b2 at 90 days in culture. Reproduced from [205] (Licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license: `https://creativecommons.org/licenses/by-nc-nd/4.0/`).

4. Discussion

Neural organoids, paired with single-cell genomics, have emerged as a powerful tool to study neurodevelopment and the onset and development of psychiatric diseases. The work presented in this thesis can be structured into connected contributions in three areas: infrastructure (for scRNA-seq data curation), data (a neural organoid reference atlas), and discovery (of GC exposure effects on neural organoids). These contributions build on top of each other towards improving our understanding of neural organoids as a model system and using them to study the effect of an environmental risk factor for mental illnesses on human brain development.

As the first contribution, detailed in Section 3.1, I presented sfaira, a data and model zoo that facilitates scRNA-seq data curation and analysis and enables data reuse and model sharing. sfaira's data curation capabilities helped to compile a comprehensive reference atlas of neural organoids, the second contribution of this thesis (Section 3.2). The data loading interface and the ontology-based metadata management of the sfaira data zoo was designed with organoid models in mind and greatly facilitated the collection, harmonisation, loading, and subsetting of the relevant neural organoid datasets sourced from a heterogeneous range of data repositories. Beyond supporting the curation process, sfaira improved the reproducibility of the HNOCA work by significantly reducing the complexity of the custom data loading code provided with the atlas and making the individual curated constituent datasets easily accessible by the community. In the context of the third contribution of this thesis (Section 3.3), the HNOCA served as a reference atlas to annotate the newly generated neural organoid data, which models the effect of an environmental challenge in the form of GCs on human neurodevelopment. Query-to-reference mapping of this data onto HNOCA efficiently identified the neuronal and progenitor sub-populations as derived from non-telencephalic brain regions, which are less commonly represented in unguided neural organoids. This assignment helped to contextualise the findings concerning the identified priming of the non-telencephalic inhibitory neuron lineage and the lineage-driving TFs specific to non-telencephalic brain regions.

In this thesis, I provided an overview of relevant background topics in Chapter 1, including single-cell transcriptomics and related computational analysis approaches, neural organoids and their applicability as disease models, as well as the role of GCs in neurodevelopment and psychiatric disease. In Chapter 2, I described the experimental and computational materials and methods underlying the results presented in Chapter 3. In the remainder of this chapter, I summarise the findings related to each of the three contributions of this thesis and discuss their limitations and remaining open questions.

4.1. Facilitating data curation and reuse in single-cell genomics

In Section 3.1 of the results chapter, I introduced sfaira, a data and model zoo for single-cell genomics. sfaira addresses a central open question in single-cell genomics: while an ever-growing amount of published scRNA-seq datasets is becoming available, there is a profound lack of standardisation for storage and data formats, storage location, and metadata vocabularies, limiting data reuse. The same is true for scRNA-seq data analysis, generally done from scratch for new datasets without efficiently leveraging existing pre-trained models. In this thesis, I outlined how the sfaira data zoo, through its ontology-controlled metadata management and structured data loading code repository, facilitates data curation and provides a simple interface for retrieving curated datasets, greatly facilitating data reuse. I have further described how the sfaira model zoo, the first-of-its-kind structured model repository for pre-trained scRNA-seq analysis models, enables efficient model reuse through a model versioning system and an easy-to-use interface. Additionally, I explained how sfaira can be leveraged to efficiently query summary statistics on the sfaira data zoo, contextualising new observations and enabling an automated zero-shot embedding and cell type assignment of new data, speeding up an initial analysis. Lastly, I outlined how sfaira models can be interpreted using gradient-based methods to derive further biological insight from the data. Overall, sfaira is a step towards aligning single-cell genomics data with the FAIR principles [65].

Since its publication, I have further improved sfaira with the help of my collaborators. We added a command line interface, guiding the user through an interactive data curation process while supporting the mapping of metadata to the relevant ontologies and validating all steps in the process. This addition speeds up data curation and has led to significant growth of the sfaira data zoo through the community and us authors. As of March 2024, the number of studies available through the sfaira data zoo has more than doubled. It currently

encompasses 118 studies featuring 962 datasets with over 15 million cells. As mentioned in the previous section, SFAIRA has also supported the data curation for the first human neural organoid cell atlas, boosting the reproducibility of the atlas construction and enabling downstream analyses across all datasets through harmonised metadata.

The *CZ CELLxGENE Discover Census* [70] represents a recent advancement in public scRNA-seq data infrastructure, offering access to curated scRNA-seq datasets through a unified interface. While the goals of SFAIRA and the Census, fostering data curation and reuse in single-cell genomics, are well aligned, they represent complementary approaches. While the Census focuses on creating a comprehensive, accessible cloud database of scRNA-seq datasets, SFAIRA is based on a decentralised design that enables on-premise data curation, addressing concerns around data privacy and proprietary information. SFAIRA has been one of the first tools to serve large-scale curated scRNA-seq datasets. It has supported the training of several extensive representation learning models and facilitated data curation in multiple projects. The long-term success of a curated data repository relies on the community's continuous contribution and curation of new data and requires significant engineering and infrastructure work. While SFAIRA will likely remain a valuable companion for custom on-premise data curation, ongoing large-scale efforts such as the *CZ CELLxGENE Discover Census* will likely become the *de facto* standard source for curated datasets in the future. The value of curated and easily accessible data has also led to the creation of commercial data curation and management providers such as LAMIN (`https://lamin.ai/`), offering custom solutions for commercial and academic entities. The ongoing growth and maturation of data curation efforts pave the way towards a more practical application of deep-learning approaches to single-cell genomics, which rely on large-scale, high-quality training data.

The recently emerging large-scale, often transformer-based, machine learning models (also called foundation models) in single-cell genomics [43, 44, 335] promise a new approach to scRNA-seq data analysis. Inspired by the recent breakthroughs of large language models in natural language processing, these models leverage vast amounts of data to uncover biological insights and predictive capabilities previously out of reach [45]. While the emerging model architectures are generally moving away from simple encoder-decoder-based models as served through the SFAIRA data zoo, this development highlights the importance of large-scale data curation in single-cell genomics and the value of curated data repositories. This is because the performance of large-scale foundational models is highly dependent on the availability of extensive, curated, high-quality training data. Consequently, recent approaches have relied on public curated repositories like the *CZ CELLxGENE Discover Census* [44] or extensive proprietary curated databases [335]. In a related ongoing project in our research group, the

SFAIRA repository efficiently contributed curated scRNA-seq data for close to ten million cells to the training of a foundational model on top of the data available through existing public curated repositories, significantly reducing the manual curation work required. As the volume and variety of data in curated data repositories grow, so does the potential for foundation models to revolutionise our understanding of complex biological systems, demonstrating the critical role of large-scale data curation in driving the capabilities of computational biology and single-cell genomics.

4.2. A path to improving human neural organoid models

In Section 3.2 of the results chapter, I introduced the HNOCA, a comprehensive, harmonised neural organoid reference atlas. HNOCA addresses a central challenge in the neural organoid field. While neural organoids are versatile models of early human brain development, there is still significant room for improving their fidelity and faithfulness to human neurodevelopment. This would increase their suitability as universal models of the developing human brain in health and disease. Consequently, a structured analysis of the cell type diversity and transcriptomic fidelity of current organoid models is needed to suggest concrete improvements to organoid differentiation protocols. Furthermore, a comprehensive reference dataset is lacking in the neural organoid field, complicating the analysis of new datasets.

In this thesis, I outlined how my collaborators and I constructed the first large-scale integrated cell atlas of human neural organoids, the HNOCA. Using a custom pipeline, we integrated 1.8 million cells spanning 36 scRNA-seq datasets generated by 15 laboratories worldwide using 26 differentiation protocols and various scRNA-seq technologies. I described the high complexity of neuronal, glial and non-neural cell types, revealed by the HNOCA, that can develop in neural organoids grown under existing protocol conditions. I explained how we refined the regional cell type annotations by mapping the HNOCA to a recent primary human brain atlas [46], further allowing us to identify primary neural cell states currently under-represented or absent in organoids. I outlined our DE analysis approach between organoids and primary fetal brain, which allowed us to identify a universal signature of cell stress specific to organoid cells and informed our comprehensive analysis of transcriptomic fidelity across neuronal organoid cell types. I lastly described how the HNOCA can be used as a reference dataset by the community using the example of a recently published neural organoid morphogen screen [321]: We demonstrated how the HNOCA and our analysis framework enabled a refined cell type annotation of the new data and the identification of

novel cell states previously unattainable with neural organoid differentiation protocols. The public portion of the HNOCA is available to the community as an integrated neural organoid reference atlas, including the complete integration and analysis pipeline. We anticipate to release the remaining datasets soon, too. Overall, this framework and reference atlas facilitate the quantitative and comparative analysis of scRNA-seq neural organoid data and the evaluation of novel neural organoid protocols.

One evident limitation of the presented HNOCA analysis is the use of only a single primary fetal brain dataset as a reference point. While the dataset [46] is one of the most extensive scRNA-seq studies of first-trimester fetal brain development, any technical artefacts specific to this dataset could be misinterpreted as a biological signal in the DE analysis. To rule out any significant bias, my collaborators and I repeated a section of this analysis using additional datasets and observed matching trends. Nevertheless, a valuable extension of our work would be the construction of a comprehensive, integrated fetal reference dataset comprising data from multiple labs and more mature developmental time points. Such a reference would allow more sensitive and precise detection of transcriptomic differences between neural organoids and the fetal brain, potentially uncovering additional systematic differences beyond the reported universal stress signature. Additionally, it would enable a comprehensive timing analysis of organoid age and fetal brain development. This information would ease the selection of the right organoid age for experiments based on the relevant fetal developmental stage, aiding the design of more meaningful organoid experiments.

The DE analysis in this work could be further improved by accounting for differences in developmental age between the primary and organoid data in the DE model. In the current analysis, it is not fully clear if the downregulation of neurodevelopment-associated functional terms in organoids results exclusively from the inherent difference between primary and organoid tissues or is partly also driven by differences in developmental age between the two groups. Explicitly accounting for developmental age in the model would alleviate this ambiguity, though the exact developmental mapping between primary and organoid tissue must first be established as described above. A third improvement to this study would be to create a user-friendly interactive web interface for interacting with the HNOCA resource. This interface would provide a mapping functionality for new datasets to HNOCA and generate key metrics introduced in this work, such as the maximum presence score, for all query cells without the need to write any code. Such a service would significantly broaden HNOCA's reach by making it available beyond the computational biology community.

The analysis of the HNOCA has revealed several primary fetal neural cell states under-

represented or absent in organoids. I am currently planning follow-up experiments that use this information to increase the coverage of these primary neural cell states in organoids. Such experiments could leverage the emerging experimental data from neuronal genetic perturbation screens focused on patterning [254, 336] or morphogen screens [321, 337] to identify promising approaches to improve guided organoid development. Despite only a limited number of morphogens with known roles in human neurodevelopment being currently considered for guiding organoid development, the importance of exact morphogen concentrations and timing creates a vast combinatorial space that would be extremely expensive to cover experimentally. Over the past years, machine learning approaches have emerged as powerful tools for *in silico* perturbation prediction, with the potential to drastically reduce the number of required wet-lab experiments [80, 338–340]. I am currently considering two ways to tackle this challenge. In the first approach, I would train perturbation prediction models, such as scGen [80] or CPA [338] on the abovementioned morphogen screen data. I then aim to predict the missing concentration and timing conditions for the existing morphogens *in silico*. Given a well-calibrated prediction model, quantifying the uncertainty of the predictions would allow us to define follow-up screening experiments to maximise the predictive power of the model, given new training data. Ideally, experimental screening and computational prediction would operate in an integrated and iterative *lab-in-the-loop* process where the predictions guide the experiments and the data generated by the experiments is used to improve the computational models further. In a second, more exploratory approach, I would use the Prophet model, currently being developed in our research group, to identify chemical compounds that could lead to a more primary tissue-like cell type composition. Extending beyond morphogens known from developmental biology, this approach would predict the resulting organoid cell type composition for a given cell line, compound, and concentration based on a database of available compounds. As this approach represents a more complex prediction task, it would require collecting additional training data before it can be meaningfully applied to this challenge. Overall, the combination of high-throughput genetic and chemical perturbation screens and advanced computational prediction approaches [341] paves the way for efficiently screening the differentiation protocol space, ultimately leading to neural organoids more faithfully representing fetal brain development.

The HCA project aims to map all primary cell states of the human body. Beyond this, it also seeks to chart the cell diversity of current organoid systems across tissues in a dedicated HCA BioNetwork. Together with the Human Endoderm Organoid Cell Atlas [342], the HNOCA will represent the first output of the organoid BioNetwork once officially approved. A natural extension of our work would extend a similar analysis to organoids of different tissues not covered by the two current studies. One example could be a mesoderm-derived organoid

atlas, including, for example, heart and kidney organoids. Once organoids of all major tissues are included, an integrated pan-organoid atlas could provide the complete picture of the current state of organoid technology and enable large-scale inter-tissue comparisons. In a second direction, the HNOCA could be extended towards additional data modalities, incorporating, for example, spatial resolution [343, 344] or information on chromatin state [254, 345, 346] in neural organoids. The latter addition would enable the inference of GRNs in large-scale neural organoid data, providing further insight into the intricate gene-regulatory mechanisms underlying human brain development.

4.3. Insight into the molecular mechanisms underlying the effects of environmental risk for psychiatric disease

In Section 3.3 of the results chapter, I introduced our chronic GC treatment paradigm in neural organoids. As an example of a common prenatally encountered environmental risk factor, our GC exposure paradigm provided insight into the molecular underpinnings of environmentally mediated psychiatric disease risk. This work addresses a central challenge in the field of psychiatric disease research. Psychiatric diseases are widely accepted to be caused by a complex interplay of genetic and environmental risk factors [153]. There have been several recent advances in understanding the onset and development of mental disorders in the context of genetic risk factors by leveraging technological breakthroughs such as pooled genetic perturbation screening and neural organoid model systems [143, 149]. The molecular underpinnings of environmentally conferred risk have, on the other hand, been studied to a much smaller extent.

In this thesis, I outlined a ten-day chronic GC exposure paradigm in unguided neural organoids totalling six treatment conditions, including a 20-day wash-out period, to quantify lasting effects and the impact of an additional acute GC hit. I described how this design allowed my collaborators and myself to compile a comprehensive scRNA-seq dataset across six experimental conditions replicated in organoids derived from two cell lines. I further described that an organoid-specific quality-control procedure to remove non-viable cell states, as identified in the HNOCA (Section 3.2.3), facilitated data interpretation. While GC exposure did not significantly affect cell viability, we observed a transcriptional response in many vital neurodevelopmental genes and TFs directly following chronic GC exposure. A subset of the responding genes remained dysregulated following the 20-day wash-out period. More than half of the genes involved in this longer-term response were also responsive to the acute GC

hit, suggesting that the long-term response is still directly linked to the GC effect. Next, I outlined our findings from the trajectory analysis, focusing on the excitatory and inhibitory neuron developmental trajectories. We found that the transcriptional response to chronic GC exposure correlates with an acceleration of neuronal differentiation, significantly more so for the inhibitory neuron lineage. This priming of the inhibitory neuron lineage translated into an increased abundance of inhibitory neurons at the transcriptional and protein level. Lastly, I described how we identified PBX3 as an example of a TF closely associated with priming the inhibitory neuron lineage, which responds robustly to GC exposure and regulates several inhibitory neuron lineage drivers. Overall, we provided evidence for the selective priming of the inhibitory neuron lineage by GCs.

The balance of excitatory and inhibitory neuron generation is critical for healthy brain development, and disruptions to this balance have been associated with different psychiatric diseases [191–194]. Consequently, our finding of preferential inhibitory neuron lineage priming through GC exposure provides a potential mechanism for how a prevalent environmental risk factor might affect the onset and development of disease. Our observed phenotype aligns with the findings of a recent pooled loss-of-function screen of 36 ASD risk genes [143]. This suggests a converging effect of genetic and environmental risk factors for neurodevelopmental and psychiatric disorders on cell fate determination during neurodevelopment. Five of the 36 tested ASD risk genes were also differentially expressed after chronic GC exposure in our data, and three of them were significantly associated with an increase in the ventral-to-dorsal ratio in the genetic loss-of-function screen. Interestingly, while we observed the same inhibitory lineage priming phenotype as in the loss-of-function screen, all five overlapping risk genes were upregulated following GC exposure. This divergence in regulation of these specific risk genes in the context of a converging phenotype suggests that GC exposure affects disease risk through different molecular pathways than the established risk genes. Further investigating the exact mediators of environmental disease risk would provide further insights into the complex molecular underpinnings of psychiatric disease aetiology.

Our findings have suggested the hox-gene PBX3 a candidate TF for mediating our observed phenotype of GC-induced inhibitory neuron lineage priming. The expression of PBX3 and a significant number of its target genes was altered after GC exposure, while PBX3 expression was closely and selectively correlated with the inhibitory neuron differentiation trajectory. Increased double-positive cells for PBX3 and GAD1 after GC exposure and the analysis of the PBX3 downstream target genes in our multimodal GRN provided further evidence for a tight association of PBX3 with GC-induced inhibitory lineage priming. Identifying PBX3 as a promising mediating candidate offers a starting point for further investigations into the

molecular mechanisms underlying disease risk mediation. A genetically modified neural organoid model featuring a PBX3 gene knockout could give more insight into the role of PBX3 during brain development while highlighting affected brain regions. Alternatively, a pooled genetic perturbation screen in organoids targeting PBX3, or one of the four additional TFs identified as robustly responding to chronic GC exposure, alongside the respective target genes in the GRN, could help determine the mediating pathways for the lineage priming effect of GCs.

My collaborators and I used unguided neural organoids for the GC treatment paradigm to obtain an unbiased model system of the developing brain representing various cellular lineages. While offering a more natural and self-organised environment for neurodevelopment, unguided neural organoid differentiation protocols provide little control over the generated brain regions. Using the HNOCA (Section 3.2) as a reference, our organoids' NPCs and neurons mapped exclusively to non-telencephalic brain areas (Supplementary Fig. B.13). While this focus on the generally lesser-studied non-telencephalic brain areas enabled the identification of the hindbrain-associated TF PBX3 (Fig. 3.17) as a potential mediator of chronic GC exposure effects, future work should investigate the impact of GC exposure on further brain regions. This could be achieved using regionalised organoids of specific brain regions or unguided organoids derived from different cell lines for added variability. Additionally, assembloids have recently been used to study genetic risk for ASD and NDDs [149]. Beyond the capabilities of individual organoids, assembloids can model the interaction of multiple brain regions, making them an exciting tool for further studying environmental perturbations of brain development.

4.4. Towards a mechanistic understanding of psychiatric disease with improved model systems and computational tools

The contributions outlined in this thesis, together with recent scientific advancements, pave the way towards better models of brain development and disease on multiple levels. For example, the continuous growth of curated scRNA-seq data repositories, spearheaded by the sfaira data zoo, enables a better molecular understanding of the healthy human neuronal landscape. Extending these repositories and cellular atlases to samples from disease models and primary diseased tissue will allow us to capture better the cell types involved in the onset and development of neurodevelopmental and psychiatric disorders. At the same time, the abundance of extensive single-cell genomics data covering health and disease will enable

the training of more complex and expressive predictive models in computational biology, as shown by the recently emerging foundational modelling approaches [45].

Furthermore, the HNOCA, with its detailed mapping of organoid cell states, provides a blueprint for developing improved neural organoid differentiation protocols that produce more faithful models of the developing human brain. Creating organoids that more faithfully reflect the cellular complexity and heterogeneity underlying NDDs such as ASD and schizophrenia will, in turn, provide access to an enhanced picture of the mechanisms driving these diseases when using these improved models in genetic or chemical perturbation studies. Together with the emergence of high-throughput perturbation screens in organoids [143, 254], these developments suggest a bright future for modelling NDDs and psychiatric diseases *in vitro*.

The emergence of psychiatric diseases is a result of a complex interplay of environmental and genetic risk factors [153]. As presented in this thesis, neural organoids are a powerful platform for modelling environmental challenges during brain development. They have helped us reveal converging phenotypes of environmental and genetic risk factors for disease. Combining genetic and environmental risk factors in improved organoid models would be a promising next step towards a holistic and mechanistic understanding of the complex aetiology of neurodevelopmental and psychiatric disorders. Additionally, enhanced neural organoid models would provide a promising platform to screen different interventions and rescue mechanisms, offering hope for more effective and targeted treatment and prevention approaches going forward.

A. Supporting information

This section contains additional information to support the material provided in the main part of the thesis.

A.1. Nucleotide sequences for generating the neuron-specific fluorescent reporter cell line

Nucleotide sequences used in generating the custom iPSC reporter cell lined described in Section 2.3.3:

gRNA:

5′_GGTCGAAGACGCCATCAGCT_3′

ssODN:

5′_*TGCGCACCCCTACCAGGCAGGCTCGCTGCCTTTCCTCCCTCTTGTCTCTCCAGAGCCGG*
ATCTTCAAGGGGAGCCTCCGTGCCCCCGGCTGCTCAGTCCCTCCGGTGTGCAGGACCCCG
GAAGTCCTCCCCGCACAGCTCTCGCTTCTCTTTGCAGCCTGTTTCTGCGCCGGACCAGTCG
*AGGACTCTGGACAGTAGAGGCCCCGGGACGACCGAGCTG*ATG<u></u>GTGAGCAAGGGCGAG
GAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGG
CCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGA
CCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGA
CCACCCTGACCTACGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAAGCAG
CACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTC
TTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACA
CCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATC
CTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGA

CAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGAC
GGCAGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCC
CCGTGCTGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGAC
CCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGAT
CACTCTCGGCATGGACGAGCTGTACAAGTAACTAGAGCTCGCTGATCAGCCTCGACT
GTGCCTTCTAGTTGCCAGCCATCTGTTGTTTGCCCCTCCCCCGTGCCTTCCTTGACCCT
GGAAGGTGCCACTCCCACTGTCCTTTCCTAATAAAATGAGGAAATTGCATCGCATTG
TCTGAGTAGGTGTCATTCTATTCTGGGGGGTGGGGTGGGGCAGGACAGCAAGGGGG
AGGATTGGGAAGACAATAGCAGGCATGCTGGGGATGCGGTGGGCTCTATGGCTTCT
GAGGCGGAAAGAACCAGCTGGGGCTCTAGGGGGTATCCCC*GCGTCTTCGACCCCATC*
TTCGTCCGCAACCTCCTCGAACGCGGGAGCGGACCCCAATACCACTAACCTGCGCCCCAC
AAGTAGGTCCCGCCCCAATTTTCTATCAAATGAACTGCAGGGAAGATGGGGGCGCTGGG
ACGTCGGGAGGCTGAGCTGGCGGAAAGGGAAGGGGGAGCGCGGAGATAATGGAGGCTG
GGAAATAAATGGGGCTCTGACCCCGTCCCTGCCAGAGGTCATTCGGCTGTCAGGGACGC
*TAGGTGACTCCCAGGGCACCGGAAAGCGAGGACCACGCAAGGTCCGA*_3′

Left and right homology arms are indicated in italics. eGFP start and stop codons are underlined.

GFP Protein translation:

MVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPT
LVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDT
LVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQL
ADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITLGMDELY
K*

Primer 1:

For 5′_CACTCCCACTGTCCTTTCCTAA_3′

Rev 5′_TCCTAGCTCTTCATTCCGCC_3′

Primer 2:

For 5′_GCTTCTCTTTGCAGCCTGTTTC_3′

Rev 5′_GGGCGCAGGTTAGTGGTATT_3′

A.2. Antobodies for immunohistochemistry staining of neural organoids

The following antibodies were used for the immunohistochemistry staining described in Section 2.3.5.

Antigen	Dilution	Vendor	Catalog #
DAPI	1:1000	Sigma Aldrich	D9542
GFP	1:1000	Aves Lab	1020
PBX3	1:500	Abcam	ab52903

B. Supplementary figures

This chapter contains the supplementary figures referenced in the results chapter of this thesis.

B.1. Sfaira accelerates data and model reuse in single-cell genomics

This section contains the supplemental figures for Section 3.1 of this thesis. The figures are collectively reproduced from the supplementary material of:

Fischer, D. S.*, **Dony, L.***, König, M., Moeed, A., Zappia, L., Heumos, L., Tritschler, S., Holmberg, O., Aliee, H. & Theis, F. J. Sfaira accelerates data and model reuse in single cell genomics. *Genome Biology* **22**, 248. doi:10.1186/s13059-021-02452-6 (2021)

"*" denotes an equal contribution.

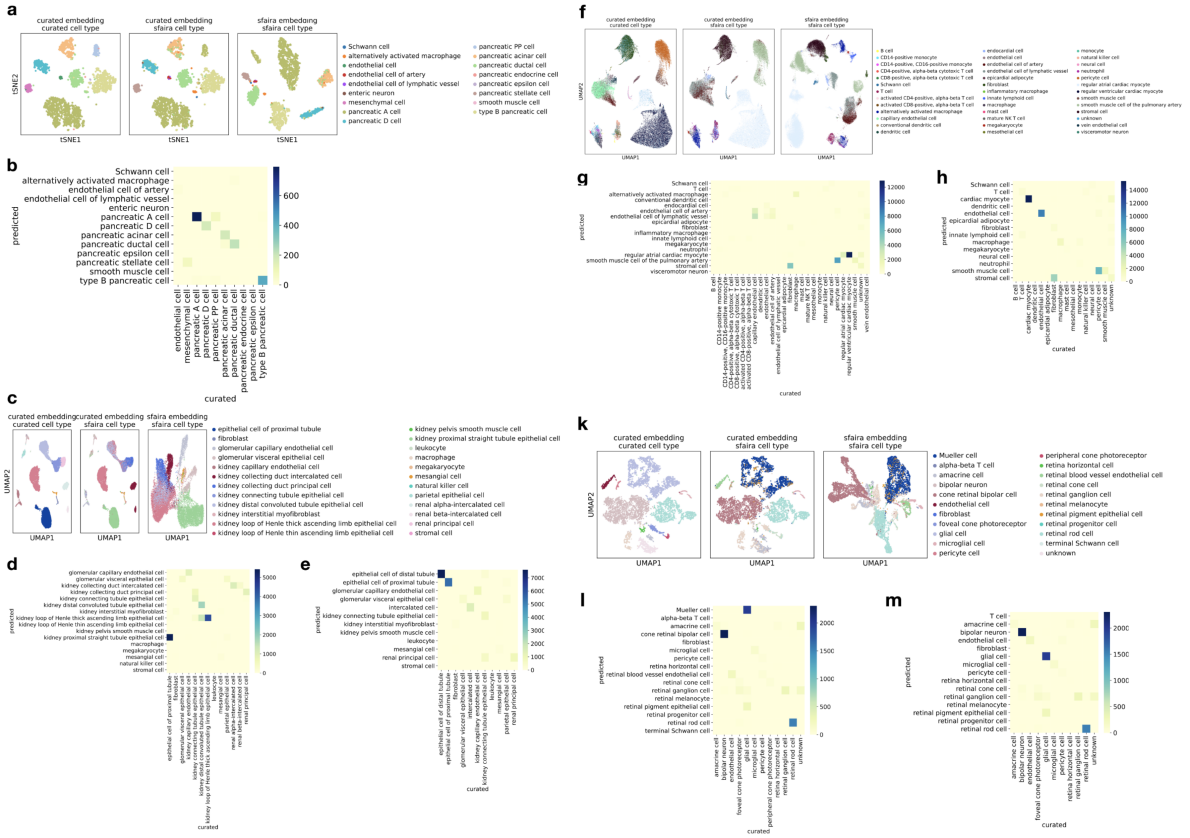

Figure B.1.: **Zero-shot analysis example cases.** Shown are summaries of four zero-shot analysis scenarios which were not included in the training and testing data of the SFAIRA model zoo presented in Section 3.1.1: A pancreas dataset (**a.**, **b.**), a kidney dataset (**c.-e.**), a heart dataset (**f.-h.**), and an eye dataset (**k.-m.**). Shown are UMAPs of either the curated embedding published by the authors or a UMAP of a SFAIRA embedding model with curated cell types or SFAIRA predicted cell types superimposed (**a.**, **c.**, **f.**, **k.**). Note that predicted and curated cell types are not always the same but often have an ontological relationship derived from annotation coarsity differences between model training data and these new datasets. The relation between these predicted and observed cell types is further characterised in confusion matrices (**b.**, **d.**, **g.**, **l.**), simplified to coarser cell ontology parent nodes in (**e.**, **h.**, **m.**). The analyses presented here were performed after the final fitting of the models, which are also available on ZENODO, and thus represent the actual behaviour of this system on new data. Minor inaccuracies in prediction are likely due to insufficient training data annotation or are related to issues in generalisation to the unseen domain effects in these test cases and can be resolved quickly in secondary analyses following this preliminary analysis. Reproduced from [258] (Licensed under the Creative Commons Attribution 4.0 International license: `https://creativecommons.org/licenses/by/4.0/`).

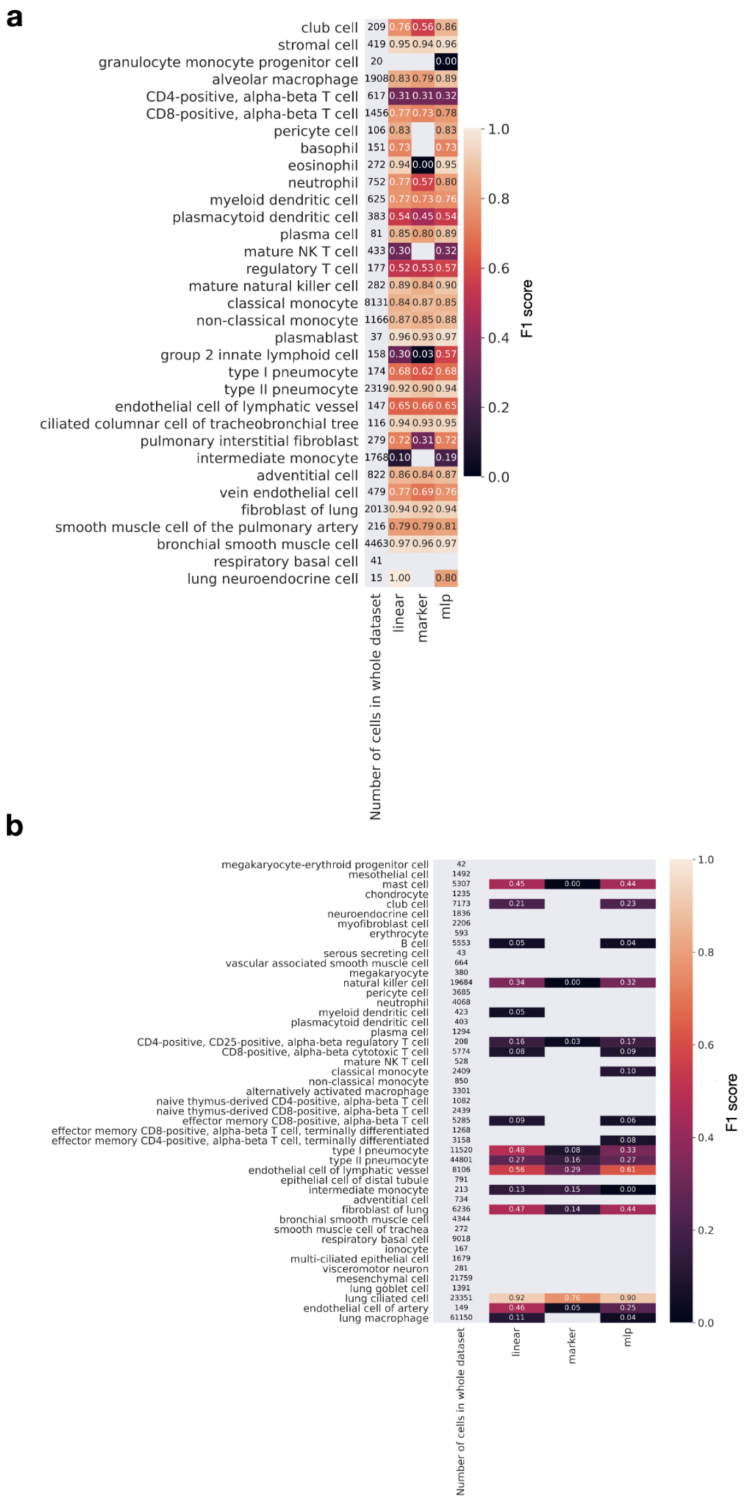

Figure B.2.: **Characterisation of the cell type classification task.** Cell type-wise accuracies for lung samples from human (**a.**) and mouse (**b.**) data. Reproduced from [258] (Licensed under the Creative Commons Attribution 4.0 International license: `https://creativecommons.org/licenses/by/4.0/`).

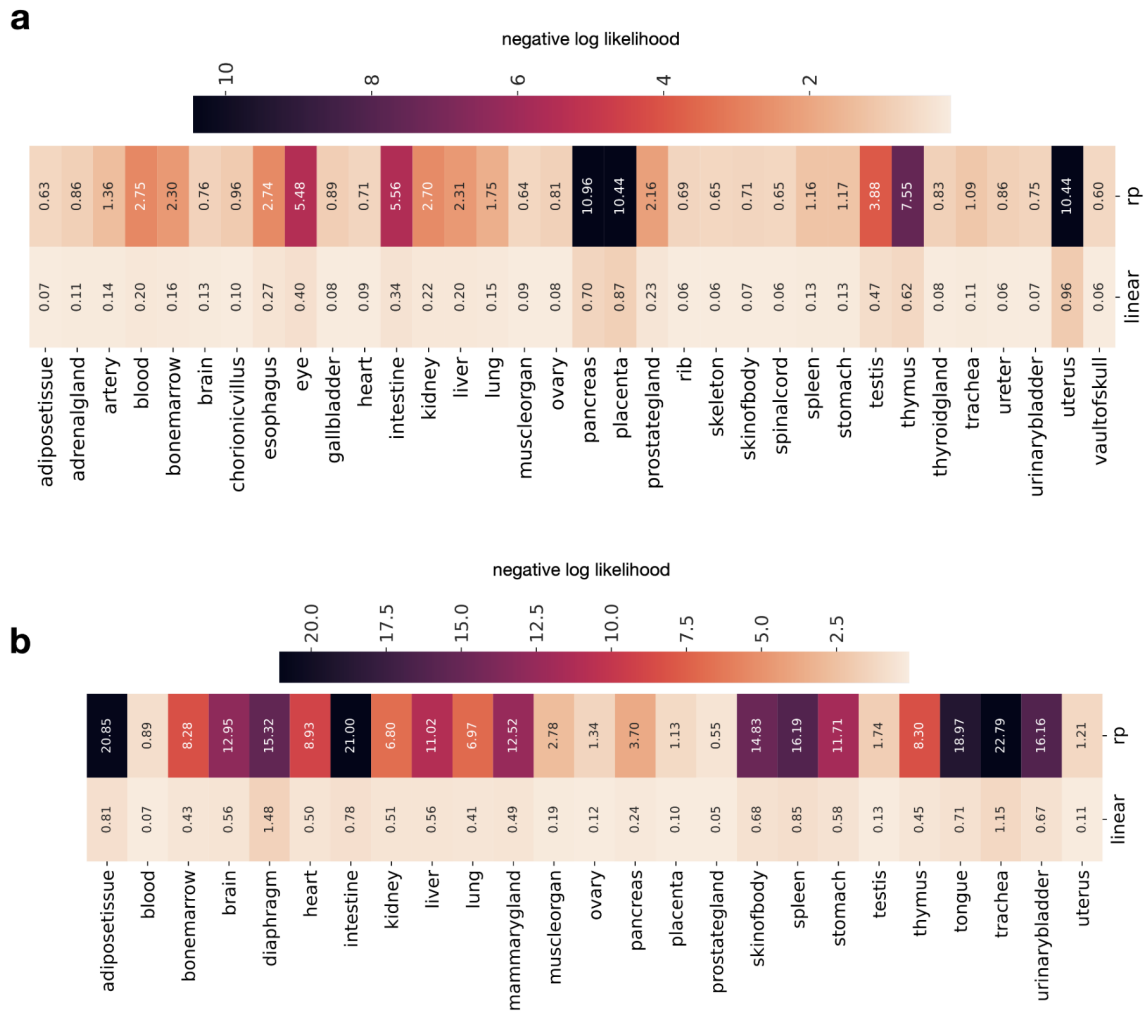

Figure B.3.: **Random projection models as baseline models for cell embedding models.** Shown is the negative log-likelihood on the held-out test data of a linear embedding model as in Fig. 3.6, and a random projection model (Section 2.1.6) for human (**a.**) and mouse (**b.**) data. Across all organs, the random projection model had significantly lower negative log-likelihoods than the linear model when comparing the mean performance across cross-validations in a one-sided paired t-test (human: p = 1.11e-05, mouse: p = 2.11e-07). Reproduced from [258] (Licensed under the Creative Commons Attribution 4.0 International license: `https://creativecommons.org/licenses/by/4.0/`).

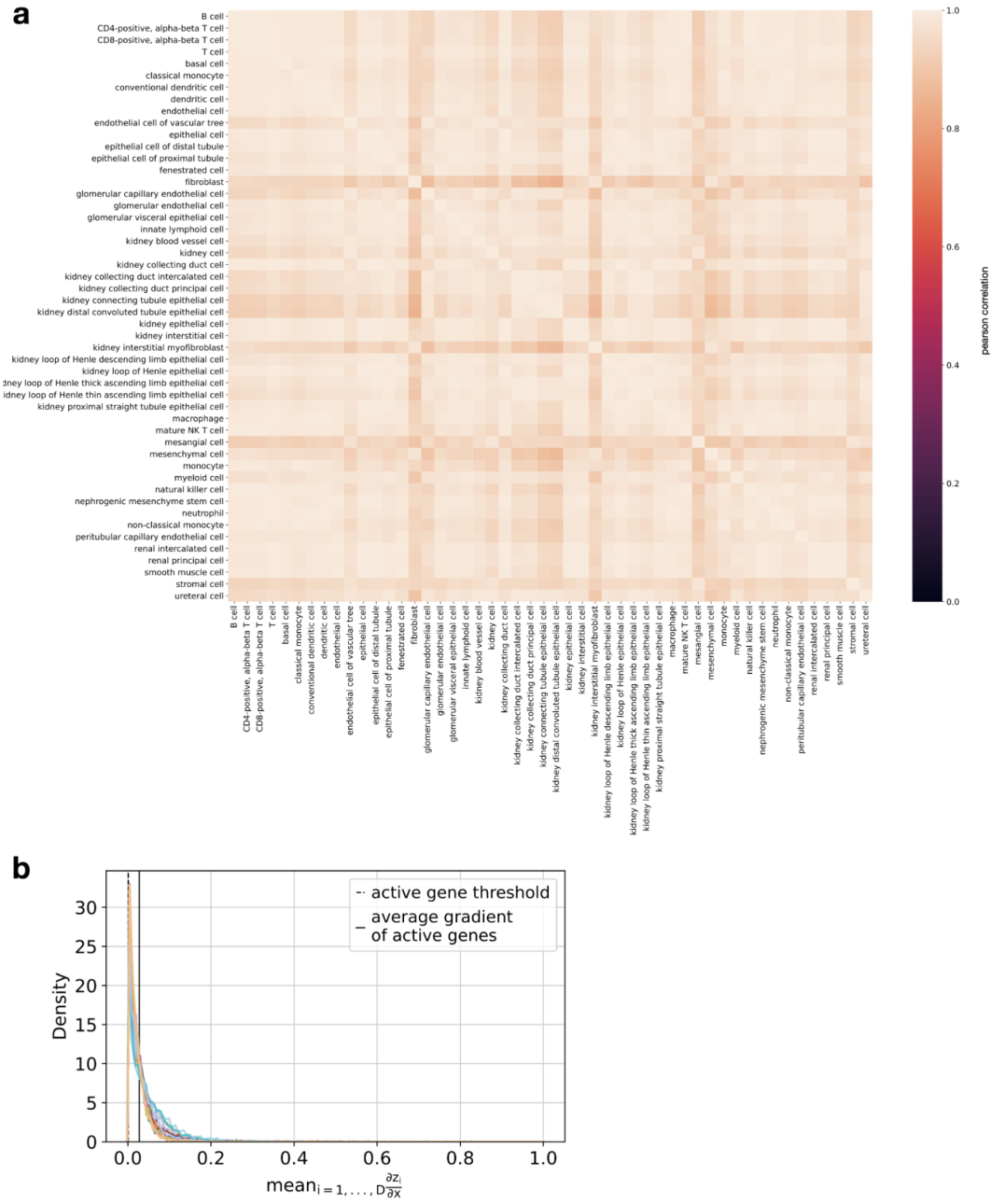

Figure B.4.: **Saliency-based interpretation of models trained on human kidney.** Shown are results for an AE. **a.** Correlation of cell type wise aggregated gradients of embedding with respect to input features. **b.** Distribution of feature-wise aggregated gradients of embedding with respect to input features by cell type (colour). Reproduced from [258] (Licensed under the Creative Commons Attribution 4.0 International license: `https://creativecommons.org/licenses/by/4.0/`).

B.2. An integrated transcriptomic cell atlas of human neural organoids

This section contains the supplemental figures for Section 3.2 of this thesis. The figures are collectively reproduced from the Extended Data Figures of:

He, Z.*, **Dony, L.***, Fleck, J. S.*, Szałata, A., Li, K. X., Sliškovic, I., Lin, H. -C., Santel, M., Atamian, A., Quadrato, G., Sun, J., Pasca, S. P., Camp, J. G., Theis, F. & Treutlein, B. An integrated transcriptomic cell atlas of human neural organoids. *In revision.* Preprint [204] doi: 10.1101/2023.10.05.561097 (2023)

"*" denotes an equal contribution.

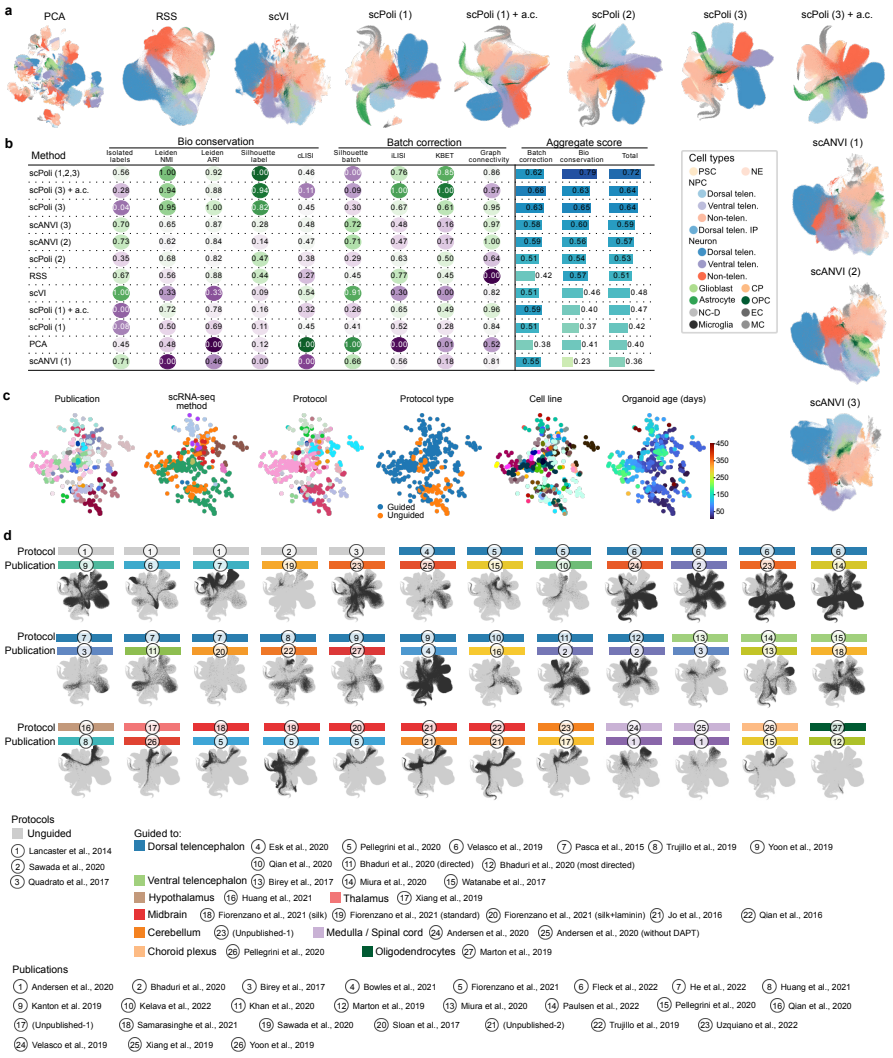

Figure B.5.: **Benchmark of data integration. a.** UMAPs of the HNOCA, either without any data integration (PCA) or with different data integration methods applied. The number in parenthesis indicates which level of RSS-based SNAPSEED annotation labels were provided as input to the model for methods which support semi-supervised data integration. Dots in all UMAP embeddings, each representing a cell, are coloured by the cell type annotation introduced in Fig. 3.8. a.c., *aggrecell* algorithm. **b.** scIB benchmarking metrics on all tested integration methods. **c.** PCA of the SCPOLI sample embeddings from the final SCPOLI integration of the HNOCA presented throughout Section 3.2, coloured by publications, scRNA-seq methods, organoid protocols, protocol types, cell lines, and sample ages. **d.** UMAPs of the HNOCA based on the final SCPOLI integration, each with one dataset highlighted. A dataset is defined as data from one protocol in one publication. The protocol and publication of each dataset are shown by the colour bar and indices on top of the UMAP. Reproduced from [204] (Licensed under the Creative Commons Attribution-NonCommercial 4.0 International license: `https://creativecommons.org/licenses/by-nc/4.0/`).

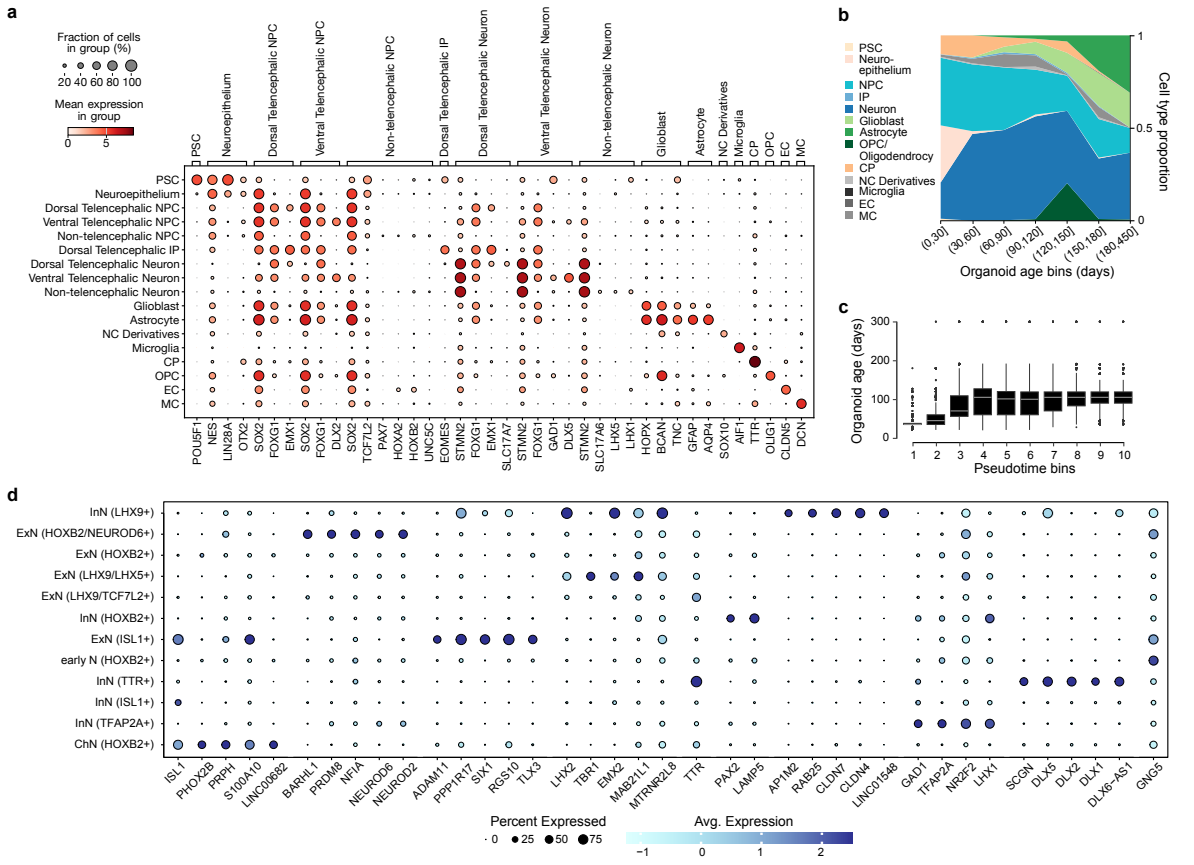

Figure B.6.: **Characterisation of the HNOCA. a.** Expression of selected marker genes used in the semi-automatic annotation of cell types for Figure 3.8. **b.** Mean cell type proportion over all datasets per organoid age bin. **c.** Distribution of sample real-time age in days over deciles of computed pseudo time. **d.** Expression of top markers in different non-telencephalic neural cell types. Markers are defined as genes with AUROC > 0.7, in-out detection rate difference > 20 %, in-out detection rate ratio > 2, and log2FC > 1.2. When more than five markers are found, only the top five (with the highest in-out detection rate ratio) are shown. Reproduced from [204] (Licensed under the Creative Commons Attribution-NonCommercial 4.0 International license: `https://creativecommons.org/licenses/by-nc/4.0/`).

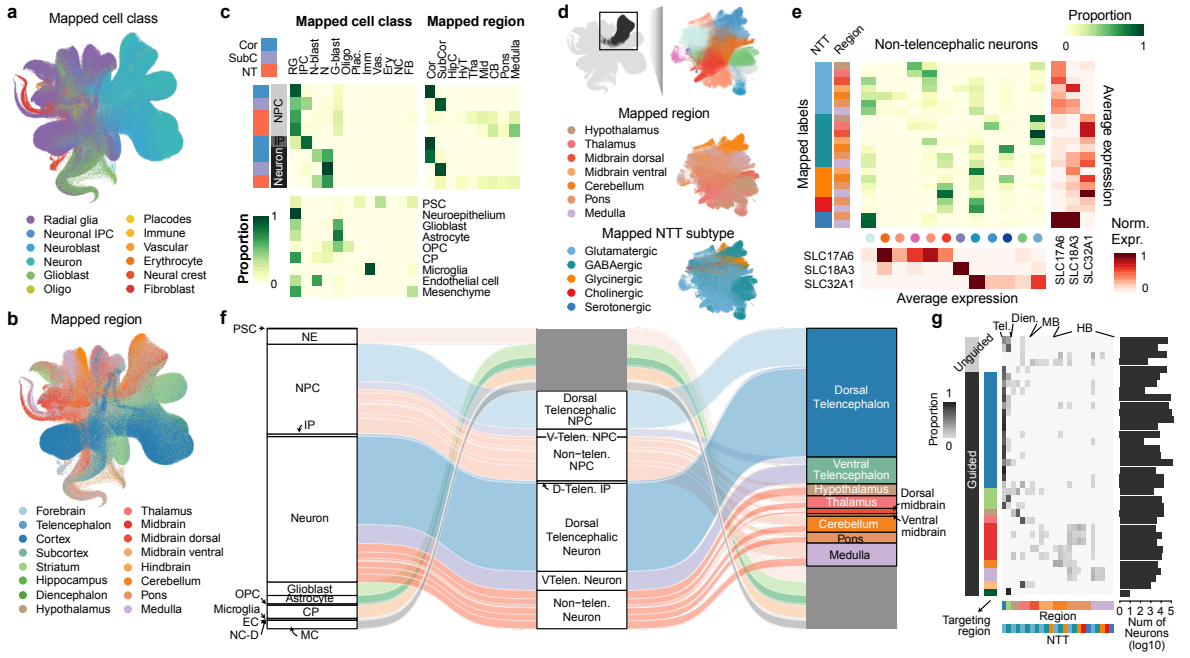

Figure B.7.: **Mapping-assisted annotation refinement of the HNOCA. a., b.** UMAP of the HNOCA coloured by the mapped cell classes (a) and brain regions (b) from the human developing brain cell atlas [46]. **c.** Comparison of the HNOCA cell type annotation with the primary reference mapping-based transferred cell class and brain region labels. The darkness of cells indicates the proportions of each HNOCA cell type being assigned to different cell class and brain region categories. Brain region labels are only shown for the HNOCA neural cell types. **d.** UMAP of non-telencephalic neurons, coloured by clusters (top), mapped brain regions (middle) and mapped NTT subtypes (bottom). **e.** Comparison of non-telencephalic neural cell types, defined as the concatenation of the mapped brain region and NTT subtype, with the clusters. The middle heatmap shows the contributions of different clusters to different neural cell types. The bar on the left shows the neural cell types; dots under the heatmap show clusters. The heatmaps at the bottom and on the right show the average expression of three NTTs SLC17A6, SLC18A3, and SLC32A1 in clusters (bottom) and neural cell types (right). **f.** Overview of the HNOCA cell type composition for the first two levels of the cell annotation (left: level-1, middle: level-2) and the refined regional annotation assisted by mapping non-telencephalic NPC and neurons to the primary reference (right). **g.** Neural cell type composition of different datasets (rows). The darkness of the heatmap shows the proportions of different neural cell types per HNOCA dataset. The bars on the left show organoid protocol types of different datasets. The bars on the bottom show neural cell types. Bars on the right show total neuron numbers across datasets. Reproduced from [204] (Licensed under the Creative Commons Attribution-NonCommercial 4.0 International license: `https://creativecommons.org/licenses/by-nc/4.0/`).

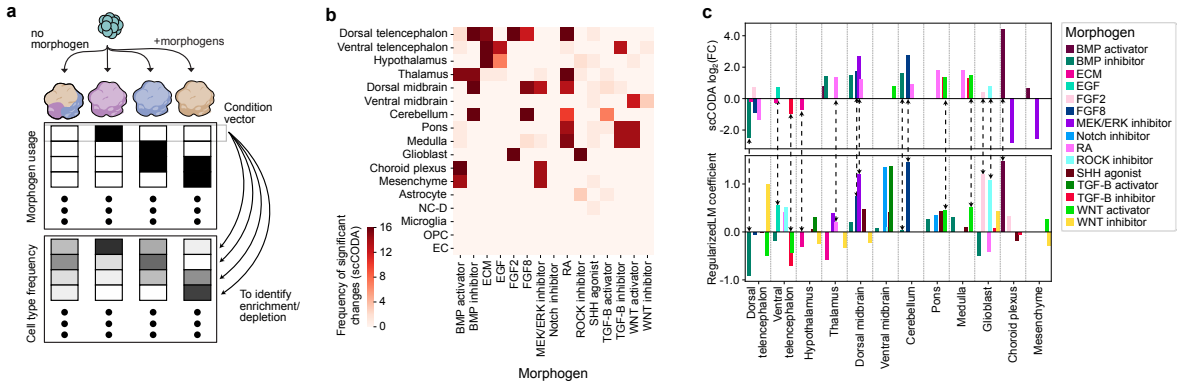

Figure B.8.: **Relationship between morphogen usage and cell type as well as regional composition. a.** Schematic of estimating cell type enrichment with different morphogen usages. **b.** This heatmap indicates in how many of the 17 scCODA iterations (using each of the 17 regional cell identities as a reference once) the respective morphogen was found to lead to compositional changes with respect to the reference regional cell identity. A morphogen effect was called significant in this consensus approach if it significantly affected cell type composition with respect to more than half of the reference cell types. **c.** Effect of different morphogens on regional organoid composition in the HNOCA. Positive values correspond to a higher abundance of cells from the indicated regional cell identity in cases where the respective morphogen was used in the differentiation protocol. Top: log2-fold-effect sizes of morphogens per regional cell identity as computed by the scCODA model. Bottom: L1-regularised linear model coefficients. The dashed arrows show consistent enrichment/depletion identified by the two methods. Reproduced from [204] (Licensed under the Creative Commons Attribution-NonCommercial 4.0 International license: `https://creativecommons.org/licenses/by-nc/4.0/`).

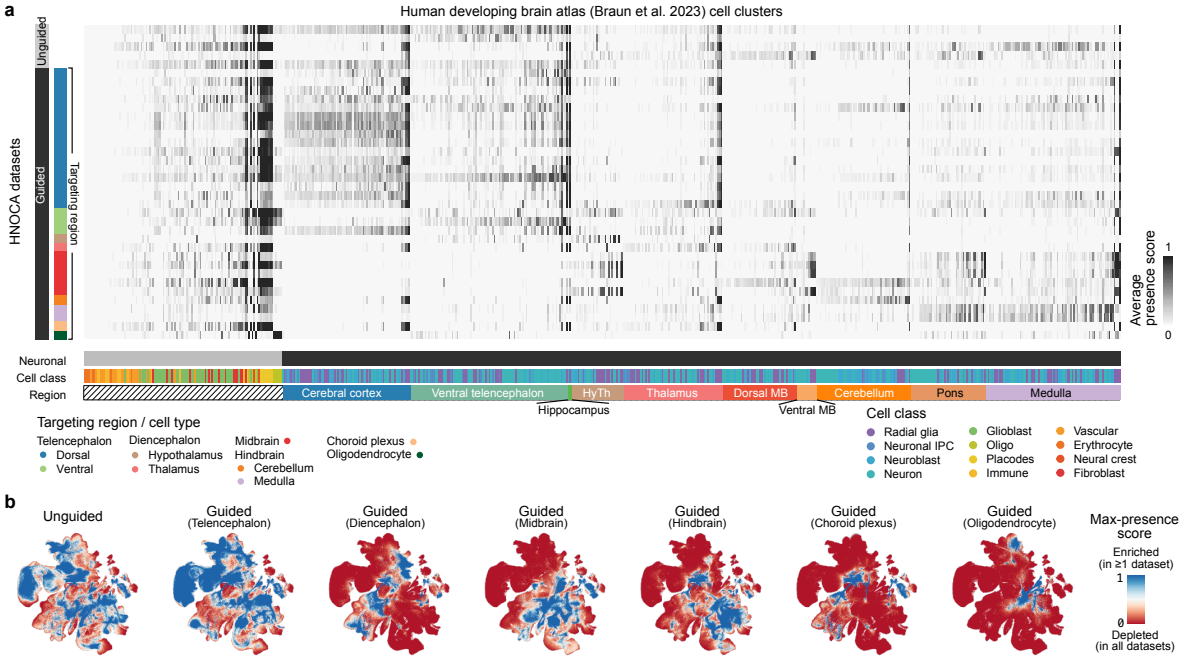

Figure B.9.: **Presence scores per HNOCA dataset. a.** Average normalised presence scores of different HNOCA datasets (rows) in different cell clusters in the primary reference of the human developing brain atlas [46] (columns). The bars on the left show organoid differentiation protocol types of the HNOCA datasets. The bars underneath show cell class and the commonest region information of the cell clusters in the primary reference. HyTh, hypothalamus; MB, midbrain. **b.** UMAP of the primary reference, coloured by the max presence scores across different HNOCA data subsets, split by organoid protocol types. A high max presence score suggests enrichment of the corresponding primary cell state in at least one HNOCA dataset among the datasets based on the specific type of organoid protocols. A low score means an under-representation of the cell state in all datasets in the subset. Reproduced from [204] (Licensed under the Creative Commons Attribution-NonCommercial 4.0 International license: `https://creativecommons.org/licenses/by-nc/4.0/`).

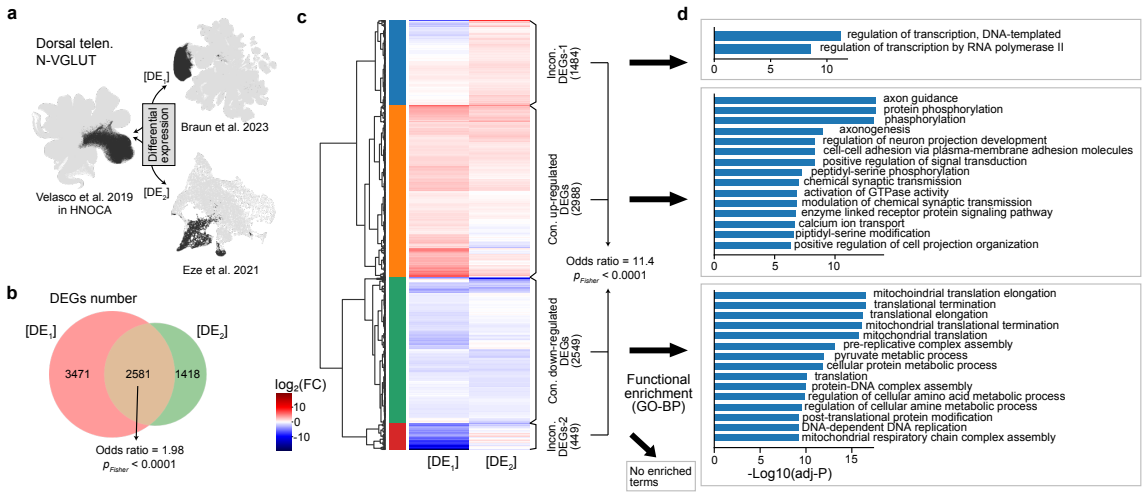

Figure B.10.: **Robustness of organoid-primary DEGs and diseases-associated organoid-primary DEGs. a.** Schematic of DE analysis between a subset of dorsal telencephalic neurons in the HNOCA [99] with counterparts in two developing human brain atlases [46, 48]. **b.** Number of DEGs between cortical neurons from *Velasco et al.* organoid data [99] and primary fetal cortical neurons from *Braun et al.* [46] (10x 3′ v2 chemistry only, [DE1]) or *Eze et al.* [48] ([DE2]) respectively. **c.** Log2FCs across all 7470 DEGs between dorsal telencephalic neurons from *Velasco et al.* [99] and either primary fetal cortical neurons from *Braun et al.* [46] (10x 3′ v2 chemistry only) or *Eze et al.* [48] The dendrogram shows the hierarchical clustering of DEGs based on their log2FC against the two primary data. **d.** Functional enrichment analysis of the four clusters of DEGs. The widths of the bars show log10-transformed adjusted p-value of Fisher's exact test. Only the first 15 terms with adjusted $p < 0.05$ are shown. Reproduced from [204] (Licensed under the Creative Commons Attribution-NonCommercial 4.0 International license: `https://creativecommons.org/licenses/by-nc/4.0/`).

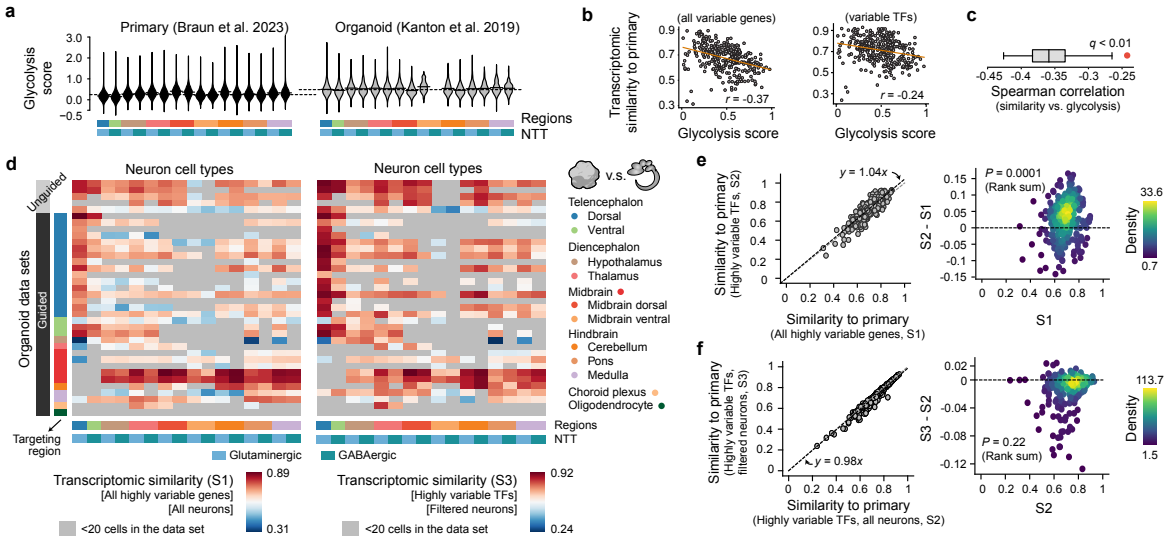

Figure B.11.: **Transcriptomic fidelity of neurons and cell stress. a.** Glycolysis scores of different neural cell types in primary (left, *Braun et al.* [46]) and a selected organoid dataset (right, *Kanton et al.* [97]). **b.** Correlation between average glycolysis scores and transcriptomic similarities (Spearman correlation) to primary counterparts. Each dot represents one neural cell type generated by one protocol. The correlation is calculated based on all HVGs (left) or TFs (right). **c.** The correlation between glycolysis scores and transcriptomic similarities to primary is significantly weaker when only TFs are considered. The box shows the distribution of correlation when a random subset of HVGs, with the same number as the variable TFs, are used. The red dot shows the correlation using variable TFs. **d.** Spearman correlation between average gene expression profiles of neural cell types in the HNOCA and those in the primary reference of human developing brain atlas [46], across all HVGs (left, S1) or variable TFs (right, S3). The average gene expression profile per neural cell type was calculated with all cells (S1) or cells with low glycolysis scores (glycolysis score < 0.6, S3). **e.** Core transcriptomic fidelity of organoid neurons (S2, shown in Fig. 3.10), which only considers TFs, is higher than the global transcriptomic fidelity (S1), which considers all the HVGs. Core transcriptomic fidelity and global transcriptomic fidelity are highly correlated (left, x-axis - S1, y-axis - S2, each dot represents one neural cell type in one HNOCA dataset), while core transcriptomic fidelity is significantly higher (right, x-axis: S1, y-axis: S2 - S1, dots are coloured by density estimated with Gaussian kernel). **f.** Core transcriptomic fidelity of organoid neurons is not improved by filtering cells based on glycolysis scores. The left panel shows core transcriptomic fidelities without (x-axis, S2) and with glycolysis score filtering (glycolysis score < 0.6, y-axis, S3). The Wilcoxon test suggests S2 and S3 are not significantly different (right, x-axis: S2, y-axis: S3 - S2, dots are coloured by density estimated with Gaussian kernel). Reproduced from [204] (Licensed under the Creative Commons Attribution-NonCommercial 4.0 International license: `https://creativecommons.org/licenses/by-nc/4.0/`).

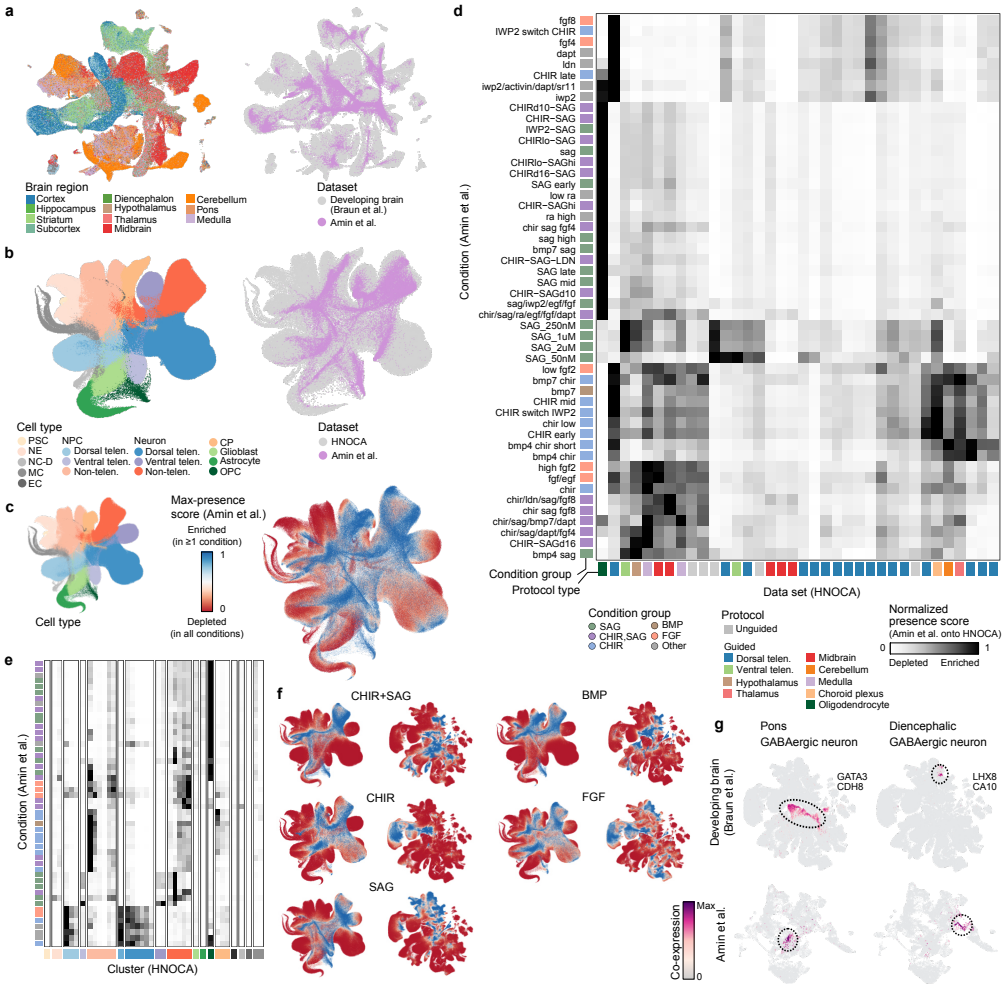

Figure B.12.: **Reference mapping of the neural organoid morphogen screen data to the HNOCA and the human developing brain atlas. a.** UMAP embedding of the human developing brain atlas and neural organoid morphogen screen [321] datasets based on the joint scANVI latent space coloured by brain region (left) and dataset (right). **b.** UMAP embedding of the HNOCA and the screen datasets based on the joint scPoli latent space coloured by annotated cell type (left) and dataset (right). **c.** scPoli UMAP embedding of the HNOCA coloured by cell type (left) and max presence score across all datasets (right). **d.** Heatmap showing min-max scaled average presence scores of each condition in the screen dataset in the HNOCA datasets. **e.** Heatmap showing min-max scaled average presence scores of each condition in the screen dataset in each *Leiden* cluster in the HNOCA, ordered by annotated cell type. **f.** UMAP embeddings of the HNOCA (left) and the human developing brain atlas (right) coloured by presence scores for each condition group in the screen dataset. **g.** UMAP embeddings of the human developing brain atlas (upper) and screen dataset (lower) coloured by coexpression scores of clusters with gained coverage in the screen dataset. Reproduced from [204] (Licensed under the Creative Commons Attribution-NonCommercial 4.0 International license: `https://creativecommons.org/licenses/by-nc/4.0/`).

B.3. Chronic exposure to glucocorticoids amplifies inhibitory neuron cell fate during human neurodevelopment in organoids

This section contains the supplemental figures for Section 3.3 of this thesis. The figures are collectively reproduced from the Supplementary Figures of:

Dony, L., Krontira, A. C., Kaspar, L., Ahmad, R., Demirel, I. S., Grochowicz, M., Schäfer, T., Begum, F., Sportelli, V., Raimundo, C., Koedel, M., Labeur, M., Cappello, S., Theis, F. J., Cruceanu, C.*, Binder, E. B.* Chronic exposure to glucocorticoids amplifies inhibitory neuron cell fate during human neurodevelopment in organoids. *In review.* Preprint [205] doi: 10.1101/2024.01.21.576532 (2024)

"*" denotes an equal contribution.

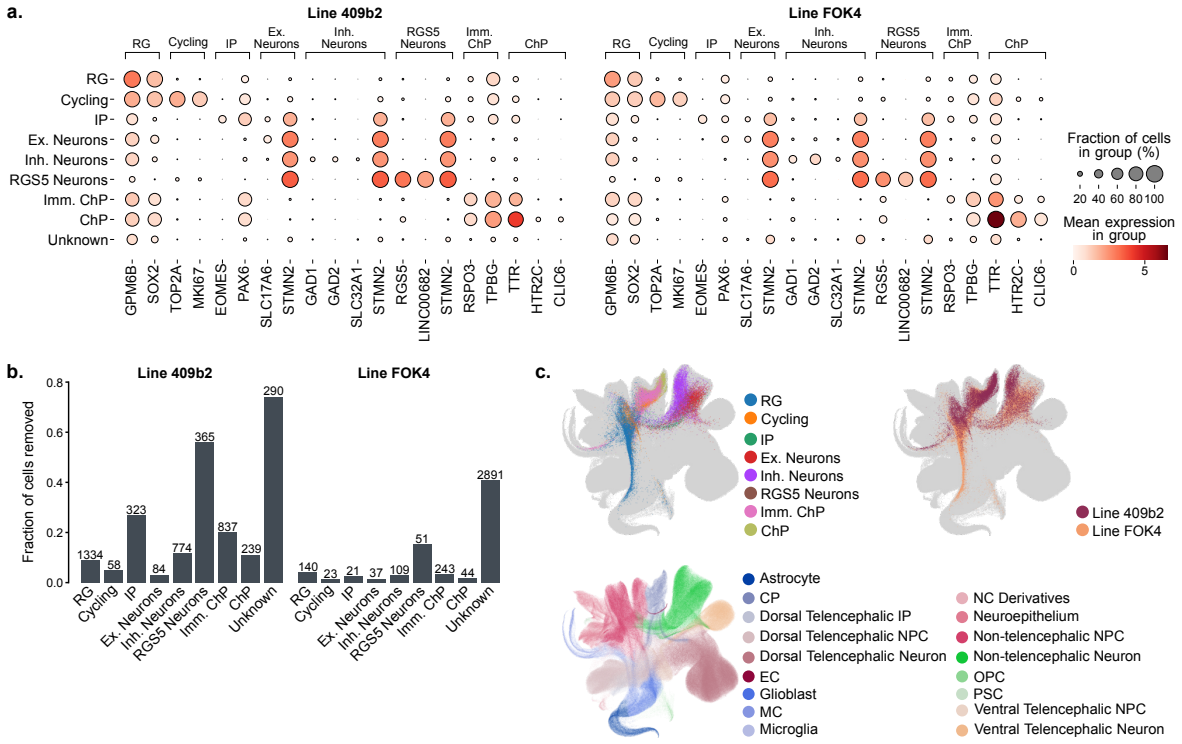

Figure B.13.: **Chronic GC exposure in neural organoids does not induce significant metabolic stress in cells.** **a.** Selected marker gene expression per cell type and cell line. **b.** Fraction of non-viable cells per cell type and cell line. The absolute number of non-viable cells per cluster is displayed above each bar. The *Unknown* clusters were removed from the datasets in their entirety (394 cells in Line 409b2; 7039 cells in Line FOK4). **c.** Top: cells from this study projected to the HNOCA [204]. Cells are coloured by their origin dataset (left) and cell types assigned in this study (right). Bottom: HNOCA cell type labels. CP, choroid plexus; NPC, neural progenitor cell; EC, endothelial cell; MC, mesenchymal cell; NC, neural crest; OPC, oligodendrocyte progenitor cell; PSC, pluripotent stem cell. Reproduced from [205] (Licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license: `https://creativecommons.org/licenses/by-nc-nd/4.0/`).

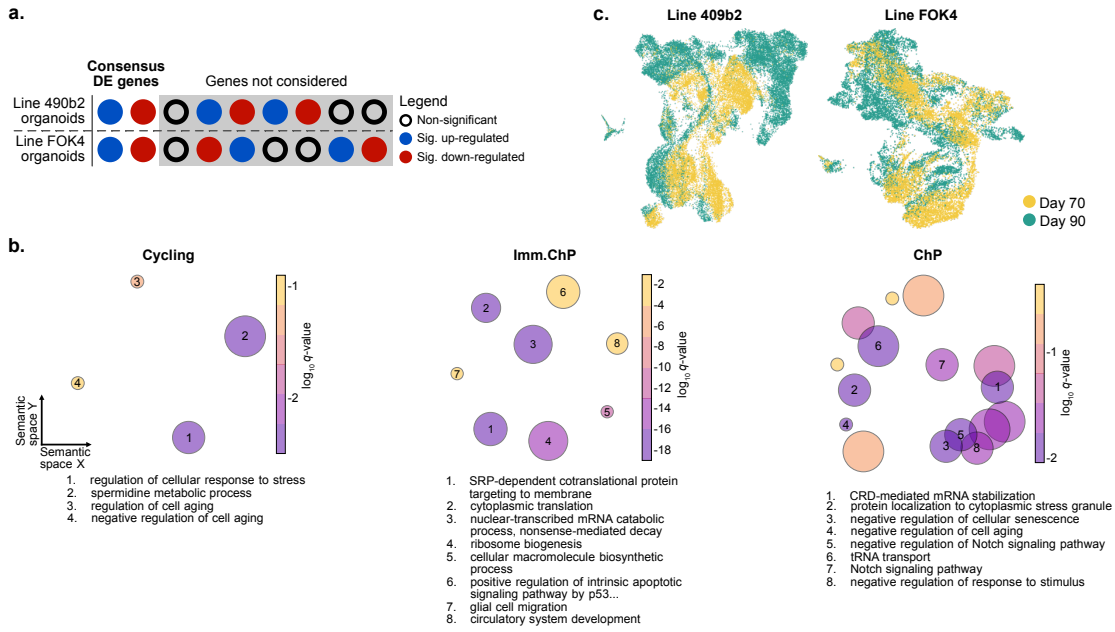

Figure B.14.: **Transcriptional response following chronic GC treatment in organoids includes key neurodevelopmental genes. a.** Filtering scheme used to identify consensus DEGs between organoids from the two genetic backgrounds. **b.** Grouped semantic space representation of the Gene Ontology Biological Process enrichment analysis for the three cell types with the least detected DEGs. The size of the circles corresponds to the number of terms in the cluster; their colour corresponds to the $\log_{10}$(q-value) of the representative term for each cluster. The integers within the circles enumerate the eight most significant clusters, and their representative term is written out in the legend below each plot. **c.** UMAP embedding of Line 409b2 and Line FOK4 data coloured by organoid age. Reproduced from [205] (Licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license: `https://creativecommons.org/licenses/by-nc-nd/4.0/`).

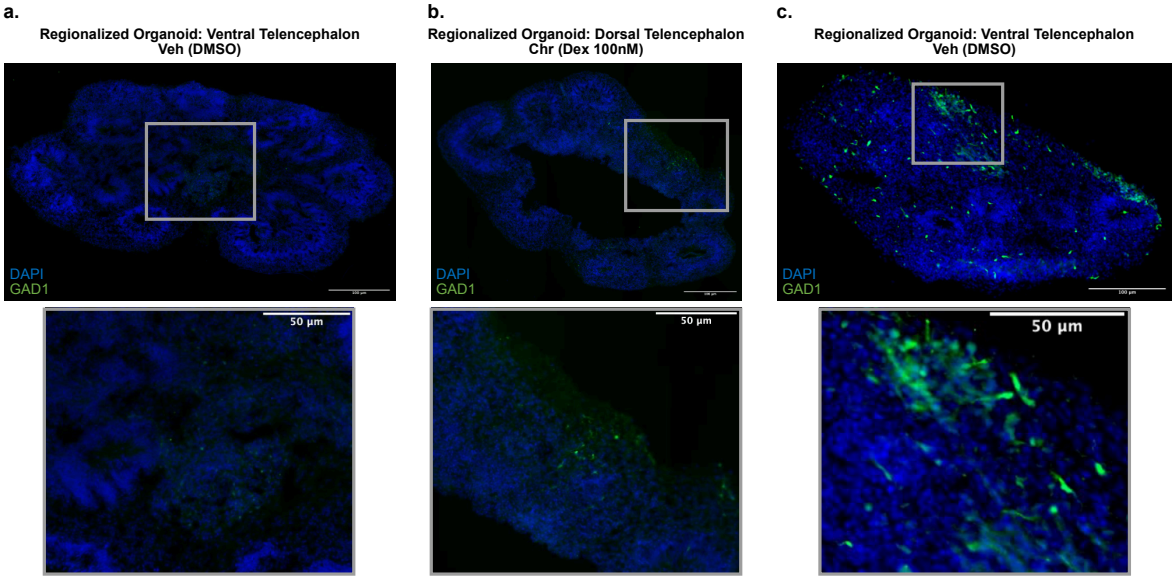

Figure B.15.: **GC exposure results in an increased abundance of inhibitory neurons in organoids. a.** Representative image of whole slice dorsalised Line 409b2 (GFP-GAD1) control organoids at day 70 in culture (*Veh* condition). Only very few GAD1-positive cells are visible. Lower panel: zoomed-in inserts. DMSO, dimethyl sulfoxide; Dex, dexamethasone. **b.** Representative image of whole slice dorsalised Line 409b2 (GFP-GAD1) organoids at day 70 in culture, following ten days of chronic treatment with GCs (100nM dexamethasone; *Chr* condition). Only very few GAD1-positive cells are visible, but slightly more than in the dorsalised control organoids. **c.** Representative image of whole slice ventralised Line 409b2 (GFP-GAD1) control organoids at day 70 in culture (*Veh* condition). A larger number of GAD1-positive cells than in the dorsalised organoids indicates a successful ventralisation. Reproduced from [205] (Licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license: `https://creativecommons.org/licenses/by-nc-nd/4.0/`).

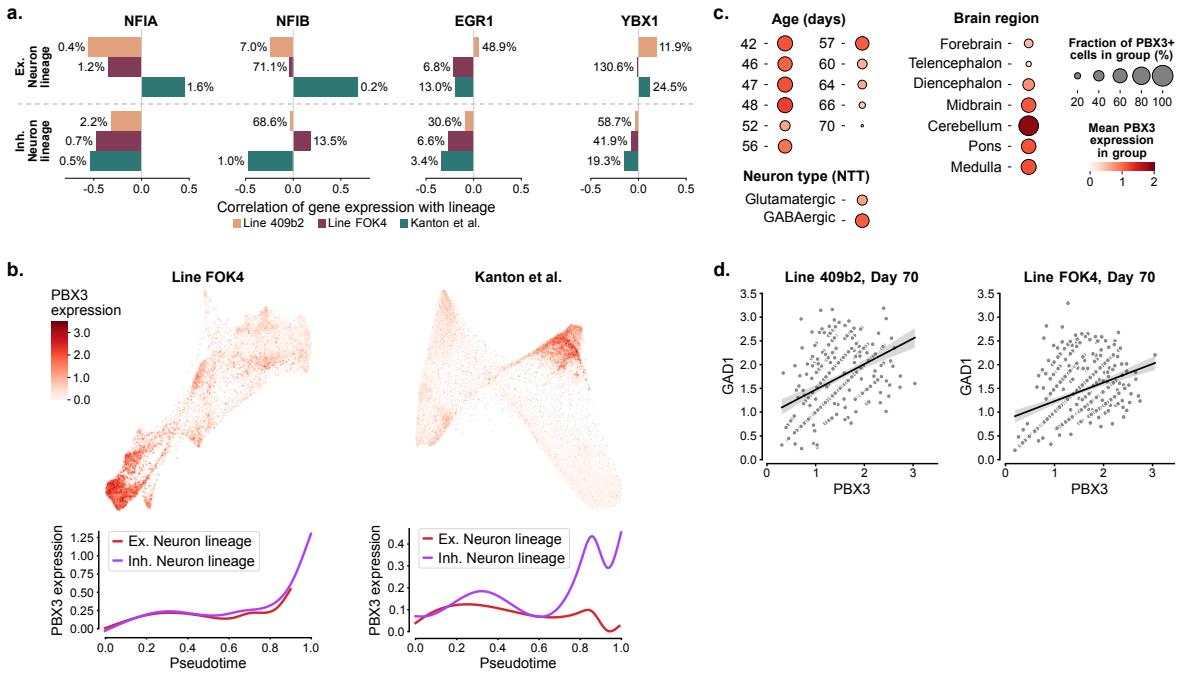

Figure B.16.: **PBX3 regulation through chronic GC exposure supports inhibitory neuron priming. a.** Correlation of NFIA, NFIB, YBX1, and EGR1 expression with lineage probability across the excitatory and inhibitory neuronal lineages in all three datasets. The percentile of each gene among all significant driver genes ranked by driver strength is shown on the side of every bar. **b.** Expression of PBX3 in Line FOK4 and the validation data [97] on a force-directed graph embedding (top). Expression patterns of PBX3 across pseudo time for each of the three lineage endpoints in Line FOK4 and the validation data [97] (bottom). **c.** Expression of PBX3 in the fetal brain atlas [46] neurons and NPCs across age (in days), dissected brain region, and NTT expression (left to right). **d.** Expression of GAD1 and PBX3 in day-70 double-positive cells with fitted linear regression line. Left: Line 409b2. Right: Line FOK4. Reproduced from [205] (Licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license: `https://creativecommons.org/licenses/by-nc-nd/4.0/`).

Acronyms

AE autoencoder. 5, 7, 31, 33, 70, 73, 76–78, 125

ASD autism spectum disorder. 1, 16, 17, 21, 27, 96, 98–100, 106, 115–117

AUROC area under the receiver operating characteristic curve. 37, 128

cDNA complementary DNA. 2, 3

ChP choroid plexus. 14, 15, 54, 57, 82, 93, 94, 97

CVAE conditional variational autoencoder. 37, 38, 41, 42

DE differential expression. 5, 7, 24, 45, 46, 62, 79, 87, 93, 94, 96–100, 104, 105, 107, 111, 112, 132

DEG differentially expressed gene. 5, 45, 55–57, 62, 84, 87, 94, 96–101, 107, 132, 137

eGFP enhanced green fluorescent protein. 49, 50, 119

FAIR Findable, Accessible, Interoperable, and Reusable. 8, 109

FDR false discovery rate. 46, 55–57, 62, 94

GC glucocorticoid. ii, iii, 1, 17–21, 23, 27, 28, 49, 50, 59, 63, 90–99, 101–109, 114–116, 136–139

GFP green fluorescent protein. 51, 99, 103, 119, 120, 138

GR glucocorticoid receptor. 18, 94, 95

GRN gene regulatory network. 5, 60–62, 75, 91, 104, 106, 107, 114–116

gRNA guide RNA. 49, 50, 118

HCA Human Cell Atlas. 9, 10, 113

HNOCA Human Neural Organoid Cell Atlas. iii, v, 23–25, 27, 28, 36, 37, 41, 43–47, 54, 55, 63, 79–89, 91, 94, 108, 111–114, 116, 117, 127–134, 136

HPA hypothalamic-pituitary-adrenal. 18–21

HVG highly variable gene. 40, 46, 53, 54, 56, 57, 59, 133

ImmChP immature choroid plexus. 93, 96, 97

IP intermediate progenitor. 12, 13, 83, 85, 93, 94

iPSC induced pluripotent stem cell. 12, 14, 48, 49, 91, 97, 118

kNN k-nearest neighbour. 7, 36, 39, 41, 43, 45, 53–57, 59, 60, 83

log2FC log2 fold-change. 46, 55, 84, 87, 94, 96, 100, 101, 105, 107, 128, 132

mRNA messenger RNA. 2, 4, 8, 25, 30

NDD neurodevelopmental disorder. 16, 116, 117

NPC neural progenitor cell. 12, 13, 15, 43, 57, 81–83, 85, 93, 94, 116, 129, 136, 139

NTT neurotransmitter transporter. 40, 43, 58, 83, 85, 129, 139

OPC oligodendrocyte precursor cell. 13, 81, 82

PBMC peripheral blood mononuclear cell. 48, 70, 71, 73

PCA principle component analysis. 30, 36, 39, 45, 53, 54, 56, 57, 62, 68, 71, 73, 75, 76, 127

PCR polymerase chain reaction. 2, 50

QC quality control. 5, 36, 40, 53, 59, 80, 91

RG radial glia. 11–13, 15, 16, 57, 82, 83, 93, 94, 96, 97

RSS reference similarity spectrum. 36, 38, 39, 80, 127

scATAC-seq single-cell assay for transposase-accessible chromatin using sequencing. 25, 52, 59–62, 90, 91, 104, 107

scRNA-seq single-cell ribonucleic acid sequencing. 1–10, 13, 22–25, 27, 28, 30, 31, 33, 35, 37, 39, 40, 44, 47, 52, 54, 59–65, 67, 68, 70–73, 75, 79–81, 84, 88–91, 97, 103, 104, 107–112, 114, 116, 127

SELU Scaled Exponential Linear Unit. 33

sGC synthetic glucocorticoid. iii, 20, 91

ssODN single-stranded oligo DNA nucleotide. 50, 118

TF transcription factor. 18, 46, 56, 60–62, 86, 87, 91, 94, 96, 98, 100, 101, 103–105, 107, 114–116, 133

uDEG ubiquitous differentially expressed gene. 84, 86, 87

UMAP Uniform Manifold Approximation and Projection. 30, 36, 45, 54–57, 59, 60, 70, 71, 76, 81, 82, 85, 89, 95, 98, 107, 122, 127, 129, 131, 134, 137

UMI unique molecular identifier. 3, 30, 36, 53, 69

VAE variational autoencoder. 27, 33, 37, 70, 73, 76

Bibliography

1. Willsey, H. R., Willsey, A. J., Wang, B. & State, M. W. Genomics, convergent neuroscience and progress in understanding autism spectrum disorder. *Nature Reviews Neuroscience* **23,** 323–341. doi:10.1038/s41583-022-00576-7 (2022).
2. Kato, M. & Dobyns, W. B. Lissencephaly and the molecular basis of neuronal migration. *Human Molecular Genetics* **12,** R89–96. doi:10.1093/hmg/ddg086 (Issue suppl_1 2003).
3. Eberwine, J., Yeh, H., Miyashiro, K. *et al.* Analysis of gene expression in single live neurons. *Proceedings of the National Academy of Sciences* **89,** 3010–3014. doi:10.1073/pnas.89.7.3010 (1992).
4. Lambolez, B., Audinat, E., Bochet, P., Crépel, F. & Rossier, J. AMPA receptor subunits expressed by single purkinje cells. *Neuron* **9,** 247–258. doi:10.1016/0896-6273(92)90164-9 (1992).
5. Crino, P. B., Trojanowski, J. Q., Dichter, M. A. & Eberwine, J. Embryonic neuronal markers in tuberous sclerosis: Single-cell molecular pathology. *Proceedings of the National Academy of Sciences* **93,** 14152–14157. doi:10.1073/pnas.93.24.14152 (1996).
6. Zhong, J. F., Chen, Y., Marcus, J. S. *et al.* A microfluidic processor for gene expression profiling of single human embryonic stem cells. *Lab Chip* **8,** 68–74. doi:10.1039/B712116D (2008).
7. Marcus, J. S., Anderson, W. F. & Quake, S. R. Microfluidic Single-Cell mRNA Isolation and Analysis. *Analytical Chemistry* **78,** 3084–3089. doi:10.1021/ac0519460 (2006).
8. Tang, F., Barbacioru, C., Wang, Y. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* **6,** 377–382. doi:10.1038/nmeth.1315 (2009).
9. Cao, J., O'Day, D. R., Pliner, H. A. *et al.* A human cell atlas of fetal gene expression. *Science* **370,** eaba7721. doi:10.1126/science.aba7721 (2020).
10. Siletti, K., Hodge, R., Mossi Albiach, A. *et al.* Transcriptomic diversity of cell types across the adult human brain. *Science* **382,** eadd7046. doi:10.1126/science.add7046 (2023).

11. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols* **13,** 599–604. doi:10.1038/nprot.2017.149 (2018).

12. Saiki, R. K., Gelfand, D. H., Stoffel, S. *et al.* Primer-Directed Enzymatic Amplification of DNA with a Thermostable DNA Polymerase. *Science* **239,** 487–491. doi:10.1126/science.2448875 (1988).

13. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Reports* **2,** 666–673. doi:10.1016/j.celrep.2012.08.003 (2012).

14. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell* **58,** 610–620. doi:10.1016/j.molcel.2015.04.005 (2015).

15. Gupta, I., Collier, P. G., Haase, B. *et al.* Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nature Biotechnology* **36,** 1197–1202. doi:10.1038/nbt.4259 (2018).

16. Islam, S., Zeisel, A., Joost, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods* **11,** 163–166. doi:10.1038/nmeth.2772 (2014).

17. Aird, D., Ross, M. G., Chen, W.-S. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology* **12,** R18. doi:10.1186/gb-2011-12-2-r18 (2011).

18. Kivioja, T., Vähärautio, A., Karlsson, K. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods* **9,** 72–74. doi:10.1038/nmeth.1778 (2012).

19. Hahaut, V., Pavlinic, D., Carbone, W. *et al.* Fast and highly sensitive full-length single-cell RNA sequencing using FLASH-seq. *Nature Biotechnology* **40,** 1447–1451. doi:10.1038/s41587-022-01312-3 (2022).

20. Hagemann-Jensen, M., Ziegenhain, C., Chen, P. *et al.* Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nature Biotechnology* **38,** 708–714. doi:10.1038/s41587-020-0497-0 (2020).

21. Hashimshony, T., Senderovich, N., Avital, G. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biology* **17,** 77. doi:10.1186/s13059-016-0938-8 (2016).

22. Ramsköld, D., Luo, S., Wang, Y.-C. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology* **30,** 777–782. doi:10.1038/nbt.2282 (2012).

23. Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343,** 776–779. doi:10.1126/science.1247651 (2014).

24. Goldstein, L. D., Chen, Y.-J. J., Dunne, J. *et al.* Massively parallel nanowell-based single-cell gene expression profiling. *BMC Genomics* **18,** 519. doi:10.1186/s12864-017-3893-1 (2017).

25. Picelli, S., Björklund, Å. K., Faridani, O. R. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods* **10,** 1096–1098. doi:10.1038/nmeth.2639 (2013).

26. Klein, A. M., Mazutis, L., Akartuna, I. *et al.* Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161,** 1187–1201. doi:10.1016/j.cell.2015.04.044 (2015).

27. Macosko, E. Z., Basu, A., Satija, R. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161,** 1202–1214. doi:10.1016/j.cell.2015.05.002 (2015).

28. Zheng, G. X. Y., Terry, J. M., Belgrader, P. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8,** 14049. doi:10.1038/ncomms14049 (2017).

29. Martin, B. K., Qiu, C., Nichols, E. *et al.* Optimized single-nucleus transcriptional profiling by combinatorial indexing. *Nature Protocols* **18,** 188–207. doi:10.1038/s41596-022-00752-0 (2023).

30. Trapnell, C., Cacchiarelli, D., Grimsby, J. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology* **32,** 381–386. doi:10.1038/nbt.2859 (2014).

31. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine* **9,** 75. doi:10.1186/s13073-017-0467-4 (2017).

32. Ziegenhain, C., Vieth, B., Parekh, S. *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell* **65,** 631–643.e4. doi:10.1016/j.molcel.2017.01.023 (2017).

33. Kaminow, B., Yunusov, D. & Dobin, A. *STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data* bioRxiv:2021.05.05.442755. 2021. doi:10.1101/2021.05.05.442755.

34. Melsted, P., Booeshaghi, A. S., Liu, L. *et al.* Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nature Biotechnology* **39,** 813–818. doi:10.1038/s41587-021-00870-2 (2021).

35. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* **19,** 15. doi:10.1186/s13059-017-1382-0 (2018).

36. Hao, Y., Stuart, T., Kowalski, M. H. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology.* doi:10.1038/s41587-023-01767-y (2023).

37. Zappia, L. & Theis, F. J. Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape. *Genome Biology* **22,** 301. doi:10.1186/s13059-021-02519-4 (2021).

38. Heumos, L., Schaar, A. C., Lance, C. *et al.* Best practices for single-cell analysis across modalities. *Nature Reviews Genetics* **24,** 550–572. doi:10.1038/s41576-023-00586-w (2023).

39. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology* **15,** e8746. doi:10.15252/msb.20188746 (2019).

40. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature Methods* **15,** 1053–1058. doi:10.1038/s41592-018-0229-2 (2018).

41. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications* **10,** 390. doi:10.1038/s41467-018-07931-2 (2019).

42. Lotfollahi, M., Naghipourfar, M., Luecken, M. D. *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology* **40,** 121–130. doi:10.1038/s41587-021-01001-7 (2022).

43. Theodoris, C. V., Xiao, L., Chopra, A. *et al.* Transfer learning enables predictions in network biology. *Nature* **618,** 616–624. doi:10.1038/s41586-023-06139-9 (2023).

44. Cui, H., Wang, C., Maan, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods.* doi:10.1038/s41592-024-02201-0 (2024).

45. Consens, M. E., Dufault, C., Wainberg, M. *et al. To Transformers and Beyond: Large Language Models for the Genome* 2023. arXiv: 2311.07621.

46. Braun, E., Danan-Gotthold, M., Borm, L. E. *et al.* Comprehensive cell atlas of the first-trimester developing human brain. *Science* **382,** eadf1226. doi:10.1126/science.adf1226 (2023).

47. Polioudakis, D., De La Torre-Ubieta, L., Langerman, J. *et al.* A Single-Cell Transcriptomic Atlas of Human Neocortical Development during Mid-gestation. *Neuron* **103,** 785–801.e8. doi:10.1016/j.neuron.2019.06.011 (2019).

48. 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. *Nature Neuroscience* **24,** 584–594. doi:10.1038/s41593-020-00794-1 (2021).

49. Ma, Z., Zhang, X., Zhong, W. *et al.* Deciphering early human pancreas development at the single-cell level. *Nature Communications* **14,** 5354. doi:10.1038/s41467-023-40893-8 (2023).

50. Olaniru, O. E., Kadolsky, U., Kannambath, S. *et al.* Single-cell transcriptomic and spatial landscapes of the developing human pancreas. *Cell Metabolism* **35,** 184–199.e5. doi:10.1016/j.cmet.2022.11.009 (2023).

51. Cao, S., Feng, H., Yi, H. *et al.* Single-cell RNA sequencing reveals the developmental program underlying proximal–distal patterning of the human lung at the embryonic stage. *Cell Research* **33,** 421–433. doi:10.1038/s41422-023-00802-6 (2023).

52. Sountoulidis, A., Marco Salas, S., Braun, E. *et al.* A topographic atlas defines developmental origins of cell heterogeneity in the human embryonic lung. *Nature Cell Biology* **25,** 351–365. doi:10.1038/s41556-022-01064-x (2023).

53. Mižíková, I. & Thébaud, B. Looking at the developing lung in single-cell resolution. *American Journal of Physiology-Lung Cellular and Molecular Physiology* **320,** L680–L687. doi:10.1152/ajplung.00385.2020 (2021).

54. Moignard, V., Woodhouse, S., Haghverdi, L. *et al.* Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nature Biotechnology* **33,** 269–276. doi:10.1038/nbt.3154 (2015).

55. Ranzoni, A. M., Tangherloni, A., Berest, I. *et al.* Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Human Developmental Hematopoiesis. *Cell Stem Cell* **28,** 472–487.e7. doi:10.1016/j.stem.2020.11.015 (2021).

56. Tritschler, S., Büttner, M., Fischer, D. S. *et al.* Concepts and limitations for learning developmental trajectories from single cell genomics. *Development* **146,** dev170506. doi:10.1242/dev.170506 (2019).

57. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods* **13,** 845–848. doi:10.1038/nmeth.3971 (2016).

58. Setty, M., Kiseliovas, V., Levine, J. *et al.* Characterization of cell fate probabilities in single-cell data with Palantir. *Nature Biotechnology* **37,** 451–460. doi:10.1038/s41587-019-0068-4 (2019).

59. Wolf, F. A., Hamey, F. K., Plass, M. *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology* **20,** 59. doi:10.1186/s13059-019-1663-x (2019).

60. La Manno, G., Soldatov, R., Zeisel, A. *et al.* RNA velocity of single cells. *Nature* **560,** 494–498. doi:10.1038/s41586-018-0414-6 (2018).

61. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology* **38,** 1408–1414. doi:10.1038/s41587-020-0591-3 (2020).

62. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nature Biotechnology* **37,** 547–554. doi:10.1038/s41587-019-0071-9 (2019).

63. Lange, M., Bergen, V., Klein, M. *et al.* CellRank for directed single-cell fate mapping. *Nature Methods* **19,** 159–170. doi:10.1038/s41592-021-01346-6 (2022).

64. Herman, J. S., Sagar & Grün, D. FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nature Methods* **15,** 379–386. doi:10.1038/nmeth.4662 (2018).

65. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3,** 160018. doi:10.1038/sdata.2016.18 (2016).

66. Diehl, A. D., Meehan, T. F., Bradford, Y. M. *et al.* The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *Journal of Biomedical Semantics* **7,** 44. doi:10.1186/s13326-016-0088-7 (2016).

67. Malone, J., Holloway, E., Adamusiak, T. *et al.* Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* **26,** 1112–1118. doi:10.1093/bioinformatics/btq099 (2010).

68. Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an integrative multi-species anatomy ontology. *Genome Biology* **13,** R5. doi:10.1186/gb-2012-13-1-r5 (2012).

69. Osumi-Sutherland, D., Xu, C., Keays, M. *et al.* Cell type ontologies of the Human Cell Atlas. *Nature Cell Biology* **23,** 1129–1135. doi:10.1038/s41556-021-00787-7 (2021).

70. CZI Single-Cell Biology Program, Abdulla, S., Aevermann, B. *et al. CZ CELL×GENE Discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data* bioRxiv:2023.10.30.563174. 2023. doi:10.1101/2023.10.30.563174.

71. Moreno, P., Fexova, S., George, N. *et al.* Expression Atlas update: gene and protein expression in multiple species. *Nucleic Acids Research* **50,** D129–D140. doi:10.1093/nar/gkab1030 (D1 2022).

72. Corridoni, D., Antanaviciute, A., Gupta, T. *et al.* Single-cell atlas of colonic CD8+ T cells in ulcerative colitis. *Nature Medicine* **26,** 1480–1490. doi:10.1038/s41591-020-1003-4 (2020).

73. Emont, M. P., Jacobs, C., Essene, A. L. *et al.* A single-cell atlas of human and mouse white adipose tissue. *Nature* **603,** 926–933. doi:10.1038/s41586-022-04518-2 (2022).

74. Travaglini, K. J., Nabhan, A. N., Penland, L. *et al.* A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587,** 619–625. doi:10.1038/s41586-020-2922-4 (2020).

75. Winkler, E. A., Kim, C. N., Ross, J. M. *et al.* A single-cell atlas of the normal and malformed human brain vasculature. *Science* **375,** eabi7377. doi:10.1126/science.abi7377 (2022).

76. Regev, A., Teichmann, S. A., Lander, E. S. *et al.* The Human Cell Atlas. *eLife* **6,** e27041. doi:10.7554/eLife.27041 (2017).

77. Sikkema, L., Ramírez-Suástegui, C., Strobl, D. C. *et al.* An integrated cell atlas of the lung in health and disease. *Nature Medicine* **29,** 1563–1577. doi:10.1038/s41591-023-02327-2 (2023).

78. Luecken, M. D., Büttner, M., Chaichoompu, K. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods* **19,** 41–50. doi:10.1038/s41592-021-01336-8 (2022).

79. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nature Biotechnology* **37,** 685–691. doi:10.1038/s41587-019-0113-3 (2019).

80. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nature Methods* **16,** 715–721. doi:10.1038/s41592-019-0494-8 (2019).

81. Xu, C., Lopez, R., Mehlman, E. *et al.* Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular Systems Biology* **17,** e9620. doi:10.15252/msb.20209620 (2021).

82. De Donno, C., Hediyeh-Zadeh, S., Moinfar, A. A. *et al.* Population-level integration of single-cell datasets enables multi-scale analysis across samples. *Nature Methods* **20,** 1683–1692. doi:10.1038/s41592-023-02035-2 (2023).

83. Hansen, D. V., Lui, J. H., Parker, P. R. L. & Kriegstein, A. R. Neurogenic radial glia in the outer subventricular zone of human neocortex. *Nature* **464,** 554–561. doi:10.1038/nature08845 (2010).

84. Wang, X., Tsai, J.-W., LaMonica, B. & Kriegstein, A. R. A new subtype of progenitor cell in the mouse embryonic neocortex. *Nature Neuroscience* **14,** 555–561. doi:10.1038/nn.2807 (2011).

85. Hill, R. S. & Walsh, C. A. Molecular insights into human brain evolution. *Nature* **437,** 64–67. doi:10.1038/nature04103 (2005).

86. Hodge, R. D., Bakken, T. E., Miller, J. A. *et al.* Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573,** 61–68. doi:10.1038/s41586-019-1506-7 (2019).

87. Zhao, X. & Bhattacharyya, A. Human Models Are Needed for Studying Human Neurodevelopmental Disorders. *The American Journal of Human Genetics* **103,** 829–857. doi:10.1016/j.ajhg.2018.10.009 (2018).

88. Karzbrun, E., Kshirsagar, A., Cohen, S. R., Hanna, J. H. & Reiner, O. Human brain organoids on a chip reveal the physics of folding. *Nature Physics* **14,** 515–522. doi:10.1038/s41567-018-0046-7 (2018).

89. Riccobelli, D. & Bevilacqua, G. Surface tension controls the onset of gyrification in brain organoids. *Journal of the Mechanics and Physics of Solids* **134,** 103745. doi:10.1016/j.jmps.2019.103745 (2020).

90. Lancaster, M. A., Renner, M., Martin, C.-A. *et al.* Cerebral organoids model human brain development and microcephaly. *Nature* **501,** 373–379. doi:10.1038/nature12517 (2013).

91. Adams, J. W., Negraes, P. D., Truong, J. *et al.* Impact of alcohol exposure on neural development and network formation in human cortical organoids. *Molecular Psychiatry* **28,** 1571–1584. doi:10.1038/s41380-022-01862-7 (2023).

92. Betts, J. G., Young, K. A., Wise, J. A. *et al.* in *Anatomy and Physiology* 13.1 The Embryologic Perspective (OpenStax, Houston, Texas, 2013).

93. Silbereis, J. C., Pochareddy, S., Zhu, Y., Li, M. & Sestan, N. The Cellular and Molecular Landscapes of the Developing Human Central Nervous System. *Neuron* **89,** 248–268. doi:10.1016/j.neuron.2015.12.008 (2016).

94. Gordon, A., Yoon, S.-J., Tran, S. S. *et al.* Long-term maturation of human cortical organoids matches key early postnatal transitions. *Nature Neuroscience* **24,** 331–342. doi:10.1038/s41593-021-00802-y (2021).

95. Kadoshima, T., Sakaguchi, H., Nakano, T. *et al.* Self-organization of axial polarity, inside-out layer pattern, and species-specific progenitor dynamics in human ES cell–derived neocortex. *Proceedings of the National Academy of Sciences* **110,** 20284–20289. doi:10.1073/pnas.1315710110 (2013).

96. Qian, X., Su, Y., Adam, C. D. *et al.* Sliced Human Cortical Organoids for Modeling Distinct Cortical Layer Formation. *Cell Stem Cell* **26,** 766–781.e9. doi:10.1016/j.stem.2020.02.002 (2020).

97. Kanton, S., Boyle, M. J., He, Z. *et al.* Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature* **574,** 418–422. doi:10.1038/s41586-019-1654-9 (2019).

98. Camp, J. G., Badsha, F., Florio, M. *et al.* Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proceedings of the National Academy of Sciences* **112,** 15672–15677. doi:10.1073/pnas.1520760112 (2015).

99. Velasco, S., Kedaigle, A. J., Simmons, S. K. *et al.* Individual brain organoids reproducibly form cell diversity of the human cerebral cortex. *Nature* **570,** 523–527. doi:10.1038/s41586-019-1289-x (2019).

100. Giandomenico, S. L., Mierau, S. B., Gibbons, G. M. *et al.* Cerebral organoids at the air–liquid interface generate diverse nerve tracts with functional output. *Nature Neuroscience* **22,** 669–679. doi:10.1038/s41593-019-0350-2 (2019).

101. Sloan, S. A., Darmanis, S., Huber, N. *et al.* Human Astrocyte Maturation Captured in 3D Cerebral Cortical Spheroids Derived from Pluripotent Stem Cells. *Neuron* **95,** 779–790.e6. doi:10.1016/j.neuron.2017.07.035 (2017).

102. Tanaka, Y., Cakir, B., Xiang, Y., Sullivan, G. J. & Park, I.-H. Synthetic Analyses of Single-Cell Transcriptomes from Multiple Brain Organoids and Fetal Brain. *Cell Reports* **30,** 1682–1689.e3. doi:10.1016/j.celrep.2020.01.038 (2020).

103. Bhaduri, A., Andrews, M. G., Mancia Leon, W. *et al.* Cell stress in cortical organoids impairs molecular subtype specification. *Nature* **578,** 142–148. doi:10.1038/s41586-020-1962-0 (2020).

104. Vértesy, Á., Eichmüller, O. L., Naas, J. *et al.* Gruffi: an algorithm for computational removal of stressed cells from brain organoid transcriptomic datasets. *The EMBO Journal* **41,** e111118. doi:10.15252/embj.2022111118 (2022).

105. Cowan, M. & Petri, W. A. Microglia: Immune Regulators of Neurodevelopment. *Frontiers in Immunology* **9,** 2576. doi:10.3389/fimmu.2018.02576 (2018).

106. Paolicelli, R. C., Bolasco, G., Pagani, F. *et al.* Synaptic Pruning by Microglia Is Necessary for Normal Brain Development. *Science* **333,** 1456–1458. doi:10.1126/science.1202529 (2011).

107. Matsui, T. K., Tsuru, Y., Hasegawa, K. & Kuwako, K.-i. Vascularization of Human Brain Organoids. *Stem Cells* **39,** 1017–1024. doi:10.1002/stem.3368 (2021).

108. Cakir, B., Xiang, Y., Tanaka, Y. *et al.* Engineering of human brain organoids with a functional vascular-like system. *Nature Methods* **16,** 1169–1175. doi:10.1038/s41592-019-0586-5 (2019).

109. Sun, X.-Y., Ju, X.-C., Li, Y. *et al.* Generation of vascularized brain organoids to study neurovascular interactions. *eLife* **11,** e76707. doi:10.7554/eLife.76707 (2022).

110. Xu, R., Boreland, A. J., Li, X. *et al.* Developing human pluripotent stem cell-based cerebral organoids with a controllable microglia ratio for modeling brain development and pathology. *Stem Cell Reports* **16,** 1923–1937. doi:10.1016/j.stemcr.2021.06.011 (2021).

111. Zhang, W., Jiang, J., Xu, Z. *et al.* Microglia-containing human brain organoids for the study of brain development and pathology. *Molecular Psychiatry* **28,** 96–107. doi:10.1038/s41380-022-01892-1 (2023).

112. Renner, M., Lancaster, M. A., Bian, S. *et al.* Self-organized developmental patterning and differentiation in cerebral organoids. *The EMBO Journal* **36,** 1316–1329. doi:10.15252/embj.201694700 (2017).

113. Lancaster, M. A. & Knoblich, J. A. Generation of cerebral organoids from human pluripotent stem cells. *Nature Protocols* **9,** 2329–2340. doi:10.1038/nprot.2014.158 (2014).

114. Quadrato, G., Nguyen, T., Macosko, E. Z. *et al.* Cell diversity and network dynamics in photosensitive human brain organoids. *Nature* **545,** 48–53. doi:10.1038/nature22047 (2017).

115. Paşca, A. M., Sloan, S. A., Clarke, L. E. *et al.* Functional cortical neurons and astrocytes from human pluripotent stem cells in 3D culture. *Nature Methods* **12,** 671–678. doi:10.1038/nmeth.3415 (2015).

116. Trujillo, C. A., Gao, R., Negraes, P. D. *et al.* Complex Oscillatory Waves Emerging from Cortical Organoids Model Early Human Brain Network Development. *Cell Stem Cell* **25,** 558–569.e7. doi:10.1016/j.stem.2019.08.002 (2019).

117. Yoon, S.-J., Elahi, L. S., Paşca, A. M. *et al.* Reliability of human cortical organoid generation. *Nature Methods* **16,** 75–78. doi:10.1038/s41592-018-0255-0 (2019).

118. Birey, F., Andersen, J., Makinson, C. D. *et al.* Assembly of functionally integrated human forebrain spheroids. *Nature* **545,** 54–59. doi:10.1038/nature22330 (2017).

119. Miura, Y., Li, M.-Y., Birey, F. *et al.* Generation of human striatal organoids and cortico-striatal assembloids from human pluripotent stem cells. *Nature Biotechnology* **38,** 1421–1430. doi:10.1038/s41587-020-00763-w (2020).

120. Xiang, Y., Tanaka, Y., Cakir, B. *et al.* hESC-Derived Thalamic Organoids Form Reciprocal Projections When Fused with Cortical Organoids. *Cell Stem Cell* **24,** 487–497.e7. doi:10.1016/j.stem.2018.12.015 (2019).

121. Huang, W.-K., Wong, S. Z. H., Pather, S. R. *et al.* Generation of hypothalamic arcuate organoids from human induced pluripotent stem cells. *Cell Stem Cell* **28,** 1657–1670.e10. doi:10.1016/j.stem.2021.04.006 (2021).

122. Fiorenzano, A., Sozzi, E., Birtele, M. *et al.* Single-cell transcriptomics captures features of human midbrain development and dopamine neuron diversity in brain organoids. *Nature Communications* **12,** 7302. doi:10.1038/s41467-021-27464-5 (2021).

123. Andersen, J., Revah, O., Miura, Y. *et al.* Generation of Functional Human 3D Cortico-Motor Assembloids. *Cell* **183,** 1913–1929.e26. doi:10.1016/j.cell.2020.11.017 (2020).

124. Pellegrini, L., Bonfio, C., Chadwick, J. *et al.* Human CNS barrier-forming organoids with cerebrospinal fluid production. *Science* **369,** eaaz5626. doi:10.1126/science.aaz5626 (2020).

125. Marton, R. M., Miura, Y., Sloan, S. A. *et al.* Differentiation and maturation of oligodendrocytes in human three-dimensional neural cultures. *Nature Neuroscience* **22,** 484–491. doi:10.1038/s41593-018-0316-9 (2019).

126. Bardy, C., Van Den Hurk, M., Eames, T. *et al.* Neuronal medium that supports basic synaptic functions and activity of human neurons in vitro. *Proceedings of the National Academy of Sciences* **112,** E2725–E2734. doi:10.1073/pnas.1504393112 (2015).

127. Kanton, S. & Paşca, S. P. Human assembloids. *Development* **149,** dev201120. doi:10.1242/dev.201120 (2022).

128. Xiang, Y., Tanaka, Y., Patterson, B. *et al.* Fusion of Regionally Specified hPSC-Derived Organoids Models Human Brain Development and Interneuron Migration. *Cell Stem Cell* **21,** 383–398.e7. doi:10.1016/j.stem.2017.07.007 (2017).

129. Revah, O., Gore, F., Kelley, K. W. *et al.* Maturation and circuit integration of transplanted human cortical organoids. *Nature* **610,** 319–326. doi:10.1038/s41586-022-05277-w (2022).

130. Klaus, J., Kanton, S., Kyrousi, C. *et al.* Altered neuronal migratory trajectories in human cerebral organoids derived from individuals with neuronal heterotopia. *Nature Medicine* **25,** 561–568. doi:10.1038/s41591-019-0371-0 (2019).

131. Jabali, A., Hoffrichter, A., Uzquiano, A. *et al.* Human cerebral organoids reveal progenitor pathology in EML1-linked cortical malformation. *EMBO reports* **23,** e54027. doi:10.15252/embr.202154027 (2022).

132. Bressan, C., Snapyan, M., Snapyan, M. *et al.* Metformin rescues migratory deficits of cells derived from patients with periventricular heterotopia. *EMBO Molecular Medicine* **15,** e16908. doi:10.15252/emmm.202216908 (2023).

133. Samarasinghe, R. A., Miranda, O. A., Buth, J. E. *et al.* Identification of neural oscillations and epileptiform changes in human brain organoids. *Nature Neuroscience* **24,** 1488–1500. doi:10.1038/s41593-021-00906-5 (2021).

134. Yildirim, M., Delepine, C., Feldman, D. *et al.* Label-free three-photon imaging of intact human cerebral organoids for tracking early events in brain development and deficits in Rett syndrome. *eLife* **11,** e78079. doi:10.7554/eLife.78079 (2022).

135. Bershteyn, M., Nowakowski, T. J., Pollen, A. A. *et al.* Human iPSC-Derived Cerebral Organoids Model Cellular Features of Lissencephaly and Reveal Prolonged Mitosis of Outer Radial Glia. *Cell Stem Cell* **20,** 435–449.e4. doi:10.1016/j.stem.2016.12.007 (2017).

136. Vanova, T., Sedmik, J., Raska, J. *et al.* Cerebral organoids derived from patients with Alzheimer's disease with PSEN1/2 mutations have defective tissue patterning and altered development. *Cell Reports* **42,** 113310. doi:10.1016/j.celrep.2023.113310 (2023).

137. Kim, H., Park, H. J., Choi, H. *et al.* Modeling G2019S-LRRK2 Sporadic Parkinson's Disease in 3D Midbrain Organoids. *Stem Cell Reports* **12,** 518–531. doi:10.1016/j.stemcr.2019.01.020 (2019).

138. Garcez, P. P., Loiola, E. C., Madeiro Da Costa, R. *et al.* Zika virus impairs growth in human neurospheres and brain organoids. *Science* **352,** 816–818. doi:10.1126/science.aaf6116 (2016).

139. Qian, X., Nguyen, H. N., Jacob, F., Song, H. & Ming, G.-l. Using brain organoids to understand Zika virus-induced microcephaly. *Development* **144,** 952–957. doi:10.1242/dev.140707 (2017).

140. Sun, G., Chiuppesi, F., Chen, X. *et al.* Modeling Human Cytomegalovirus-Induced Microcephaly in Human iPSC-Derived Brain Organoids. *Cell Reports Medicine* **1,** 100002. doi:10.1016/j.xcrm.2020.100002 (2020).

141. Notaras, M., Lodhi, A., Dündar, F. *et al.* Schizophrenia is defined by cell-specific neuropathology and multiple neurodevelopmental mechanisms in patient-derived cerebral organoids. *Molecular Psychiatry* **27,** 1416–1434. doi:10.1038/s41380-021-01316-6 (2022).

142. Paulsen, B., Velasco, S., Kedaigle, A. J. *et al.* Autism genes converge on asynchronous development of shared neuron classes. *Nature* **602,** 268–273. doi:10.1038/s41586-021-04358-6 (2022).

143. Li, C., Fleck, J. S., Martins-Costa, C. *et al.* Single-cell brain organoid screening identifies developmental defects in autism. *Nature* **621,** 373–380. doi:10.1038/s41586-023-06473-y (2023).

144. Arzua, T., Yan, Y., Jiang, C. *et al.* Modeling alcohol-induced neurotoxicity using human induced pluripotent stem cell-derived three-dimensional cerebral organoids. *Translational Psychiatry* **10,** 347. doi:10.1038/s41398-020-01029-4 (2020).

145. Dang, J., Tiwari, S. K., Agrawal, K. *et al.* Glial cell diversity and methamphetamine-induced neuroinflammation in human cerebral organoids. *Molecular Psychiatry* **26,** 1194–1207. doi:10.1038/s41380-020-0676-x (2021).

146. Monk, C., Lugo-Candelas, C. & Trumpff, C. Prenatal Developmental Origins of Future Psychopathology: Mechanisms and Pathways. *Annual Review of Clinical Psychology* **15,** 317–344. doi:10.1146/annurev-clinpsy-050718-095539 (2019).

147. Keshavan, M., Lizano, P. & Prasad, K. The synaptic pruning hypothesis of schizophrenia: promises and challenges. *World Psychiatry* **19,** 110–111. doi:10.1002/wps.20725 (2020).

148. Sellgren, C. M., Gracias, J., Watmuff, B. *et al.* Increased synapse elimination by microglia in schizophrenia patient-derived models of synaptic pruning. *Nature Neuroscience* **22,** 374–385. doi:10.1038/s41593-018-0334-7 (2019).

149. Meng, X., Yao, D., Imaizumi, K. *et al.* Assembloid CRISPR screens reveal impact of disease genes in human neurodevelopment. *Nature* **622,** 359–366. doi:10.1038/s41586-023-06564-w (2023).

150. Geschwind, D. H. & State, M. W. Gene hunting in autism spectrum disorder: on the path to precision medicine. *The Lancet Neurology* **14,** 1109–1120. doi:10.1016/S1474-4422(15)00044-7 (2015).

151. Levitt, P. & Campbell, D. B. The genetic and neurobiologic compass points toward common signaling dysfunctions in autism spectrum disorders. *Journal of Clinical Investigation* **119,** 747–754. doi:10.1172/JCI37934 (2009).

152. Sullivan, P. F., Daly, M. J. & O'Donovan, M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature Reviews Genetics* **13,** 537–551. doi:10.1038/nrg3240 (2012).

153. Binder, E. B. Dissecting the molecular mechanisms of gene x environment interactions: implications for diagnosis and treatment of stress-related psychiatric disorders. *European Journal of Psychotraumatology* **8,** 1412745. doi:10.1080/20008198.2017.1412745 (sup5 2017).

154. Genovese, A. & Butler, M. G. The Autism Spectrum: Behavioral, Psychiatric and Genetic Associations. *Genes* **14,** 677. doi:10.3390/genes14030677 (2023).

155. Corfield, E. C., Yang, Y., Martin, N. G. & Nyholt, D. R. A continuum of genetic liability for minor and major depression. *Translational Psychiatry* **7,** e1131–e1131. doi:10.1038/tp.2017.99 (2017).

156. Flint, J. The genetic basis of major depressive disorder. *Molecular Psychiatry* **28,** 2254–2265. doi:10.1038/s41380-023-01957-9 (2023).

157. O'Donnell, K. J. & Meaney, M. J. Fetal Origins of Mental Health: The Developmental Origins of Health and Disease Hypothesis. *American Journal of Psychiatry* **174,** 319–328. doi:10.1176/appi.ajp.2016.16020138 (2017).

158. Kotsiri, I., Resta, P., Spyrantis, A. *et al.* Viral Infections and Schizophrenia: A Comprehensive Review. *Viruses* **15,** 1345. doi:10.3390/v15061345 (2023).

159. Cheroni, C., Caporale, N. & Testa, G. Autism spectrum disorder at the crossroad between genes and environment: contributions, convergences, and interactions in ASD developmental pathophysiology. *Molecular Autism* **11,** 69. doi:10.1186/s13229-020-00370-1 (2020).

160. Davies, C., Segre, G., Estradé, A. *et al.* Prenatal and perinatal risk and protective factors for psychosis: a systematic review and meta-analysis. *The Lancet Psychiatry* **7,** 399–410. doi:10.1016/S2215-0366(20)30057-2 (2020).

161. Ursini, G., Punzi, G., Langworthy, B. W. *et al.* Placental genomic risk scores and early neurodevelopmental outcomes. *Proceedings of the National Academy of Sciences* **118,** e2019789118. doi:10.1073/pnas.2019789118 (2021).

162. O'Donnell, K. J., Glover, V., Barker, E. D. & O'Connor, T. G. The persisting effect of maternal mood in pregnancy on childhood psychopathology. *Development and Psychopathology* **26,** 393–403. doi:10.1017/S0954579414000029 (2014).

163. Davis, E. P., Hankin, B. L., Glynn, L. M. *et al.* Prenatal Maternal Stress, Child Cortical Thickness, and Adolescent Depressive Symptoms. *Child Development* **91,** e432–e450. doi:10.1111/cdev.13252 (2020).

164. Moog, N. K., Entringer, S., Rasmussen, J. M. *et al.* Intergenerational Effect of Maternal Exposure to Childhood Maltreatment on Newborn Brain Anatomy. *Biological Psychiatry* **83,** 120–127. doi:10.1016/j.biopsych.2017.07.009 (2018).

165. Krontira, A. C., Cruceanu, C. & Binder, E. B. Glucocorticoids as Mediators of Adverse Outcomes of Prenatal Stress. *Trends in Neurosciences* **43,** 394–405. doi:10.1016/j.tins.2020.03.008 (2020).

166. Love, M. I., Huska, M. R., Jurk, M. *et al.* Role of the chromatin landscape and sequence in determining cell type-specific genomic glucocorticoid receptor binding and gene regulation. *Nucleic Acids Research* **45,** 1805–1819. doi:10.1093/nar/gkw1163 (2017).

167. Martens, C., Bilodeau, S., Maira, M., Gauthier, Y. & Drouin, J. Protein-Protein Interactions and Transcriptional Antagonism between the Subfamily of NGFI-B/Nur77 Orphan Nuclear Receptors and Glucocorticoid Receptor. *Molecular Endocrinology* **19,** 885–897. doi:10.1210/me.2004-0333 (2005).

168. Bridges, J. P., Sudha, P., Lipps, D. *et al.* Glucocorticoid regulates mesenchymal cell differentiation required for perinatal lung morphogenesis and function. *American Journal of Physiology-Lung Cellular and Molecular Physiology* **319,** L239–L255. doi:10.1152/ajplung.00459.2019 (2020).

169. Jellyman, J. K., Fletcher, A. J., Fowden, A. L. & Giussani, D. A. Glucocorticoid Maturation of Fetal Cardiovascular Function. *Trends in Molecular Medicine* **26,** 170–184. doi:10.1016/j.molmed.2019.09.005 (2020).

170. Edwards, P. D. & Boonstra, R. Glucocorticoids and CBG during pregnancy in mammals: diversity, pattern, and function. *General and Comparative Endocrinology* **259,** 122–130. doi:10.1016/j.ygcen.2017.11.012 (2018).

171. Carson, R., Monaghan-Nichols, A. P., DeFranco, D. B. & Rudine, A. C. Effects of antenatal glucocorticoids on the developing brain. *Steroids* **114,** 25–32. doi:10.1016/j.steroids.2016.05.012 (2016).

172. Bassil, K., Krontira, A. C., Leroy, T. *et al.* In vitro modeling of the neurobiological effects of glucocorticoids: A review. *Neurobiology of Stress* **23,** 100530. doi:10.1016/j.ynstr.2023.100530 (2023).

173. Chan, J. C., Nugent, B. M. & Bale, T. L. Parental Advisory: Maternal and Paternal Stress Can Impact Offspring Neurodevelopment. *Biological Psychiatry* **83,** 886–894. doi:10.1016/j.biopsych.2017.10.005 (2018).

174. Harris, A. & Seckl, J. Glucocorticoids, prenatal stress and the programming of disease. *Hormones and Behavior* **59,** 279–289. doi:10.1016/j.yhbeh.2010.06.007 (2011).

175. Graham, A. M., Rasmussen, J. M., Entringer, S. *et al.* Maternal Cortisol Concentrations During Pregnancy and Sex-Specific Associations With Neonatal Amygdala Connectivity and Emerging Internalizing Behaviors. *Biological Psychiatry* **85,** 172–181. doi:10.1016/j.biopsych.2018.06.023 (2019).

176. Provençal, N., Arloth, J., Cattaneo, A. *et al.* Glucocorticoid exposure during hippocampal neurogenesis primes future stress response by inducing changes in DNA methylation. *Proceedings of the National Academy of Sciences* **117,** 23280–23285. doi:10.1073/pnas.1820842116 (2020).

177. Buss, C., Entringer, S., Moog, N. K. *et al.* Intergenerational Transmission of Maternal Childhood Maltreatment Exposure: Implications for Fetal Brain Development. *Journal of the American Academy of Child & Adolescent Psychiatry* **56,** 373–382. doi:10.1016/j.jaac.2017.03.001 (2017).

178. Lajic, S., Karlsson, L. & Nordenström, A. Prenatal Treatment of Congenital Adrenal Hyperplasia: Long-Term Effects of Excess Glucocorticoid Exposure. *Hormone Research in Paediatrics* **89,** 362–371. doi:10.1159/000485100 (2018).

179. Vidavalur, R., Hussain, Z. & Hussain, N. Association of Survival at 22 Weeks' Gestation With Use of Antenatal Corticosteroids and Mode of Delivery in the United States. *JAMA Pediatrics* **177,** 90. doi:10.1001/jamapediatrics.2022.3951 (2023).

180. Ohuma, E. O., Moller, A.-B., Bradley, E. *et al.* National, regional, and global estimates of preterm birth in 2020, with trends from 2010: a systematic analysis. *The Lancet* **402,** 1261–1271. doi:10.1016/S0140-6736(23)00878-4 (2023).

181. Ninan, K., Liyanage, S. K., Murphy, K. E., Asztalos, E. V. & McDonald, S. D. Evaluation of Long-term Outcomes Associated With Preterm Exposure to Antenatal Corticosteroids: A Systematic Review and Meta-analysis. *JAMA Pediatrics* **176,** e220483. doi:10.1001/jamapediatrics.2022.0483 (2022).

182. Tsiarli, M. A., Rudine, A., Kendall, N. *et al.* Antenatal dexamethasone exposure differentially affects distinct cortical neural progenitor cells and triggers long-term changes in murine cerebral architecture and behavior. *Translational Psychiatry* **7,** e1153–e1153. doi:10.1038/tp.2017.65 (2017).

183. Lin, Y.-H., Lin, C.-H., Lin, M.-C., Hsu, Y.-C. & Hsu, C.-T. Antenatal Corticosteroid Exposure is Associated with Childhood Mental Disorders in Late Preterm and Term Infants. *The Journal of Pediatrics* **253,** 245–251.e2. doi:10.1016/j.jpeds.2022.09.050 (2023).

184. Melamed, N., Asztalos, E., Murphy, K. *et al.* Neurodevelopmental disorders among term infants exposed to antenatal corticosteroids during pregnancy: a population-based study. *BMJ Open* **9,** e031197. doi:10.1136/bmjopen-2019-031197 (2019).

185. Räikkönen, K., Gissler, M. & Kajantie, E. Associations Between Maternal Antenatal Corticosteroid Treatment and Mental and Behavioral Disorders in Children. *JAMA* **323,** 1924. doi:10.1001/jama.2020.3937 (2020).

186. Räikkönen, K., Gissler, M., Tapiainen, T. & Kajantie, E. Associations Between Maternal Antenatal Corticosteroid Treatment and Psychological Developmental and Neurosensory Disorders in Children. *JAMA Network Open* **5,** e2228518. doi:10.1001/jamanetworkopen.2022.28518 (2022).

187. Shoener, J. A., Baig, R. & Page, K. C. Prenatal exposure to dexamethasone alters hippocampal drive on hypothalamic-pituitary-adrenal axis activity in adult male rats. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* **290,** R1366–R1373. doi:10.1152/ajpregu.00757.2004 (2006).

188. Noorlander, C. W., Tijsseling, D., Hessel, E. V. S. *et al.* Antenatal Glucocorticoid Treatment Affects Hippocampal Development in Mice. *PLoS ONE* **9,** e85671. doi:10.1371/journal.pone.0085671 (2014).

189. Antonow-Schlorke, I., Kühn, B., Müller, T. *et al.* Antenatal betamethasone treatment reduces synaptophysin immunoreactivity in presynaptic terminals in the fetal sheep brain. *Neuroscience Letters* **297,** 147–150. doi:10.1016/S0304-3940(00)01605-0 (2001).

190. Wei, M., Gao, Q., Liu, J. *et al.* Development programming: Stress during gestation alters offspring development in sheep. *Reproduction in Domestic Animals* **58,** 1497–1511. doi:10.1111/rda.14465 (2023).

191. Jourdon, A., Wu, F., Mariani, J. *et al.* Modeling idiopathic autism in forebrain organoids reveals an imbalance of excitatory cortical neuron subtypes during early neurogenesis. *Nature Neuroscience* **26,** 1505–1515. doi:10.1038/s41593-023-01399-0 (2023).

192. Mariani, J., Coppola, G., Zhang, P. *et al.* FOXG1-Dependent Dysregulation of GABA/-Glutamate Neuron Differentiation in Autism Spectrum Disorders. *Cell* **162,** 375–390. doi:10.1016/j.cell.2015.06.034 (2015).

193. Sawada, T., Chater, T. E., Sasagawa, Y. *et al.* Developmental excitation-inhibition imbalance underlying psychoses revealed by single-cell analyses of discordant twins-derived cerebral organoids. *Molecular Psychiatry* **25,** 2695–2711. doi:10.1038/s41380-020-0844-z (2020).

194. Tai, D. J., Razaz, P., Erdin, S. *et al.* Tissue- and cell-type-specific molecular and functional signatures of 16p11.2 reciprocal genomic disorder across mouse brain and human neuronal models. *The American Journal of Human Genetics* **109,** 1789–1813. doi:10.1016/j.ajhg.2022.08.012 (2022).

195. Sonnenschein, S. F., Gomes, F. V. & Grace, A. A. Dysregulation of Midbrain Dopamine System and the Pathophysiology of Schizophrenia. *Frontiers in Psychiatry* **11,** 613. doi:10.3389/fpsyt.2020.00613 (2020).

196. Zhu, Y., Owens, S. J., Murphy, C. E. *et al.* Inflammation-related transcripts define "high" and "low" subgroups of individuals with schizophrenia and bipolar disorder in the midbrain. *Brain, Behavior, and Immunity* **105,** 149–159. doi:10.1016/j.bbi.2022.06.012 (2022).

197. Elvsåshagen, T., Shadrin, A., Frei, O. *et al.* The genetic architecture of the human thalamus and its overlap with ten common brain disorders. *Nature Communications* **12,** 2909. doi:10.1038/s41467-021-23175-z (2021).

198. Hwang, W. J., Kwak, Y. B., Cho, K. I. K. *et al.* Thalamic Connectivity System Across Psychiatric Disorders: Current Status and Clinical Implications. *Biological Psychiatry Global Open Science* **2,** 332–340. doi:10.1016/j.bpsgos.2021.09.008 (2022).

199. Fischer, S. The hypothalamus in anxiety disorders. *Handbook of Clinical Neurology* **180,** 149–160. doi:10.1016/B978-0-12-820107-7.00009-4 (2021).

200. Terlevic, R., Isola, M., Ragogna, M. *et al.* Decreased hypothalamus volumes in generalized anxiety disorder but not in panic disorder. *Journal of Affective Disorders* **146,** 390–394. doi:10.1016/j.jad.2012.09.024 (2013).

201. Bast, N., Poustka, L. & Freitag, C. M. The locus coeruleus–norepinephrine system as pacemaker of attention – a developmental mechanism of derailed attentional function in autism spectrum disorder. *European Journal of Neuroscience* **47,** 115–125. doi:10.1111/ejn.13795 (2018).

202. Bloomer, B. F., Morales, J. J., Bolbecker, A. R., Kim, D.-J. & Hetrick, W. P. Cerebellar Structure and Function in Autism Spectrum Disorder. *Journal of Psychiatry and Brain Science* **7,** e220003. doi:10.20900/jpbs.20220003 (2022).

203. Mapelli, L., Soda, T., D'Angelo, E. & Prestori, F. The Cerebellar Involvement in Autism Spectrum Disorders: From the Social Brain to Mouse Models. *International Journal of Molecular Sciences* **23,** 3894. doi:10.3390/ijms23073894 (2022).

204. He, Z., Dony, L., Fleck, J. S. *et al. An integrated transcriptomic cell atlas of human neural organoids* bioRxiv:2023.10.05.561097. 2023. doi:10.1101/2023.10.05.561097.

205. Dony, L., Krontira, A. C., Kaspar, L. *et al. Chronic exposure to glucocorticoids amplifies inhibitory neuron cell fate during human neurodevelopment in organoids* bioRxiv:2024.01.21.576532. 2024. doi:10.1101/2024.01.21.576532.

206. Avsec, Ž., Kreuzhuber, R., Israeli, J. *et al.* The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nature Biotechnology* **37,** 592–600. doi:10.1038/s41587-019-0140-0 (2019).

207. European Organization For Nuclear Research & OpenAIRE. *Zenodo: Research. Shared.* doi: 10.25495/7gxk-rd71. 2013.

208. McInnes, L., Healy, J. & Melville, J. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* 2020. arXiv: 1802.03426.

209. Eraslan, G., Avsec, Ž., Gagneur, J. & Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics* **20,** 389–403. doi:10.1038/s41576-019-0122-6 (2019).

210. Pedregosa, F., Varoquaux, G., Gramfort, A. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).

211. 10x Genomics Datasets Repository. *10k PBMCs from a Healthy Donor (v3 chemistry)* https://10xgenomics.com/datasets/10-k-pbm-cs-from-a-healthy-donor-v-3-chemistry-3-standard-3-0-0. 2018.

212. Aizarani, N., Saviano, A., Sagar *et al.* A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature* **572,** 199–204. doi:10.1038/s41586-019-1373-2 (2019).

213. Baron, M., Veres, A., Wolock, S. L. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Systems* **3,** 346–360.e4. doi:10.1016/j.cels.2016.08.011 (2016).

214. Enge, M., Arda, H. E., Mignardi, M. *et al.* Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. *Cell* **171,** 321–330.e14. doi:10.1016/j.cell.2017.09.004 (2017).

215. Guo, J., Grow, E. J., Mlcochova, H. *et al.* The adult human testis transcriptional cell atlas. *Cell Research* **28,** 1141–1157. doi:10.1038/s41422-018-0099-2 (2018).

216. Habermann, A. C., Gutierrez, A. J., Bui, L. T. *et al.* Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. *Science Advances* **6,** eaba1972. doi:10.1126/sciadv.aba1972 (2020).

217. Habib, N., Avraham-Davidi, I., Basu, A. *et al.* Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nature Methods* **14,** 955–958. doi:10.1038/nmeth.4407 (2017).

218. Han, X., Zhou, Z., Fei, L. *et al.* Construction of a human cell landscape at single-cell level. *Nature* **581,** 303–309. doi:10.1038/s41586-020-2157-4 (2020).

219. Hao, Y., Hao, S., Andersen-Nissen, E. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184,** 3573–3587.e29. doi:10.1016/j.cell.2021.04.048 (2021).

220. Henry, G. H., Malewska, A., Joseph, D. B. *et al.* A Cellular Anatomy of the Normal Adult Human Prostate and Prostatic Urethra. *Cell Reports* **25,** 3530–3542.e5. doi:10.1016/j.celrep.2018.11.086 (2018).

221. James, K. R., Gomes, T., Elmentaite, R. *et al.* Distinct microbial and immune niches of the human colon. *Nature Immunology* **21,** 343–353. doi:10.1038/s41590-020-0602-z (2020).

222. Kinchen, J., Chen, H. H., Parikh, K. *et al.* Structural Remodeling of the Human Colonic Mesenchyme in Inflammatory Bowel Disease. *Cell* **175,** 372–386.e17. doi:10.1016/j.cell.2018.08.067 (2018).

223. Lake, B. B., Chen, S., Hoshi, M. *et al.* A single-nucleus RNA-sequencing pipeline to decipher the molecular anatomy and pathophysiology of human kidneys. *Nature Communications* **10,** 2832. doi:10.1038/s41467-019-10861-2 (2019).

224. Liao, J., Yu, Z., Chen, Y. *et al.* Single-cell RNA sequencing of human kidney. *Scientific Data* **7,** 4. doi:10.1038/s41597-019-0351-8 (2020).

225. Lukassen, S., Chua, R. L., Trefzer, T. *et al.* SARS-CoV-2 receptor ACE2 and TMPRSS2 are primarily expressed in bronchial transient secretory cells. *The EMBO Journal* **39,** e105114. doi:10.15252/embj.20105114 (2020).

226. Lukowski, S. W., Lo, C. Y., Sharov, A. A. *et al.* A single-cell transcriptome atlas of the adult human retina. *The EMBO Journal* **38,** e100811. doi:10.15252/embj.2018100811 (2019).

227. MacParland, S. A., Liu, J. C., Ma, X.-Z. *et al.* Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nature Communications* **9,** 4383. doi:10.1038/s41467-018-06318-7 (2018).

228. Madissoon, E., Wilbrey-Clark, A., Miragaia, R. J. *et al.* scRNA-seq assessment of the human lung, spleen, and esophagus tissue stability after cold preservation. *Genome Biology* **21,** 1. doi:10.1186/s13059-019-1906-x (2020).

229. Martin, J. C., Chang, C., Boschetti, G. *et al.* Single-Cell Analysis of Crohn's Disease Lesions Identifies a Pathogenic Cellular Module Associated with Resistance to Anti-TNF Therapy. *Cell* **178,** 1493–1508.e20. doi:10.1016/j.cell.2019.08.008 (2019).

230. Menon, M., Mohammadi, S., Davila-Velderrain, J. *et al.* Single-cell transcriptomic atlas of the human retina identifies cell types associated with age-related macular degeneration. *Nature Communications* **10,** 4902. doi:10.1038/s41467-019-12780-8 (2019).

231. Miller, A. J., Yu, Q., Czerwinski, M. *et al.* In Vitro and In Vivo Development of the Human Airway at Single-Cell Resolution. *Developmental Cell* **53,** 117–128.e6. doi:10.1016/j.devcel.2020.01.033 (2020).

232. Park, J.-E., Botting, R. A., Domínguez Conde, C. *et al.* A cell atlas of human thymic development defines T cell repertoire formation. *Science* **367,** eaay3224. doi:10.1126/science.aay3224 (2020).

233. Popescu, D.-M., Botting, R. A., Stephenson, E. *et al.* Decoding human fetal liver haematopoiesis. *Nature* **574,** 365–371. doi:10.1038/s41586-019-1652-y (2019).

234. Ramachandran, P., Dobie, R., Wilson-Kanamori, J. R. *et al.* Resolving the fibrotic niche of human liver cirrhosis at single-cell level. *Nature* **575,** 512–518. doi:10.1038/s41586-019-1631-3 (2019).

235. Rozenblatt-Rosen, O., Li, B., Kowalczyk, M. S. *et al. A single cell immune cell atlas of human hematopoietic system* https://explore.data.humancellatlas.org/projects/cc95ff89-2e68-4a08-a234-480eca21ce79. 2020.

236. Segerstolpe, Å., Palasantza, A., Eliasson, P. *et al.* Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metabolism* **24,** 593–607. doi:10.1016/j.cmet.2016.08.020 (2016).

237. Smillie, C. S., Biton, M., Ordovas-Montanes, J. *et al.* Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. *Cell* **178,** 714–730.e22. doi:10.1016/j.cell.2019.06.029 (2019).

238. Stewart, B. J., Ferdinand, J. R., Young, M. D. *et al.* Spatiotemporal immune zonation of the human kidney. *Science* **365,** 1461–1466. doi:10.1126/science.aat5031 (2019).

239. Szabo, P. A., Levitin, H. M., Miron, M. *et al.* Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. *Nature Communications* **10,** 4706. doi:10.1038/s41467-019-12464-3 (2019).

240. Vento-Tormo, R., Efremova, M., Botting, R. A. *et al.* Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature* **563,** 347–353. doi:10.1038/s41586-018-0698-6 (2018).

241. Vieira Braga, F. A., Kar, G., Berg, M. *et al.* A cellular census of human lungs identifies novel cell states in health and in asthma. *Nature Medicine* **25,** 1153–1163. doi:10.1038/s41591-019-0468-5 (2019).

242. Voigt, A. P., Mulfaul, K., Mullin, N. K. *et al.* Single-cell transcriptomics of the human retinal pigment epithelium and choroid in health and macular degeneration. *Proceedings of the National Academy of Sciences* **116,** 24100–24107. doi:10.1073/pnas.1914143116 (2019).

243. Wang, Y., Song, W., Wang, J. *et al.* Single-cell transcriptome analysis reveals differential nutrient absorption functions in human intestine. *Journal of Experimental Medicine* **217,** e20191130. doi:10.1084/jem.20191130 (2020).

244. Han, X., Wang, R., Zhou, Y. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172,** 1091–1107.e17. doi:10.1016/j.cell.2018.02.001 (2018).

245. The Tabula Muris Consortium, Almanzar, N., Antony, J. *et al.* A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* **583,** 590–595. doi:10.1038/s41586-020-2496-1 (2020).

246. Thompson, P. J., Shah, A., Ntranos, V. *et al.* Targeted Elimination of Senescent Beta Cells Prevents Type 1 Diabetes. *Cell Metabolism* **29,** 1045–1060.e10. doi:10.1016/j.cmet.2019.01.021 (2019).

247. Van Hove, H., Martens, L., Scheyltjens, I. *et al.* A single-cell atlas of mouse brain macrophages reveals unique transcriptional identities shaped by ontogeny and tissue environment. *Nature Neuroscience* **22,** 1021–1035. doi:10.1038/s41593-019-0393-4 (2019).

248. Muraro, M. J., Dharmadhikari, G., Grün, D. *et al.* A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Systems* **3,** 385–394.e3. doi:10.1016/j.cels.2016.09.002 (2016).

249. Voigt, A., Whitmore, S., Flamme-Wiese, M. *et al.* Molecular characterization of foveal versus peripheral human retina by single-cell RNA sequencing. *Experimental Eye Research* **184,** 234–242. doi:10.1016/j.exer.2019.05.001 (2019).

250. Muto, Y., Wilson, P. C., Ledru, N. *et al.* Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney. *Nature Communications* **12,** 2190. doi:10.1038/s41467-021-22368-w (2021).

251. Litviňuková, M., Talavera-López, C., Maatz, H. *et al.* Cells of the adult human heart. *Nature* **588,** 466–472. doi:10.1038/s41586-020-2797-4 (2020).

252. Khan, T. A., Revah, O., Gordon, A. *et al.* Neuronal defects in a human cellular model of 22q11.2 deletion syndrome. *Nature Medicine* **26,** 1888–1898. doi:10.1038/s41591-020-1043-9 (2020).

253. Bowles, K. R., Silva, M. C., Whitney, K. *et al.* ELAVL4, splicing, and glutamatergic dysfunction precede neuron loss in MAPT mutation cerebral organoids. *Cell* **184,** 4547–4563.e17. doi:10.1016/j.cell.2021.07.003 (2021).

254. Fleck, J. S., Jansen, S. M. J., Wollny, D. *et al.* Inferring and perturbing cell fate regulomes in human brain organoids. *Nature* **621,** 365–372. doi:10.1038/s41586-022-05279-8 (2023).

255. He, Z., Maynard, A., Jain, A. *et al.* Lineage recording in human cerebral organoids. *Nature Methods* **19,** 90–99. doi:10.1038/s41592-021-01344-8 (2022).

256. Kelava, I., Chiaradia, I., Pellegrini, L., Kalinka, A. T. & Lancaster, M. A. Androgens increase excitatory neurogenic potential in human brain organoids. *Nature* **602,** 112–116. doi:10.1038/s41586-021-04330-4 (2022).

257. Uzquiano, A., Kedaigle, A. J., Pigoni, M. *et al.* Proper acquisition of cell class identity in organoids allows definition of fate specification programs of the human cerebral cortex. *Cell* **185,** 3770–3788.e27. doi:10.1016/j.cell.2022.09.010 (2022).

258. Fischer, D. S., Dony, L., König, M. *et al.* Sfaira accelerates data and model reuse in single cell genomics. *Genome Biology* **22,** 248. doi:10.1186/s13059-021-02452-6 (2021).

259. Virshup, I., Rybakov, S., Theis, F. J., Angerer, P. & Wolf, F. A. *anndata: Annotated data* bioRxiv:2021.12.16.473007. 2021. doi:10.1101/2021.12.16.473007.

260. Cunningham, F., Allen, J. E., Allen, J. *et al.* Ensembl 2022. *Nucleic Acids Research* **50,** D988–D995. doi:10.1093/nar/gkab1049 (D1 2022).

261. Kinsella, R. J., Kahari, A., Haider, S. *et al.* Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* **2011,** bar030–bar030. doi:10.1093/database/bar030 (2011).

262. Traag, V. A., Waltman, L. & Van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* **9,** 5233. doi:10.1038/s41598-019-41695-z (2019).

263. Nolet, C., Lal, A., Ilango, R. *et al. Accelerating single-cell genomic analysis with GPUs* bioRxiv:2022.05.26.493607. 2022. doi:10.1101/2022.05.26.493607.

264. Klein, D., Uscidda, T., Theis, F. & Cuturi, M. *Entropic (Gromov) Wasserstein Flow Matching with GENOT* 2024. arXiv: 2310.09254.

265. Klein, D., Palla, G., Lange, M. *et al. Mapping cells through time and space with moscot* bioRxiv:2023.05.11.540374. 2023. doi:10.1101/2023.05.11.540374.

266. Gayoso, A., Lopez, R., Xing, G. *et al.* A Python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology* **40,** 163–166. doi:10.1038/s41587-021-01206-w (2022).

267. Virshup, I., Bredikhin, D., Heumos, L. *et al.* The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nature Biotechnology* **41,** 604–606. doi:10.1038/s41587-023-01733-8 (2023).

268. Büttner, M., Ostner, J., Müller, C. L., Theis, F. J. & Schubert, B. scCODA is a Bayesian model for compositional single-cell data analysis. *Nature Communications* **12,** 6876. doi:10.1038/s41467-021-27150-6 (2021).

269. Tay, J. K., Narasimhan, B. & Hastie, T. Elastic Net Regularization Paths for All Generalized Linear Models. *Journal of Statistical Software* **106,** 1–31. doi:10.18637/jss.v106.i01 (2023).

270. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323. doi:10.1186/1471-2105-12-323 (2011).

271. Dobin, A., Davis, C. A., Schlesinger, F. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21. doi:10.1093/bioinformatics/bts635 (2013).

272. Badia-i-Mompel, P., Vélez Santiago, J., Braunger, J. *et al.* decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinformatics Advances* **2,** vbac016. doi:10.1093/bioadv/vbac016 (2022).

273. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR : a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140. doi:10.1093/bioinformatics/btp616 (2010).

274. Fang, Z., Liu, X. & Peltz, G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* **39,** btac757. doi:10.1093/bioinformatics/btac757 (2023).

275. Shen, W.-K., Chen, S.-Y., Gan, Z.-Q. *et al.* AnimalTFDB 4.0: a comprehensive animal transcription factor database updated with variation and expression annotations. *Nucleic Acids Research* **51,** D39–D45. doi:10.1093/nar/gkac907 (D1 2023).

276. Chen, E. Y., Tan, C. M., Kou, Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14,** 128. doi:10.1186/1471-2105-14-128 (2013).

277. Kuleshov, M. V., Jones, M. R., Rouillard, A. D. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research* **44,** W90–W97. doi:10.1093/nar/gkw377 (W1 2016).

278. Koyanagi-Aoi, M., Ohnuki, M., Takahashi, K. *et al.* Differentiation-defective phenotypes revealed by large-scale analyses of human pluripotent stem cells. *Proceedings of the National Academy of Sciences* **110,** 20569–20574. doi:10.1073/pnas.1319061110 (2013).

279. Okita, K., Matsumura, Y., Sato, Y. *et al.* A more efficient method to generate integration-free human iPS cells. *Nature Methods* **8,** 409–412. doi:10.1038/nmeth.1591 (2011).

280. Brückl, T. M., Spoormaker, V. I., Sämann, P. G. *et al.* The biological classification of mental disorders (BeCOME) study: a protocol for an observational deep-phenotyping study for the identification of biological subtypes. *BMC Psychiatry* **20,** 213. doi:10.1186/s12888-020-02541-z (2020).

281. Bagley, J. A., Reumann, D., Bian, S., Lévi-Strauss, J. & Knoblich, J. A. Fused cerebral organoids model interactions between brain regions. *Nature Methods* **14,** 743–751. doi:10.1038/nmeth.4304 (2017).

282. Riesenberg, S. & Maricic, T. Targeting repair pathways with small molecules increases precise genome editing in pluripotent stem cells. *Nature Communications* **9,** 2164. doi:10.1038/s41467-018-04609-7 (2018).

283. Cunningham, F., Achuthan, P., Akanni, W. *et al.* Ensembl 2019. *Nucleic Acids Research* **47,** D745–D751. doi:10.1093/nar/gky1113 (D1 2019).

284. Cruceanu, C., Dony, L., Krontira, A. C. *et al.* Cell-Type-Specific Impact of Glucocorticoid Receptor Activation on the Developing Brain: A Cerebral Organoid Study. *American Journal of Psychiatry* **179,** 375–387. doi:10.1176/appi.ajp.2021.21010095 (2022).

285. L. Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology* **17,** 75. doi:10.1186/s13059-016-0947-7 (2016).

286. Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE* **9,** e98679. doi:10.1371/journal.pone.0098679 (2014).

287. Ashburner, M., Ball, C. A., Blake, J. A. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics* **25,** 25–29. doi:10.1038/75556 (2000).

288. Gene Ontology Consortium, Aleksander, S. A., Balhoff, J. *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* **224,** iyad031. doi:10.1093/genetics/iyad031 (2023).

289. Finak, G., McDavid, A., Yajima, M. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology* **16,** 278. doi:10.1186/s13059-015-0844-5 (2015).

290. Müller-Dott, S., Tsirvouli, E., Vazquez, M. *et al.* Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. *Nucleic Acids Research* **51,** 10934–10949. doi:10.1093/nar/gkad841 (2023).

291. Reijnders, M. J. M. F. & Waterhouse, R. M. Summary Visualizations of Gene Ontology Terms With GO-Figure! *Frontiers in Bioinformatics* **1,** 638255. doi:10.3389/fbinf.2021.638255 (2021).

292. Polański, K., Young, M. D., Miao, Z. *et al.* BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* **36,** 964–965. doi:10.1093/bioinformatics/btz625 (2020).

293. Maaten, L. v. d. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9,** 2579–2605 (2008).

294. Weiler, P., Lange, M., Klein, M., Pe'er, D. & Theis, F. J. *Unified fate mapping in multiview single-cell data* bioRxiv:2023.07.19.549685. 2023. doi:10.1101/2023.07.19.549685.

295. Reuter, B., Fackeldey, K. & Weber, M. Generalized Markov modeling of nonreversible molecular kinetics. *The Journal of Chemical Physics* **150,** 174103. doi:10.1063/1.5064530 (2019).

296. Virtanen, P., Gommers, R., Oliphant, T. E. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17,** 261–272. doi:10.1038/s41592-019-0686-2 (2020).

297. Wood, S. N. Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **73,** 3–36. doi:10.1111/j.1467-9868.2010.00749.x (2011).

298. Satpathy, A. T., Granja, J. M., Yost, K. E. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nature Biotechnology* **37,** 925–936. doi:10.1038/s41587-019-0206-z (2019).

299. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nature Methods* **18,** 1333–1341. doi:10.1038/s41592-021-01282-5 (2021).

300. Thibodeau, A., Eroglu, A., McGinnis, C. S. *et al.* AMULET: a novel read count-based method for effective multiplet detection from single nucleus ATAC-seq data. *Genome Biology* **22,** 252. doi:10.1186/s13059-021-02469-x (2021).

301. Stuart, T., Butler, A., Hoffman, P. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177,** 1888–1902.e21. doi:10.1016/j.cell.2019.05.031 (2019).

302. Waltman, L. & Van Eck, N. J. A smart local moving algorithm for large-scale modularity-based community detection. *The European Physical Journal B* **86,** 471. doi:10.1140/epjb/e2013-40829-0 (2013).

303. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nature Methods* **14,** 975–978. doi:10.1038/nmeth.4401 (2017).

304. Fornes, O., Castro-Mondragon, J. A., Khan, A. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **48,** D87–D92. doi:10.1093/nar/gkz1001 (D1 2019).

305. Cao, Z.-J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology* **40,** 1458–1466. doi:10.1038/s41587-022-01284-4 (2022).

306. Stark, S. G., Ficek, J., Locatello, F. *et al.* SCIM: universal single-cell matching with unpaired feature sets. *Bioinformatics* **36,** i919–i927. doi:10.1093/bioinformatics/btaa843 (Supplement_2 2020).

307. Van Dijk, D., Sharma, R., Nainys, J. *et al.* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174,** 716–729.e27. doi:10.1016/j.cell.2018.05.061 (2018).

308. Badia-I-Mompel, P., Wessels, L., Müller-Dott, S. *et al.* Gene regulatory network inference in the era of single-cell multi-omics. *Nature Reviews Genetics* **24,** 739–754. doi:10.1038/s41576-023-00618-5 (2023).

309. Wickham, H. *ggplot2* doi: 10.1007/978-3-319-24277-4 (Springer International Publishing, Cham, 2016).

310. Vasilevsky, N. A., Matentzoglu, N. A., Toro, S. *et al. Mondo: Unifying diseases for the world, by the world* medRxiv:2022.04.13.22273750. 2022. doi:10.1101/2022.04.13.22273750.

311. Bairoch, A. The Cellosaurus, a Cell-Line Knowledge Resource. *Journal of Biomolecular Techniques : JBT* **29,** 25–38. doi:10.7171/jbt.18-2902-002 (2018).

312. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Communications of the ACM* **60,** 84–90. doi:10.1145/3065386 (2017).

313. Zappia, L., Phipson, B. & Oshlack, A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLOS Computational Biology* **14,** e1006245. doi:10.1371/journal.pcbi.1006245 (2018).

314. Wang, S., Pisco, A. O., McGeever, A. *et al.* Leveraging the Cell Ontology to classify unseen cell types. *Nature Communications* **12,** 5556. doi:10.1038/s41467-021-25725-x (2021).

315. Köhler, N. D., Büttner, M., Andriamanga, N. & Theis, F. J. *Deep learning does not outperform classical machine learning for cell-type annotation* bioRxiv:653907. 2019. doi:10.1101/653907.

316. Abdelaal, T., Michielsen, L., Cats, D. *et al.* A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biology* **20,** 194. doi:10.1186/s13059-019-1795-z (2019).

317. Stein-O'Brien, G. L., Clark, B. S., Sherman, T. *et al.* Decomposing Cell Identity for Transfer Learning across Cellular Measurements, Platforms, Tissues, and Species. *Cell Systems* **8,** 395–411.e8. doi:10.1016/j.cels.2019.04.004 (2019).

318. Wang, J., Agarwal, D., Huang, M. *et al.* Data denoising with transfer learning in single-cell transcriptomics. *Nature Methods* **16,** 875–878. doi:10.1038/s41592-019-0537-1 (2019).

319. Dony, L., König, M., Fischer, D. S. & Theis, F. J. *Variational autoencoders with flexible priors enable robust distribution learning on single-cell RNA sequencing data* in *ICML 2020 Workshop on Computational Biology Proceedings* ICML 2020 Workshop on Computational Biology (WCB). https://icml-compbio.github.io/2020/papers/WCBICML2020_paper_37.pdf (2020), Paper 37.

320. Xian, Y., Schiele, B. & Akata, Z. *Zero-Shot Learning — The Good, the Bad and the Ugly* in. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE, Honolulu, HI, 2017), 3077–3086. doi:10.1109/CVPR.2017.328.

321. Amin, N. D., Kelley, K. W., Hao, J. *et al. Generating human neural diversity with a multiplexed morphogen screen in organoids* bioRxiv:2023.05.31.541819. 2023. doi:10.1101/2023.05.31.541819.

322. Pitale, P. M., Howse, W. & Gorbatyuk, M. Neuronatin Protein in Health and Disease. *Journal of Cellular Physiology* **232,** 477–481. doi:10.1002/jcp.25498 (2017).

323. Rad, A., Altunoglu, U., Miller, R. *et al.* MAB21L1 loss of function causes a syndromic neurodevelopmental disorder with distinctive cerebellar, ocular, cranio facial and genital features (COFG syndrome). *Journal of Medical Genetics* **56,** 332–339. doi:10.1136/jmedgenet-2018-105623 (2019).

324. Schanze, I., Bunt, J., Lim, J. W. *et al.* NFIB Haploinsufficiency Is Associated with Intellectual Disability and Macrocephaly. *The American Journal of Human Genetics* **103,** 752–768. doi:10.1016/j.ajhg.2018.10.006 (2018).

325. Montalbán-Loro, R., Lassi, G., Lozano-Ureña, A. *et al.* Dlk1 dosage regulates hippocampal neurogenesis and cognition. *Proceedings of the National Academy of Sciences* **118,** e2015505118. doi:10.1073/pnas.2015505118 (2021).

326. Feng, Y., Reznik, S. E. & Fricker, L. D. ProSAAS and prohormone convertase 1 are broadly expressed during mouse development. *Gene Expression Patterns* **1,** 135–140. doi:10.1016/S1567-133X(02)00002-9 (2002).

327. Vandervore, L., Stouffs, K., Tanyalçin, I. *et al.* Bi-allelic variants in COL3A1 encoding the ligand to GPR56 are associated with cobblestone-like cortical malformation, white matter changes and cerebellar cysts. *Journal of Medical Genetics* **54,** 432–440. doi:10.1136/jmedgenet-2016-104421 (2017).

328. El Amri, M., Fitzgerald, U. & Schlosser, G. MARCKS and MARCKS-like proteins in development and regeneration. *Journal of Biomedical Science* **25,** 43. doi:10.1186/s12929-018-0445-1 (2018).

329. Rottkamp, C. A., Lobur, K. J., Wladyka, C. L., Lucky, A. K. & O'Gorman, S. Pbx3 is required for normal locomotion and dorsal horn development. *Developmental Biology* **314,** 23–39. doi:10.1016/j.ydbio.2007.10.046 (2008).

330. Shen, Y., Zhang, C., Xiao, K., Liu, D. & Xie, G. CELF4 regulates spine formation and depression-like behaviors of mice. *Biochemical and Biophysical Research Communications* **605,** 39–44. doi:10.1016/j.bbrc.2022.03.067 (2022).

331. Salamon, I., Park, Y., Miškić, T. *et al.* Celf4 controls mRNA translation underlying synaptic development in the prenatal mammalian neocortex. *Nature Communications* **14,** 6025. doi:10.1038/s41467-023-41730-8 (2023).

332. Barone, R., Fichera, M., De Grandi, M. *et al.* Familial 18q12.2 deletion supports the role of RNA-binding protein CELF4 in autism spectrum disorders. *American Journal of Medical Genetics Part A* **173,** 1649–1655. doi:10.1002/ajmg.a.38205 (2017).

333. Blankvoort, S., Olsen, L. C. & Kentros, C. G. Single Cell Transcriptomic and Chromatin Profiles Suggest Layer Vb Is the Only Layer With Shared Excitatory Cell Types in the Medial and Lateral Entorhinal Cortex. *Frontiers in Neural Circuits* **15,** 806154. doi:10.3389/fncir.2021.806154 (2022).

334. Delgado, R. N., Allen, D. E., Keefe, M. G. *et al.* Individual human cortical progenitors can produce excitatory and inhibitory neurons. *Nature* **601,** 397–403. doi:10.1038/s41586-021-04230-7 (2022).

335. Heimberg, G., Kuo, T., DePianto, D. *et al. Scalable querying of human cell atlases via a foundational model reveals commonalities across fibrosis-associated macrophages* bioRxiv:2023.07.18.549537. 2023. doi:10.1101/2023.07.18.549537.

336. Lin, H.-C., Janssens, J., Kroell, A.-S. *et al. Human neuron subtype programming through combinatorial patterning with scRNA-seq readouts* bioRxiv:2023.12.12.571318. 2023. doi:10.1101/2023.12.12.571318.

337. Sanchís-Calleja, F., Jain, A., He, Z. *et al. Decoding morphogen patterning of human neural organoids with a multiplexed single-cell transcriptomic screen* bioRxiv:2024.02.08.579413. 2024. doi:10.1101/2024.02.08.579413.

338. Lotfollahi, M., Klimovskaia Susmelj, A., De Donno, C. *et al.* Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology* **19,** e11517. doi:10.15252/msb.202211517 (2023).

339. Hetzel, L., Boehm, S., Kilbertus, N. *et al. Predicting Cellular Responses to Novel Drug Perturbations at a Single-Cell Resolution* in. Advances in Neural Information Processing Systems. https://openreview.net/forum?id=vRrFVHxFiXJ (2022).

340. Roohani, Y., Huang, K. & Leskovec, J. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology.* doi:10.1038/s41587-023-01905-6 (2023).

341. Ji, Y., Lotfollahi, M., Wolf, F. A. & Theis, F. J. Machine learning for perturbational single-cell omics. *Cell Systems* **12,** 522–537. doi:10.1016/j.cels.2021.05.016 (2021).

342. Xu, Q., Halle, L., Hediyeh-zadeh, S. *et al. An integrated transcriptomic cell atlas of human endoderm-derived organoids* bioRxiv:2023.11.20.567825. 2023. doi:10.1101/2023.11.20.567825.

343. Legnini, I., Emmenegger, L., Zappulo, A. *et al.* Spatiotemporal, optogenetic control of gene expression in organoids. *Nature Methods* **20,** 1544–1552. doi:10.1038/s41592-023-01986-w (2023).

344. Lozachmeur, G., Bramoulle, A., Aubert, A. *et al.* Three-dimensional molecular cartography of human cerebral organoids revealed by double-barcoded spatial transcriptomics. *Cell Reports Methods* **3,** 100573. doi:10.1016/j.crmeth.2023.100573 (2023).

345. Trevino, A. E., Sinnott-Armstrong, N., Andersen, J. *et al.* Chromatin accessibility dynamics in a model of human forebrain development. *Science* **367,** eaay1645. doi:10.1126/science.aay1645 (2020).

346. Ziffra, R. S., Kim, C. N., Ross, J. M. *et al.* Single-cell epigenomics reveals mechanisms of human cortical development. *Nature* **598,** 205–213. doi:10.1038/s41586-021-03209-8 (2021).