Inaugural dissertation
for
obtaining the doctoral degree
of the
Combined Faculty of Mathematics, Engineering
and Natural Sciences
of the
Ruprecht - Karls - University
Heidelberg

Presented by
M.Sc. Valentyna Zinchenko
Born in: Kostiantynivka, Ukraine
Oral examination: 10th January 2023

How to describe a cell: a path to automated versatile characterization of cells in imaging data

Referees: Prof. Dr. Alexis Maizel
Dr. Wolfgang Huber

I never liked the phrase
"within the limits of what's possible".
For there are no limits to what's possible.

Nataliia Pastukhova,
vice-rector of Donetsk Regional IPPO
and the teacher who introduced
me to quantitative biology

Summary

A cell is the basic functional unit of life. Most multicellular organisms, including animals, are composed of a variety of different cell types that fulfil distinct roles. Within an organism, all cells share the same genome, however, their diverse genetic programs lead them to acquire different molecular and anatomical characteristics. Describing these characteristics is essential for understanding how cellular diversity emerged and how it contributes to the organism function.

Probing cellular appearance by microscopy methods is the original way of describing cell types and the main approach to characterise cellular morphology and position in the organism. Present cutting-edge microscopy techniques generate immense amounts of data, requiring efficient automated unbiased methods of analysis. Not only can such methods accelerate the process of scientific discovery, they should also facilitate large-scale systematic reproducible analysis.

The necessity of processing big datasets has led to development of intricate image analysis pipelines, however, they are mostly tailored to a particular dataset and a specific research question. In this thesis I aimed to address the problem of creating more general fully-automated ways of describing cells in different imaging modalities, with a specific focus on deep neural networks as a promising solution for extracting rich general-purpose features from the analysed data. I further target the problem of integrating multiple data modalities to generate a detailed description of cells on the whole-organism level.

First, on two examples of cell analysis projects, I show how using automated image analysis pipelines and neural networks in particular, can assist characterising cells in microscopy data. In the first project I analyse a movie of drosophila embryo development to elucidate the difference in myosin patterns between two populations of cells with different shape fate. In the second project I develop a pipeline for automatic cell classification in a new imaging modality to show that the quality of the data is sufficient to tell apart cell types in a volume of mouse brain cortex.

Next, I present an extensive collaborative effort aimed at generating a whole-body multimodal cell atlas of a three-segmented *Platynereis dumerilii* worm, combining high resolution morphology and gene expression. To generate a multi-sided description of cells in the atlas I create a pipeline for assigning coherent denoised gene expression profiles, obtained from spatial gene expression maps, to cells segmented in the EM volume.

Finally, as the main project of this thesis, I focus on extracting comprehensive unbiased cell morphology features from an EM volume of *Platynereis dumerilii*. I design a fully unsupervised neural network pipeline for extracting rich morphological representations that enable grouping cells into morphological cell classes with characteristic gene expression. I further show how such descriptors could be used to explore the morphological diversity of cells, tissues and organs in the dataset.

Zusammenfassung

Zellen bilden die funktionelle Grundlage des Lebens. Vielzeller, wie zum Beispiel Tiere, bestehen aus vielen unterschiedlichen Zelltypen, die zur Erfuellung verschiedener hochspezialisierter Funktionen existieren. Unterschiedliche genetische Programme fuehren dazu, dass die Zellen eines Organismus verschiedene molekulare und anatomische Eigenschaften annehmen, obwohl sie dasselbe Genom besitzen. Die Beschreibung dieser Eigenschaften ist essentiell, um ein Verstaendnis darueber zu entwickeln, wie die Zellvielfalt entstanden ist und wie sie zur Funktion eines Organismus beitraegt.

Die Mikroskopie dient als Standardwerkzeug, um die Morphologien und Positionen unterschiedlicher Zelltypen in Organismen zu beschreiben. Moderne Mikroskopiemethoden generieren immense Datenmengen, welche wiederum automatisierte, unvoreingenommene Algorithmen zur Verarbeitung benoetigen. Die Verwendung solcher Algorithmen kann nicht nur das Finden wissenschaftlicher Entdeckungen beschleunigen, sondern auch die Reproduzierbarkeit der damit verbundenen Analysen ermoeglichen.

Die Notwendigkeit der automatisierten Analyse von groen Mikroskopie-Datensaetzen hat zur Entwicklung komplexer Bildanalyse-Methoden gefuehrt. Diese Methoden sind jedoch in vielen Faellen auf den zu untersuchenden Datensatz und die unterliegende Fragestellung zugeschnitten. Aus diesem Grund adressiere ich mit dieser Dissertation das Problem der automatisierten generellen Beschreibung von Zellen in verschiedenen Mikroskopie Modalitaeten. Viele der entwickelten Algorithmen basieren auf kuenstlichen neuronalen Netzen, welche eine vielversprechende Loesung fuer die Extraktion hochqualitativer vielseitiger Eigenschaften aus den zu analysierenden Daten darstellen. Des Weiteren widme ich mich dem Problem der Integration verschiedener Daten-Modalitaeten, um eine detaillierte Beschreibung von Zellen im Kontext des gesamten Organismus zu geben.

Konkret zeige ich zunaechst anhand von zwei Projekten, wie die Verwendung automatisierter Bildanalyse Algorithmen die Charakterisierung von Zellen in Mikroskopie-Daten erleichtern kann. Im ersten Projekt analysiere ich Videos von sich entwickelnden Drosophila-Embryos, um den Unterschied in den Myosin-Mustern zwischen zwei Zellpopulationen mit verschiedenen Endformen zu erklaeren. Im zweiten Projekt entwickle ich einen Algorithmus, basierend auf neuronalen Netzen, um die automatisierte Klassifizierung von Zellen in einer neuen Mikroskopie-Modalitaet zu ermoeglichen. Dies soll dazu dienen, um zu untersuchen, ob die Qualitaet der neuen Modalitaet ausreicht, um Zelltypen in einem Volumen der Hirnrinde einer Maus zu unterscheiden.

Anschlieend stelle ich meinen Teil im Rahmen eines umfangreichen Kollaborativen Projektes vor. Das Projekt zielte darauf ab, einen multimodalen Ganzkoerper-Zellatlas eines dreisegmentigen *Platynereis dumerilii* Wurms zu erstellen, indem hochaufgeloeste Mikroskopie-Daten der Zellmorphologie mit deren Genexpressionen kombiniert wurden. Um in diesem Zellatlas eine vielschichtige Beschreibung von Zellen zu ermoeglichen, habe ich eine Methode entwickelt, mit der Genexpressionen auf kohaerente und entrauschte Weise segmentierten Zellen aus dem Elektronenmikroskopie-Volumen zugeordnet werden koennen.

Schlielich, als mein Hauptprojekt, habe ich mich auf die umfangreiche unvoreingenommene Extraktion von Zellmorphologie Eigenschaften aus dem Elektronenmikroskopie-Volumen von *Platynereis dumerilii* konzentriert. Ich habe einen auf neuronalen Netzen basierten Algorithmus entworfen, der vollkommen unsupervised, d.h. ohne Annotationen, trainiert werden kann. Dieser Algorithmus extrahiert hochqualitative vielseitige Repraesentationen der Zellmorphologie, welche die Gruppierung von Zellen in morphologische Klassen mit charakteristischen Genexpressionen ermoeglichen. Zudem veranschauliche ich weiter, wie diese Repraesentationen von Zellen benutzt werden koennen, um die morphologische Diversitaet von Zellen, Geweben und Organen zu untersuchen.

Acknowledgements

As I look back on the years I have spent in Anna Kreshuk's lab at EMBL, I realise that this thesis is only a minor outcome of my PhD. What is more important is the way I grew both as a researcher and a person, strong and independent. With this I want to thank all people who supported me on this journey, making it an actually enjoyable experience.

First and foremost I want to thank Anna for being an exemplary advisor and a role model for me. My transition from a student to an independent researcher was a smooth and natural process, exactly because Anna patiently guided me when I needed it, but let me follow my own path when I was ready. Besides being an admirable advisor, Anna is also a great person and I truly value all the scientific and nonscientific discussions we had.

My PhD would have probably been a lonely and miserable experience if I were not surrounded by my amazing colleagues. I want to thank Constantin for selflessly babysitting me in the beginning of my journey (and further), Adrian for random mind blowing topics, Fynn for pragmatic work- and life-related advice, Dominik for deep conversations, Tomaz for being the best random fun topic generator, Emil for having heated deep discussions, Maxim for his pure practical vision, Artem and Neha for walking the first steps at EMBL with me, Alex for talks about life, Ricardo for being a source of amusing irony, Qin for just being Qin, Valentyna for taking the responsibility of being the next Ukrainian Valentyna of the lab, Alyona for the amazing stickers, Joris, Paul, Ashwin, Joel, Jonas, Mae and every present and past lab member for great conversations. You all reverse-contributed to my PhD progress by making me want to go to lunch earlier, have longer coffee breaks or drink one beer more, just because it was always so much fun with you. I additionally have to thank Anna for bringing all these amazing people together.

I also would like to thank my collaborators who were also pure pleasure to work with and who helped me better form my research interests. At the very beginning of my PhD I had the fortune of being involved in a big collaborative project of building a whole-animal multimodal atlas of *Platynereis* and this incredible experience made me wonder why not all scientific projects are done as a collaboration of multiple groups. So I am deeply grateful to Hernando M. Vergara, Constantin Pape, Kimberly I. Meechan, Kevin Nzumbi Mutemi, Rachel M. Templin, Yannick Schwab, Detlev Arendt, Christian Tischer and everyone involved for giving me this positive experience. I will never forget the Platy meeting where we spent an hour discussing whether *Platynereis* has a high degree of stereotypy, stereotypicity or stereotypicality.

I want to further thank Maria Leptin, Denisa Gombalova, Sourabh Bhide, Srishti Dar, Timothy Fuqua, Aaron Kuan and all the people I got to discuss and conduct research with. You all contributed to the fact that I still strongly believe that collaborating is the most enjoyable part of doing research.

Convinced of that, later during my PhD I also had a chance to lead my own collaboration and I am greatly indebted to Johannes Hugger, Detlev Arendt and Virginie Uhlmann for helping me transform my young naive idea into a beautiful mature research project. You all showed me that in a scientific environment one plus one actually equals three.

Special thanks goes to Detlev Arendt, who not only took on the responsibility to co-supervise my main project, but also became a true scientific mentor for me during that time. I am incredibly grateful to him for supporting me, believing in me and my ideas and helping me shape my scientific vision and interests.

During my PhD I also came to realise that I truly enjoy teaching, so I would like to warmly thank everyone who shared these rewarding experiences with me and helped me improve my teaching skills. Special thanks goes to all the EMBL Deep Learning course teaching team, including Pejman Rasti, David Rousseau, Szymon Stoma, Simon Norrelykke, Uwe Schmidt and Martin Weigert, as well as the teaching teams of ZIDAS and DL@MBL.

What was also important for me to have a productive, yet enjoyable PhD is the amazing EMBL environment. I am thankful to the EMBL Predoc Office for truly investing time and effort in making our PhD experience smooth and easy and to all the EMBL predocs for being a remarkably supportive community.

Additionally, I would like to thank the members of my thesis advisory committee - Alexis Maizel, Yannick Schwab and Wolfgang Huber - for their support and valuable advice on my projects. I am further grateful to Wolgang Huber and Alexis Maizel for agreeing to review my thesis and to Friedrich Frischknecht and Detlev Arendt for agreeing to serve on my thesis defence committee.

Finally, a big thanks goes to my family and friends for supporting me through my journey. I am extremely grateful to my parents, Lena and Nikolay, and my sister Ania for their unconditional love and support. I am deeply thankful to my amazing husband Johannes for being my happiness, my comfort, my safe haven. I would also like to warmly thank my son Nikita for bringing an unquantifiable amount of joy, and once again thank my mom and my husband for taking care of baby Nikita while I was writing this thesis.

Contents

Chapter 1

Introduction

1.1 Cellular diversity

Life starts at the cellular level. A single cell is necessary and sufficient for all the main attributes of life that include growth and reproduction, conversion of energy and chemical compounds, and reaction to external and internal stimuli [1, 2]. Furthermore, the omnipresence and abundance of unicellular organisms, including archaea, bacteria, protists and yeasts, show it is in fact a successful strategy to have one single cell independently fulfil all the above mentioned functions. Nevertheless, groups of cells that briefly or permanently depend on each other have repeatedly appeared across the main branches of the tree of life [3, 4, 5].

Making such multicellularity permanent allowed cells to progressively specialise in fulfilling a particular function, giving origin to increasingly diverse cell types. For example, choanoflagellates, the closest living unicellular relatives of animals, have only one cell type that can switch between two forms - flagellate and amoeboid - that are presumed to give rise to epithelial and crawling cells in animals [6]. The epithelial cell family in sponges already shows gene regulatory modules necessary for actomyosin contractility and photo- and mechanosensation, suggesting that they might further differentiate into sensory cells and myocytes [7]. The latter group further diverges into smooth and striated muscles in bilaterians, with smooth muscles regaining striation proteins to form cardiomyocytes in vertebrates [8].

Such progressive functional diversification generated organisms comprising tens or even hundreds of cell types. Still for every single individual all this diversity emerges from one cell only - a fertilised egg, or zygote. A series of cell divisions will eventually transform a zygote into an entire organism composed of up to trillions of different cells. All these cells share the same genome, still under the influence of internal and/or external factors they will end up expressing different genes: their transcriptomes will be slowly diverging, reflecting their specialisations. This differential gene expression will in turn lead to the production of different proteins, lipids and other metabolites, which will cause cells to have

different morphological and physiological characteristics and traits, generally referred to as cellular phenotypes. Such phenotypes, combined with cellular position will ultimately define cell functions in the organism.

1.2 Describing morphological cell types with microscopy

The first ever attempt to describe cells was made in 1665 by Robert Hooke, who observed a piece of cork under a self-made microscope and noticed regular structures that resembled a honey-comb. Soon after, another scientist working on improving microscopes, Antonie van Leeuwenhoek, for the first time described unicellular organisms. Centuries afterwards microscopy would remain the main method of characterising cells, and cells would be classified and even named based on their morphology and position.

At the moment, constantly improving light microscopy (LM) methods allow for diverse ways of describing cellular phenotypes (reviewed in [9]). In transmission light microscopy, used by Hooke and Leeuwenhoek, the image is formed when a sample absorbs the light passing through it. However, most animal cells do not absorb enough light, so the resulting image has relatively low contrast. Still, so far it is the easiest and the least phototoxic way of imaging cells, so this low contrast is often compensated by powerful image post-processing algorithms, making the method widely applicable for various medical screening applications [10, 11, 12].

A big milestone for imaging biological samples has been achieved by using fluorescent labels to target specific molecules of interest inside living cells. This allowed us, for the first time, to directly investigate not just visible structures, but also molecular characteristics of cells, such as subcellular localisation of various proteins and their interactions [13, 14, 15]. Another breakthrough was the development of light-sheet microscopes that use a thin sheet of light to illuminate the sample, drastically reducing phototoxicity, thus allowing for much longer imaging periods. This method is widely used for describing cellular morphology and dynamics during organism development [16, 17, 18, 19, 20]. A further major step forward was breaking the diffraction limit of light by superresolution methods [21, 22, 23, 24, 25] that improved the ability to observe molecules directly under the microscope. These methods enabled, for example, better characterization of chromatin organisation [26], centrioles [27] or the nuclear pore complex [28].

While such fluorescent light microscopy methods allow following specific molecules inside living cells with unprecedented resolution, their main limitation is the necessity to label molecules of interest. In order to obtain a detailed view of all the cellular structures at once the method of choice is electron microscopy (EM) that uses electrons instead of light, enabling much higher resolutions. The image is generated by electrons either transmitted through a thin sample (transmission electron microscopy, TEM) or reflected from the sample surface (scanning electron microscopy, SEM).

Soon after the first electron microscope was produced by Ernst Ruska and Max Knoll in 1931, it became a widely adopted method for examining biological tissues and cells [29,

30]. Many cell types have been firstly described by scientists looking at EM samples [31, 32, 33]. Recent advances in instrumentation, sample preparation and postprocessing methods allowed acquiring whole volumes of tissues, organs or even smaller organisms in 3D [34]. For the first time this gave access to fine 3D ultrastructure of cells at nanometer resolution, allowing detailed morphological characterisation of membrane-bound and membraneless organelles [35, 36, 37]. Furthermore, such resolution in 3D allows researchers to study cellular neighbourhoods and interconnections, which is essential for describing cell types where such connectivity also defines their function, e.g. neurons. Overall, methods that profile both high resolution ultrastructure and interactions with other cells are crucial to determine how cells function separately and as part of the whole organism.

1.3 Describing molecular cell types

The rise of single-cell transcriptomics has revolutionised the way cell types are defined [38]. High-throughput methods for profiling gene expression have shown that there is much more cellular diversity within an organism than we previously expected based on morphological appearance only. This led to discovering multiple new cell types in well-studied tissues and organs, as well as in non-model organisms [39, 40, 41, 42]. The transcription information can be further supplemented by profiling chromatin modifications, protein expression and metabolite composition, generating an even more detailed molecular view of a cell.

However, at this level of precision it is not clear which differences truly define cell types and which ones rather characterise temporary cell states. Consequently, there is no generally accepted definition of cell type. Some researchers would define it based on the origin of cellular diversity - complexes of transcription factors that drive cellular gene expression programs, while the others would take into account its end goal - varying functions fulfilled by different cell types [43, 44, 45, 46, 47]. Nevertheless, regardless of which approach one chooses, it is clear that in order to get a comprehensive understanding of how different cells function, one needs to consider both the molecular and the anatomical description of a cell [48].

While single-cell omics technologies can describe cells in precise molecular details, the information about cell location, morphology and connectivity is lost due to the necessary step of tissue dissociation during sample preparation. To alleviate this problem, spatial techniques for large-scale molecular profiling have emerged recently. Reviewed in more detail in [49, 50, 51], these methods can now offer measuring the expression of thousands of genes in a tissue at subcellular resolution while retaining information about the cell location and phenotype, as visualised by a light microscope [52, 53]. At the moment the application of such methods is still limited to tissue slices, while whole-body studies of smaller model organisms typically rely on whole mount *in situ* hybridization [54]. This technique can provide cellular resolution gene expression in 3D, however is still limited to a few genes of interest at a time. Therefore, extending the available methods to probe more genes in bigger volumes is necessary to describe the variety of cell types on a large scale.

1.4 Automating image-based cell type descriptions

The exponential increase in throughput of methods for both anatomical and molecular characterisation of cells leads to the generation of massive amounts of data. The field of 'omics' could never rely on manual data analysis, thus developing more precise and efficient pipelines for fully automatic cell typing has constantly been of high priority. Analysing microscopy images, however, has been historically done manually. In recent years multiple computational solutions emerged, nevertheless many types of analysis are still performed fully or partially by hand.

Developing automated methods for image analysis is of utter importance for two main reasons. Firstly, manual examinations are often highly time consuming, making it unfeasible to analyse, for instance, big EM volumes or extensive light-sheet microscopy movies. Secondly, even when working with smaller amounts of data, manual analysis is highly prone to research bias. This includes not only confirmation bias, when a researcher unconsciously skews the analysis to confirm the desired hypothesis, but also simply focusing on the parts of data that are more likely to be important for a given question, thus potentially missing some interesting observations. For example, a researcher might acquire a whole EM volume of a smaller animal to study nervous system. However, while focusing on neurons only they might miss some rare cells of other types. Therefore, with more and more imaging data being acquired, developing automated methods for analysing cells in microscopy volumes is essential for systematic unbiased investigation of cellular diversity.

In order to quantitatively characterise cells in an imaging volume, one first needs to delineate the cells present in the data. This is the task of cell segmentation, that aims to find which pixels of the image belong to which cell. The emergence of ever increasing volumes necessitates efficient automatic segmentation pipelines. A substantial improvement to the accuracy of such pipelines has been brought by the adoption of neural network based methods, allowing segmenting and analysing cells on a previously unimaginable scale. For example, automatic segmentation of the whole Drosophila central brain helped researchers identify around 25,000 neurons and trace the connections between them that led to a better understanding of how specific fly behaviours function on a cellular level [55]. For time-resolved data, segmenting and subsequent tracking of cells in light-sheet microscopy videos of early mouse development enabled researchers to statistically quantify dynamics of tissue morphogenesis [19].

While most of the modern pipelines for automatic large-scale morphology characterisation necessarily start with cell segmentation, further analysis steps vary to a greater extent. This arises from the fact that the very definition of morphology in microscopy studies heavily depends both on the type of acquired data and on the question a researcher aims to address. In fluorescence microscopy cellular morphology can only be quantified for the labelled structures, most commonly, cell membrane and nucleus. For example, Wu et al. [56] use cell and nucleus segmentation of breast cancer cells to measure more than 200 morphological properties, which define distinct cell classes predictive of tumorigenic and metastatic potentials. Electron microscopy gives researchers a chance to focus on a

broader range of morphological characteristics. For example, in order to investigate energy management in phytoplankton, Uwizeye et al. [57] focus on morphological features of plastids, mitochondria, nuclei, vacuoles and various energy storage structures. Another interesting case of image-based cell characterisation is classifying neurons that historically required additionally describing their branching patterns [58, 59].

The multitude of different ways to describe cellular morphology led to a great diversity of manual and automatic pipelines for analysing cell imaging data. Such pipelines are often dataset-specific, so addressing a new question often requires designing a completely new pipeline. Thus, to facilitate image-based cell typing more general methods are needed that are easily transferable between different datasets. A potentially promising solution is the application of neural networks that have emerged in the last decade as a one-fits-all solution for natural image processing. Designed to mimic the way our brain learns to recognise different objects, they are able to automatically identify features important for a specific task at various levels of detail. Particularly suitable for describing fine-grained phenomena in big and complex data, neural networks have already been successfully used for automating various biological image analysis problems [60, 61, 62], so adapting them for characterising cellular morphology is likely the sought-for next step for automating cell type investigation in imaging data.

1.5 Thesis overview

This thesis begins to address the problem of automating the description of cells in various imaging modalities. I will describe different types of imaging data used to investigate cell differences and show how computational methods could be used to automate analysing such data. In particular, I will focus on neural networks as a generally applicable state-of-the-art method for image analysis. Moreover, I will show how different data modalities describing anatomical and molecular cell characteristics can be integrated, and further discuss generating unbiased many-sided representations of cells.

To start with, in Chapter 2 I introduce a small example of how I automated image processing to find the underlying cause of differences in cellular behaviour. During the first stage of gastrulation in the fruit fly, initially similar cells of developing mesoderm either expand or constrict based on their position. However, it is still not well understood, how exactly this binary fate is triggered. I analysed myosin dynamics in a light-sheet microscopy movie of Drosophila gastrulation [63], aiming to detect and quantify differences in myosin expression that could prompt such divergence in cell shape. I showed that there is no obvious difference in myosin patterns between the two populations of mesoderm cells, but their contrasting fates can rather be explained by the influence of their neighbourhoods.

In Chapter 3 I describe a small pilot project aimed at evaluating how well a new intermediate-resolution imaging modality captures cellular morphology. X-ray holographic nano-tomography (XNH) was developed to bridge the gap between light and electron mi-

croscopy and offers substantially higher resolution than LM methods, but a bigger field of view than EM imaging, that is often necessary for neural circuit discovery [64]. In order to quantify how prominent cell type specific morphology is in this type of data I set up a neural network cell classification pipeline and applied it to the mouse cortex volume acquired by XNH with annotated cell types from [64]. I showed that the data quality is sufficient to automatically classify brain cell types with high accuracy even without prior cell segmentation, concluding that XNH imaging is suitable for differentiating morphological cell types in the brain.

In Chapter 4 I present an extensive collaborative study aimed at integrating high-resolution morphology and gene expression data for an entire annelid worm on the single-cell level. In order to generate a more detailed description of different cell types, an EM volume of the whole animal was acquired, fully segmented and aligned with multiple whole-body gene expression maps [65]. As a part of this collaborative effort to generate a multimodal cellular atlas of the whole animal, I improved gene expression assignment by designing an automatic way to assign each segmented cell a coherent denoised gene expression profile. Such profiles could be further used to integrate single-cell sequencing data into the atlas, paving the way to elucidate the interconnection of gene expression and phenotype.

Finally, in Chapter 5 I describe my main PhD project aimed at designing a pipeline for automated unbiased extraction of comprehensive cellular morphological representations from EM data. Cell morphology has been mostly described using a limited number of manually defined features. However, such features are mostly restricted in the scope of visual characteristics they can represent. As a part of the effort to generate exhaustive morphological descriptors, I developed a pipeline for unsupervised deep-learning based extraction of texture representations from cells segmented in EM data. I applied the pipeline to the EM volume of *Platynereis dumerilii* and showed that these features, supplemented by similarly extracted shape features [66], quantitatively outperform a broad set of manually defined descriptors on the tasks of finding similar cells and distinguishing cell classes. I further show how these morphological representations capture visual similarities between cells, allowing to group them into meaningful morphological classes consistent with gene expression. Finally, I demonstrate how using such unbiased descriptors can assist exploring the whole-animal volume, discovering rare cell types and tissues.

I conclude this thesis with Chapter 6, where I briefly summarise my contributions and discuss the next steps to be taken to achieve the ultimate goal of fully automatic unbiased complete description of cell types.

Chapter 2

Characterising shape changes of mesodermal cells during Drosophila gastrulation

2.1 Background

Describing cells mostly implies investigating their molecular and morphological phenotypes. However, the way cells develop and function is also heavily influenced by their environment. This is especially prominent during embryo development, when individual cell fates are often defined by complex interactions between different cell populations. In this chapter I will focus on analysing shape fates of mesodermal cells during the early stages of *Drosophila* gastrulation imaged by light-sheet microscopy. I use computational analysis methods to show that the cell shape change during this process is regulated by a combination of molecular determinants and mechanical interactions between neighbouring cells.

2.1.1 How cell fate is decided

The development of most multicellular organisms starts with one cell only. The cell repeatedly divides and eventually its descendants differentiate to form the astonishing diversity of cell types that we see in complex organisms. However, the way initially identical cells choose astonishingly different cell fates is not trivial to describe. This divergence is caused by differing genetic programs that can be triggered by a multitude of factors, or even their combinations, that could be both intrinsic (e.g., cytoplasmic determinants) or extrinsic (e.g., morphogen gradients, cellular interactions). Moreover, these factors are hardly ever binary, meaning that cellular differences mostly arise not from the factor being fully present in one cell and fully absent in the other one, but rather from a small variance in the factor magnitude in different cells. Such fate decisions, provoked by subtle fluctuations, are

commonly referred to as symmetry breaking and can be observed, for example, during the establishment of the anterior-posterior body axis as a result of the *Wnt*-proteins gradient [67]. Many developmental processes based on symmetry breaking have been studied in detail and the triggering factor is already known, however, describing how the continuous factor translates into discrete outcomes is often far from straightforward.

2.1.2 Cells change their shape differently during *Drosophila* gastrulation

An example of such a well-studied, yet not fully understood process is the shape change of mesodermal cells during *Drosophila* gastrulation. Gastrulation is a process that transforms an early embryo, that is a one layer sphere of epithelial cells, into a gastrula with three layers. These three layers, ecto-, meso- and endoderm, will further give rise to all the organism tissues. In *Drosophila* this transformation starts in the lower (ventral) part of the embryo that contains prospective mesoderm. The mesoderm epithelium invaginates along the midline to form a ventral furrow. This inward folding is initiated in the central rows of mesoderm cells that accumulate substantial amounts of myosin in their apical parts. Afterwards, a series of pulsed actomyosin contractions in these cells causes them to reduce their apical surfaces. To preserve tissue integrity and compensate for this constriction the apical surfaces of the immediate neighbouring rows of lateral mesodermal cells expand accordingly [68, 69, 70, 71].

While the major molecular details of this constriction have been well-studied, what is still not clear is how the differences between the constricting and expanding mesodermal cells arise. Characterising genetic differences does not fully explain the binary fate decision: these cells seem to follow a similar developmental program and express the same genes, albeit in different quantities. The expression of many genes involved in myosin activation, such as *t48*, *fog*, and *mist*, shows a dorsoventral gradient [72] that leads to a smooth gradient in myosin concentration. The central mesodermal cells, located in the most ventral part of the embryo show the highest levels of contractile actomyosin, are the first ones to contract and have the highest degree of contraction. All these three characteristics gradually decrease in mesodermal cells in the dorsal direction. Given that, it is clear that myosin concentration is highly important for apical constriction, however, the mechanism behind the lateral mesoderm cells expanding their apical area is yet to be understood.

2.1.3 Myosin concentration alone can not explain differences in shape

As a part of the project aimed at a better understanding of the influence of actomyosin networks on constriction and expansion patterns [63], Sourabh Bhide and Denisa Gombalova imaged the apical surface of the *Drosophila* embryo ventral side during the time when the central mesodermal cells constrict. Sourabh Bhide segmented the cells in these videos and quantified cellular myosin concentrations. The analysis confirmed that at each point of time the concentration of myosin is higher in constricting cells than in the ex-

panding ones. Moreover, the rows with higher concentration of myosin - the central ones - are the first ones to constrict. However, he also observed that the expanding cells are reaching similar myosin concentrations, albeit at later stages of the process. Still, the levels of myosin that triggered constriction in the central rows do not seem sufficient to achieve a similar effect in the lateral mesoderm that expands nevertheless. This shows that although at each single time point one could correlate the constriction level of a cell to its myosin concentration, it is generally impossible to define a concentration threshold that a cell needs to reach in order to start constricting.

2.1.4 Project aims and contributions

My aim was to investigate which differences could be detected between the cells with the same concentration of contractile meshwork, that could explain the diverging fates of otherwise similar cells during the process of ventral furrow formation. More specifically, I wanted to analyse the imaged mesodermal cells to find intrinsic or extrinsic characteristics potentially affecting the cell shape changes. First, I was interested if any visual difference between the myosin loci exists and could be identified. Second, I wanted to examine if cellular environment affects the shape changes as well.

The analysis described further was performed on the above-mentioned data acquired by Sourabh Bhide and Denisa Gombalova as a part of [63] and has been included in the paper. The hypotheses that guided the analysis were generated by Maria Leptin, Sourabh Bhide and myself.

2.2 Results

2.2.1 Difference in myosin patterns

The first possibility that I wanted to explore is that given the same myosin concentration the actomyosin meshwork might assemble differently in constricting and expanding cells. In this case I expected to see some visual differences between the myosin loci in these cells. To be able to automatically recognize such differences I referred to convolutional neural networks (CNNs).

Neural networks are known to be strong feature extractors for images, with convolutional layers designed to capture local texture patterns. So in order to see whether there is a visual difference in myosin between the cells of different fates, I decided to train a network to distinguish between these two types of cells. I expected that if such a difference is noticeable, the network would be able to successfully predict if a cell will constrict or expand based on the myosin distribution of the cell. To make sure such predictions are based on the distribution only, and not the concentration, I planned to use the cells with the same myosin concentration, meaning I would take constricting cells from earlier time points and expanding cells from later ones.

The biggest advantage of using CNNs for this type of problems is that one does not have to manually define any visual descriptors, as the network decides itself which features might be important for a given task. This means we can refrain from defining what the potential difference might be and just let the network find it on its own. However, this comes at the price that the features extracted by the network are not directly explainable. Thus, it would not be straightforward to confirm whether the differences the network has found are indeed related to the myosin pattern in the cells.

For training I used myosin channel only. However, I noticed that it is still possible to infer cell shape from myosin distribution, and this could potentially give the network an easier cue of what shape changes await a certain cell. That is why for all the experiments I additionally performed a 'negative control' - I trained the network in the very same setting using cell segmentation masks instead of the myosin channel. Then, if the network can predict the cell fate not only from the myosin distribution, but also from the cell mask, I would not be able to claim there is any difference in myosin patterns.

Additionally, I considered that using the binary labels 'constricting' or 'expanding' might not be informative enough, because similarly to the myosin gradient, constriction also follows a gradient, with more central cells constricting earlier and stronger than the more lateral ones. That is why I used the row numbers defined by Sourabh Bhide as labels. The enumeration starts with row 1 of the most central cells and ends with the last layer of mesoderm cells of row 8. During gastrulation rows 1-6 gradually constrict and rows 7 and 8 expand (Figure 2.1). Due to the imaging angle only the rows 3 to 8 could be segmented, so I used labels 3-6 for constricting cells and 7-8 for expanding ones.

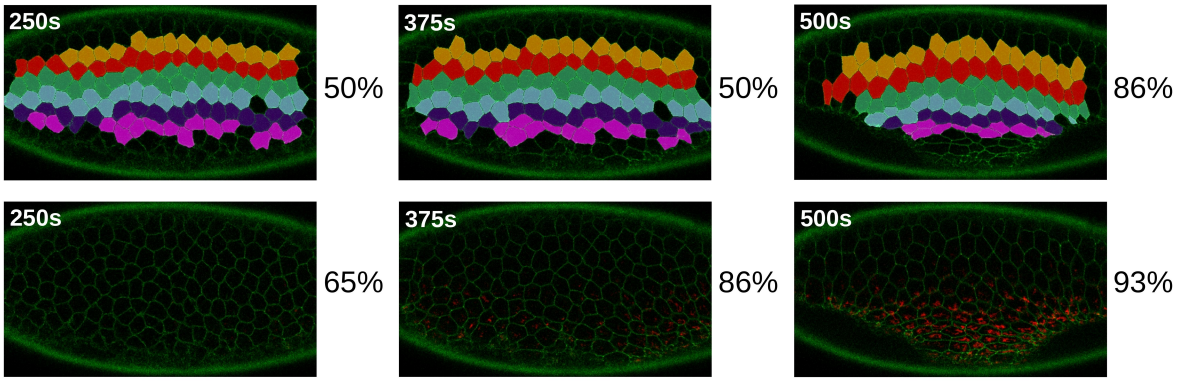


Figure 2.1: Training data and prediction accuracy for predicting the row number for cells of the same time point. Upper row: cell segmentation masks, lower row: myosin. The color coding of the rows is the following. Constricting rows: 3 - magenta, 4 - indigo, 5 - blue, 6 - medium green, expanding rows: 7 - red, 8 - orange. During the training each cell from the image was fed to the model separately. The prediction accuracy on the right of each figure shows the percentage of correct predictions of constricting vs expanding rows (see Methods for more details).

In order to show that the network setup is functional, I tried to predict the row numbers in the cells of the same timeframe. In the early timepoints of the video one can not notice any difference neither in myosin pattern or concentration, nor in cell shape yet. Over time the lower constricting rows start gradually accumulating higher concentrations of myosin, but the cell shape change comes with a delay. This is clearly reflected in the network predictions (Figure 2.1): at the early timestep of 250s the row number can hardly be predicted from both cell masks and myosin images. At the later timestep of 375s the shape is still not predictive, but the differences in myosin concentration already let the network differentiate the rows. Finally, at the late timestep of 500s the network performs well on both myosin and shapes. Not only this showed that the network is able to learn the connection between myosin concentration and constriction patterns whenever they exist, but also that it does not extract any myosin-unrelated cues from the early timepoints.

After I verified that the network is behaving as expected, I turned to the main aim of the experiment - to examine if there are any differences in myosin pattern between the constricting and expanding cells of the same myosin concentration. To make sure any pattern differences are already well-pronounced, I selected the highest myosin concentration present in a sufficient number of both constricting and expanding cells (Figure 2.2). The network scored the accuracy of 61% on cell shapes and 65% on myosin. This shows that the network struggled to learn any differences and most of its performance can be attributed to cell shape features rather than to the myosin patterning.

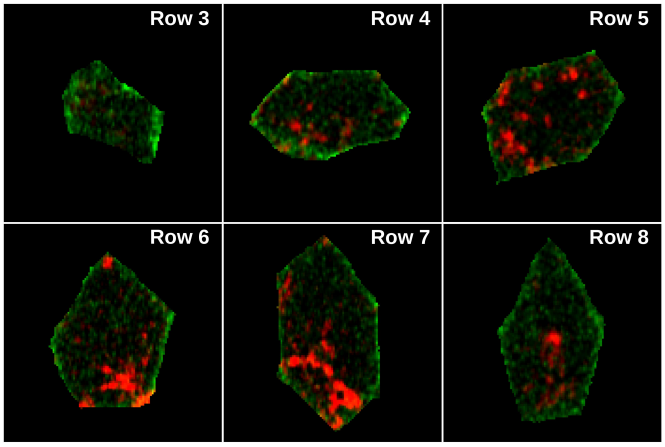


Figure 2.2: Training data for predicting the row number for cells with the same myosin concentration. Examples of cells from different rows are shown.

It is important to stress out that these negative results can not be unambiguously interpreted as no difference in myosin pattern. This outcome, however, shows that even if a difference exists, it is not visually obvious or could not be captured by the imaging methods used.

2.2.2 Neighbourhood influence

Not finding any differences in myosin pattern, I aimed to investigate the influence of cellular neighbourhoods. An observation that row 7, immediately adjacent to the constricting cells, expands more than the following row 8 led me to the hypothesis that the inability to constrict comes not from certain properties of the contractile network of the cell, but rather from the high pulling forces acting on it from the neighbours that started constricting earlier. This hypothesis could also explain why, when detached from their constricting neighbours, the expanding rows regain the ability to constrict [73]. If this hypothesis proves true, that would mean that the shape change of a cell depends on the concentration of myosin both inside the cell and in its immediate neighbourhood. Thus, I aimed to correlate the shape change to the myosin concentration inside and outside of each cell.

Again I decided to refrain from using the binary labels 'constricting' and 'expanding' and calculated the cell shape change as the area of a cell in the next time step divided by the area in the given timestep. To calculate myosin concentration outside of a cell I first considered calculating the average concentration of all the cell's immediate neighbours. However, I decided against it, since I expected that myosin loci further away from a given cell might influence it less then the ones closer to it. Thus, given that neighbouring cells might have different radii, such an approach could give an incorrect estimate of the myosin that actually influences the cell. That is why, to calculate the outside concentration I took a radius of 70 pixels (8.5 µm, approximately, one cell diameter) around the cell and divided the sum of myosin intensity in this area by the size of the area (Figure 2.4A).

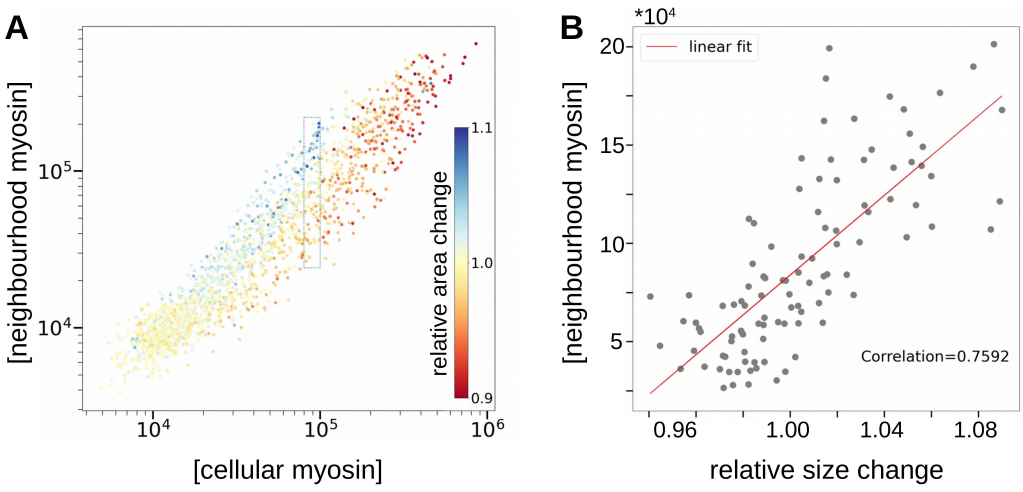


Figure 2.3: Size change depends on both cellular and neighbourhood myosin concentration. **A.** Myosin concentration inside a cell plotted against the cell's neighbourhood myosin concentration. Color encodes the change in cell size over two consecutive time points (red - constricting, blue - expanding). **B.** The neighbourhood myosin concentration plotted against the relative size change for a subset of cells boxed in A.

Finally, I plotted the concentration of myosin outside the cell vs inside the cell and colored it by cell shape change (Figure 2.3A). The plot clearly shows that given the same myosin concentration inside two cells, a cell tends to constrict more if its neighbours have lower myosin concentrations. More specifically, for cells with similar myosin concentration (around 10^5 units), the correlation between the neighbourhood myosin and the cell shape change is around 0.76 (Figure 2.3B). This strongly suggests that the cell shape fate is indeed influenced by the myosin concentration of the cell neighbourhood.

2.2.3 Calculating myosin offset

Given the demonstrated correlation, Maria Leptin expected that the pulling force of myosin in constricting cells might lead to displacement of myosin loci in neighbouring cells. If this holds true, we would expect constricting cells to pull neighbouring myosin stronger towards their membrane than the expanding ones. To check whether it is the case, for each cell in every timeframe I calculated the 'neighbourhood myosin offset' - a value that shows the average distance of surrounding myosin from a cell - with lower values indicating cells where surrounding myosin is located closer to the cell and vice versa. This was achieved by weighing the neighbourhood myosin, as defined before (Figure 2.4A, area between the two yellow lines), by the distance from the cell boundary (membrane) (Figure 2.4B). Plotting this value against cellular myosin concentration and colouring it by cell size change we see that indeed given the same myosin concentration the constricting cells pull myosin loci around them closer to their cell membrane than the expanding ones (Figure 2.4C).

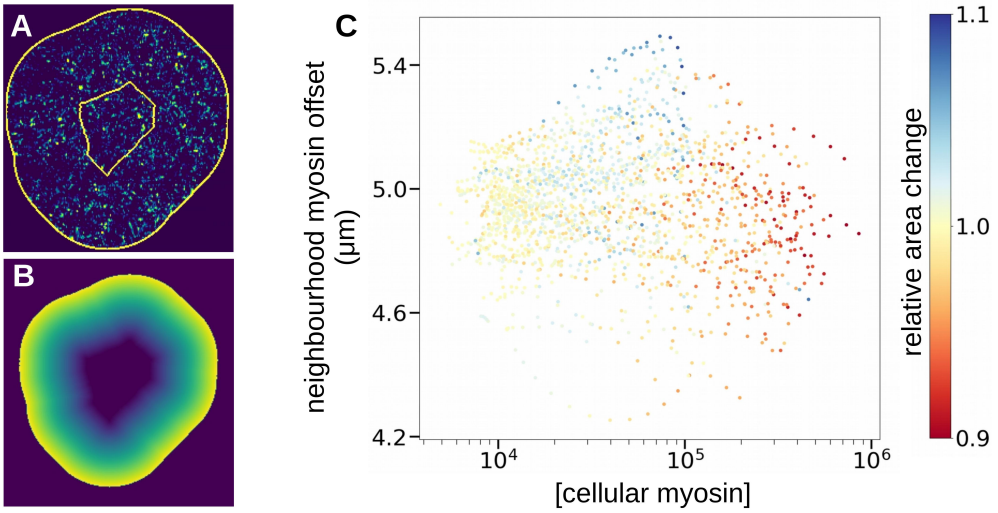


Figure 2.4: Cellular neighbourhood and myosin offset. **A.** I calculated neighbourhood myosin in a 70-pixel radius area around the cell (between two yellow lines). **B.** Distance from the cell membrane that I used to weigh neighbourhood myosin to calculate myosin offset in cell neighbourhood. **C.** For each cell I related its neighbourhood myosin offset to the concentration of cell-internal myosin.

To complement this, I also calculated myosin offset within each cell as the euclidean distance between the centre of mass of a cell and the centre of mass of the myosin expression within the cell. This also shows that bigger, expanding cells have their myosin loci further away from their centre than smaller, constricting cells (Figure 2.5). From this I could conclude that constricting cells could indeed exert a pulling force on the myosin loci of their neighbours, preventing their constriction.

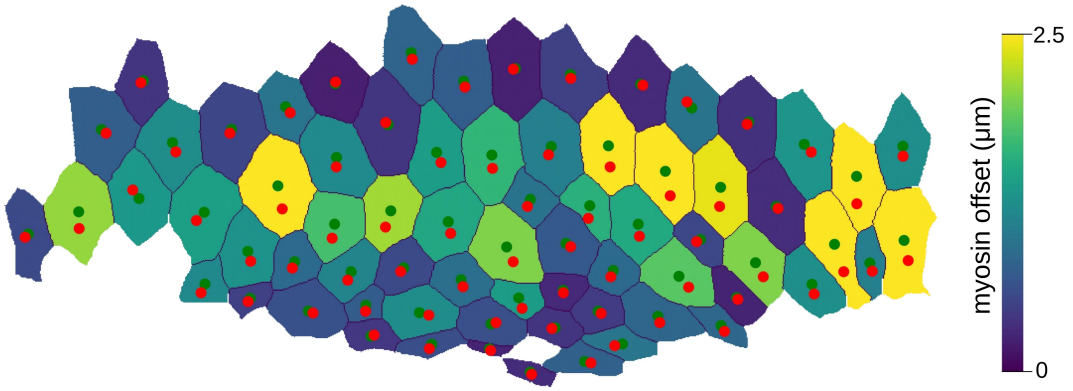


Figure 2.5: Internal myosin offset. All segmented cells of the embryo at 550s are shown coloured by the degree of myosin offset inside a cell. The centers of mass of each cell and its myosin intesity are shown as a green and red dot, respectively.

2.3 Discussion

2.3.1 Contributions

In this chapter I showed how differences in cell shape arise at the interplay of intrinsic and extrinsic factors and how hypothesis-driven automated image analysis can assist in elucidating and confirming these factors. More specifically, I have shown that constricting and expanding mesoderm cells might not have differences in the pattern of contractile meshworks, but rather take different fates due to the neighbourhood interactions. To prove this, I have detected a correlation between myosin concentration inside and outside the cell and cell area increase or decrease. Additionally, I have observed and quantified the changes in location of myosin loci caused by pulling forces of the neighbouring constricting cells.

2.3.2 Further development

While the results of the conducted analysis are unambiguous, more accuracy could still be gained in describing cellular shape and myosin concentration changes. Firstly, more

samples could be acquired to give the results more statistical significance. More data would also allow for more precise quantification of myosin concentrations inside and outside of a cell needed to trigger constriction or expansion.

Additionally, a better modelling of changes in both cell shape and myosin distribution could be achieved by analysing the original 3D data instead of its 2D views. The biggest obstacle for such analysis is segmenting cells in 3D that, even if assisted by the state-of-the-art deep neural networks, would require a substantial time investment to generate training data and to proofread the resulting segmentations. Still, obtaining 3D segmentation would allow us to analyse the role of the overall myosin distribution in a cell, not just the apical myosin. Similarly, this would facilitate describing true cell shape changes, rather than the mere changes of the cell's apical surface.

Chapter 3

Classifying cells imaged by a new intermediate resolution modality

3.1 Background

In the previous chapter I showed how cellular differences can be described by analysing fluorescent microscopy data. While such type of data is sufficient to inspect cellular shapes and examine one or a couple of proteins of interest, describing fine cellular morphology requires higher resolution than a light microscope can provide. Although many efforts are put into breaking the light diffraction limit, most studies that require characterising cellular ultrastructure or tracing thin cytoplasmic extensions are done using electron microscopy. This technique has recently seen a rapid increase in both resolution and throughput, but comes with with high costs and long acquisition times that substantially limit the size of samples that can be imaged. In this chapter I focus on a newly developed imaging method of intermediate resolution and, as a part of a pilot project, I investigate how well it can describe brain cells based on their morphology.

3.1.1 The need to bridge the gap between light and electron microscopy

Imaging biological samples requires a trade-off between imaging scale and resolution. Traditionally, the task of describing fine morphology of a limited number of cells has been solved by electron microscopy, while light microscopy was used to image bigger samples, sometimes over the course of hours or days, yet with limited resolution. However, many biological problems also require intermediate imaging techniques allowing for bigger samples than EM, but higher resolution than LM. This gap has been mostly bridged by X-Ray imaging, with micro-CT scanners being increasingly often used as a fast and simple non-destructive imaging method [74, 75]. This allowed, for example, examining the quality of samples prepared for long EM data acquisition [55] or targeting specific regions of the animal volume for further EM imaging [76].

Another example of an application that would greatly benefit from such intermediate resolution and scale imaging is the field of connectomics that aims to map all the neural connections in a brain [55, 77, 78]. Tracing axons and dendrites (neurites) and detecting their connections (synapses) requires resolution higher than light microscopy can currently offer. However, electron microscopy, which is currently the leading method in this field, heavily restricts the field of view - size of the areas that could be imaged in block, due to the high cost and time required for imaging. Thus, X-Ray imaging could be a promising direction for characterising neural circuits. However, so far the limited contrast of the method did not allow for imaging of label-free brain regions.

3.1.2 X-ray holographic nano-tomography for fast imaging of thick samples

To address this limitation Kuan et al. [64] proposed a new way of high resolution imaging by adjusting X-ray holographic nano-tomography (XNH) for imaging thick densely-stained brain samples. The authors improved both the contrast and the field of view, achieving sub-100-nm resolution on volumes of central and peripheral nervous systems of mouse and adult *Drosophila*. More specifically, they have shown that it is possible to achieve resolution below 90nm with the field of view bigger than 50µm and resolution below 200 nm with the field of view around 250µm. They also showed a possibility of imaging a whole *Drosophila* leg - a body part that is difficult to physically section, thus unsuitable for EM imaging. Stitching together multiple XNH scans of the leg, the authors obtained a volume over 1.4mm long - a size unreachable for the current EM methods.

With regards to describing neuron morphology, Kuan et al. showed the possibility to identify and automatically segment neuron somas and larger neurites in XNH data. Smaller neuron processes would still require higher resolution (5-10nm) EM imaging to be resolved, so the authors also demonstrated a way to target smaller regions of interest, e.g. areas with synaptic connections, using EM. They have additionally used the acquired EM volumes to confirm that neurons in both datasets have similar morphological features and tracing of larger neurites in the XNH volumes is reliable. The conducted work showed that while reconstructing neural connections still requires EM imaging, applying the fast, intermediate-resolution XNH imaging as the first step substantially increases the size of neuronal tissue areas that can be simultaneously analysed. Thus, such a hybrid approach can considerably simplify and accelerate describing neuronal types and their interplay.

3.1.3 Project aims and contributions

To further investigate the applicability of this promising approach to various neuroscience problems, Aaron T. Kuan, Alexandra Pacureanu, Cornelius Gross, Anna Kreshuk and myself have drafted a pilot project aimed to explore how well different cell types can be told apart at such levels of resolution. More specifically, we sought to investigate the possibility of automatically classifying cells in the lower-resolution (130nm) XNH volumes

of the mouse cortex from the study (Figure 3.1). This resolution and the volume contrast were not sufficient to automatically segment neuron somas, let alone neurites. Thus, I wanted to check if, even in the absence of segmentation, cellular morphology at this level of resolution is descriptive enough to predict the cell types.

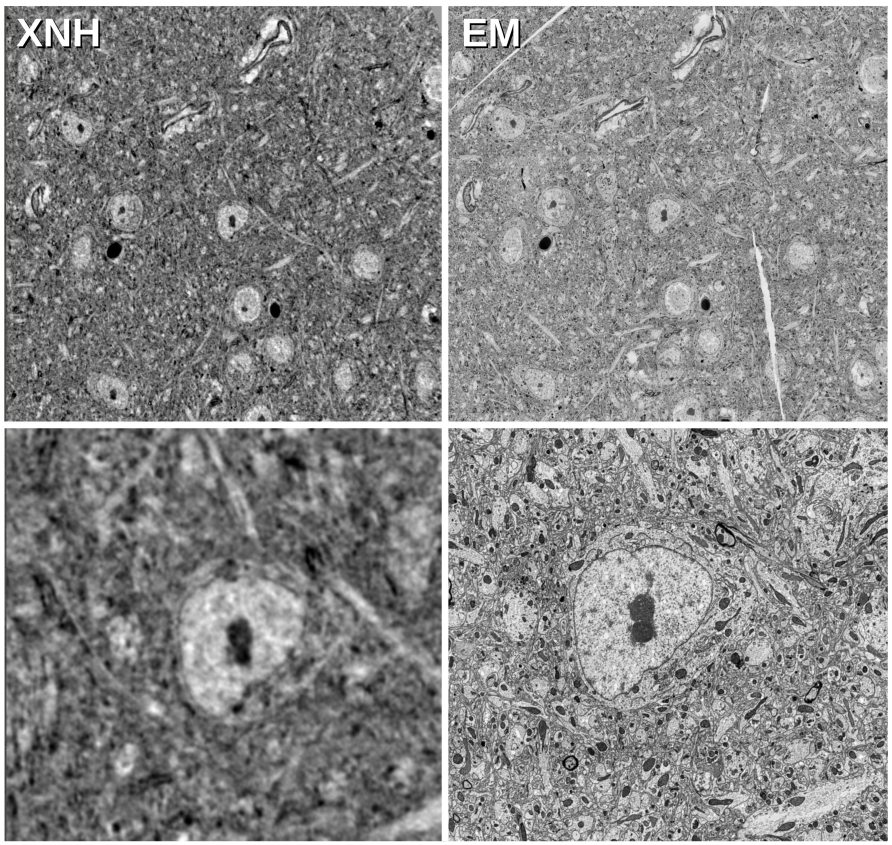


Figure 3.1: Comparison of the image resolution of the mouse cortex volume acquired by XNH (X-ray holographic nano-tomography, left) and EM (electron microscopy, right). The lower images show zoomed in central regions of the upper images.

For this Aaron T. Kuan, Jasper S. Phelps and Julie Han manually annotated cells in the XNH volume of the mouse posterior parietal cortex, assisted, whenever available, by the aligned EM volume. My goal was to see if by using these manual labels I could train an automated algorithm to distinguish between these cell types by solely relying on the XNH modality. The full description of the data and its acquisition can be found in [64]. The project was designed as a pilot project without the aim to publish the results.

3.2 Results

3.2.1 Designing a neural network classification pipeline

To see whether the data resolution is good enough to automatically tell the cell types apart, I decided to train a neural network using existing annotations as a source of supervision. Convolutional neural networks have been proven to be strong algorithms for image data classification, so I have chosen a relatively shallow network ResNet-18 [79]. Having comparatively few parameters, it can be trained on a modest amount of data, yet shows good performance across multiple image datasets.

The available annotations contained four classes: pyramidal neurons, non-pyramidal neurons, non-neuronal cells and unclassified cells (Figure 3.2). However, after inspecting the latter class I decided not to include it into the training process due to the following reasons. Firstly, some of the cells in this category were repeatedly confidently assigned to one of the other classes by multiple of my trained networks. Manual verification done by Aaron T. Kuan showed that in most cases this assignment was correct. Secondly, many of the cells in this category were located on the border of the acquired XNH volume (Figure 3.2, lower right) that led the network to predict any cell on the volume border as unclassified.

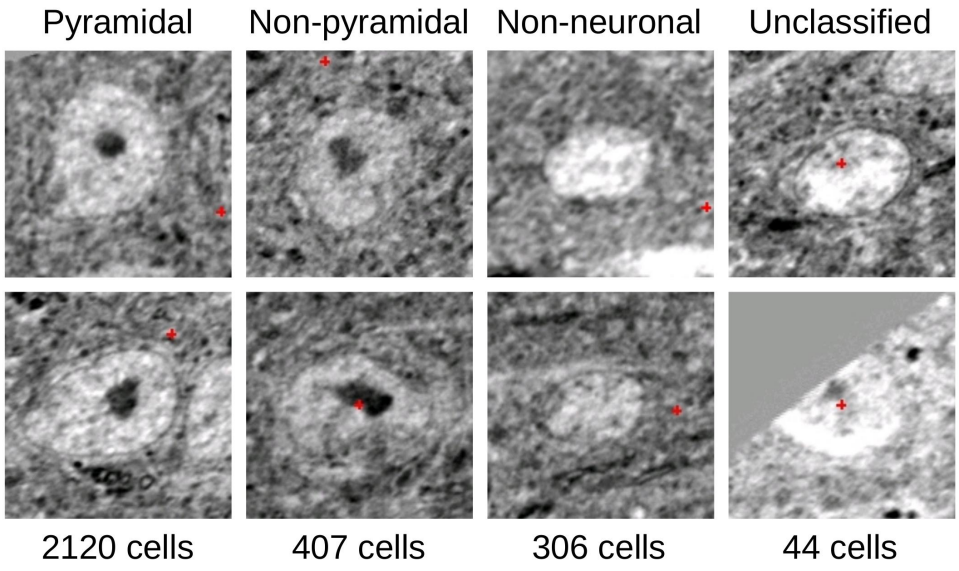


Figure 3.2: Cell class annotations used for training a classification network. For each cell class two example cells are shown.

As I have mentioned before, no segmentation was available for this dataset. The annotations were done is a form of a node ('seed') with the label, placed approximately in the centre of each cell. Thus, just taking a region of a pre-specified radius (bounding

box) around each cell was mostly sufficient to separate the cell from its surroundings. I hypothesised that taking a bigger bounding box around a cell would increase the classification accuracy, since the network would have more access to cell extensions such as axons and dendrites. However, due to high cell density in some parts of the volume, this also resulted in multiple cells being present in one bounding box, so decreasing its radius has proven to be better for the final accuracy. The problem of having multiple cells in one bounding box can be especially disadvantageous for convolutional neural networks, since they do not have an explicit concept of position. That means the cell of interest in the centre could be treated equally to any other cell on the periphery of the bounding box. That is why I have also experimented with adding positional information to the input raw volume. However, this approach did not bring any accuracy improvement.

Another problem that needed to be addressed was the high imbalance of the annotated classes (Figure 3.2). Firstly, if not properly accounted for, it might lead to predictions being biassed towards the most frequent class. To avoid this, I used weighted sampling - sampling training examples (cell volumes) with frequency inversely proportional to their class size. This ensures that the less represented classes are being sampled more frequently, so that the network sees each class with equal probability. Another problem that occurs due to the imbalance is that the low number of cells in the classes of non-pyramidal neurons and non-neuronal cells could lead to overfitting on these classes. Overfitting is a problem when due to the lack of variety in the training data a network might achieve good accuracy on the training data, but fail to generalise to the previously unseen data. To increase the variety of the data I applied extensive augmentations - various transformations that produce different, yet realistic views of the training samples, such as rotations, adding noise or moderate elastic deformations. To prevent overfitting I have also tried adding self-supervised training objectives (e.g. data reconstruction), but this did not improve the classification results.

The last problem of the dataset that I detected was the label assignment (annotation). Though performed with the help of the EM volume, it was nevertheless error-prone. Thus, both training and validation labels have some amount of noise. To ensure smoother training despite the noisy labels, I used so-called 'soft labels'. More specifically, instead of training the network to predict that, for example, a cell annotated as non neuronal should be predicted as 100% non neuronal, I trained it to predict the cell as 90% non-neuronal, 5% pyramidal and 5% non-pyramidal neuron. This approach has been shown to improve training on noisy labels [80] and also improved accuracy on the mouse cortex cell dataset.

3.2.2 Analysing results and failure cases

All the above mentioned adjustments to the training procedure led to the general classification accuracy of 85% (Figure 3.3). Closer inspection of the results showed that the network managed to separate neuronal from non-neuronal cells with almost perfect accuracy, but found it harder to distinguish pyramidal and nonpyramidal cells. This was expected, since it is harder to separate the latter two classes for a human annotator as well.

However, the overall performance of the network exceeded our expectations and showed that the data quality is indeed good enough to distinguish between these cell types with satisfactory accuracy.

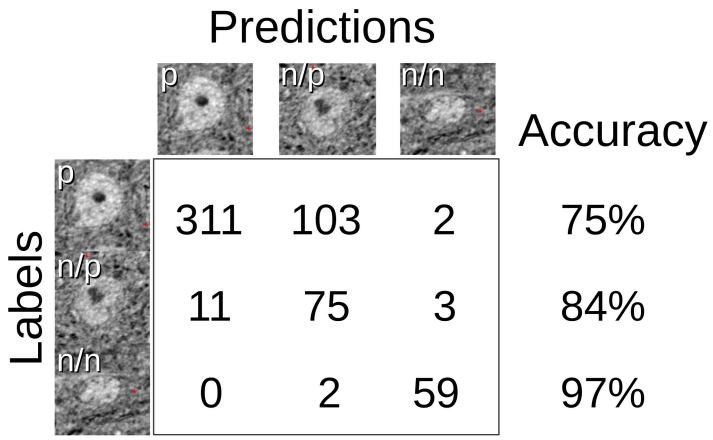


Figure 3.3: Confusion matrix of the best prediction results. The rows correspond to the annotated labels, the columns show the network predictions. The accuracy for each class is stated on the right. (p - pyramidal neurons, n/p - non-pyramidal neurons, n/n - non-neuronal cells).

To investigate the directions for further improvement I first identified the samples that were consistently misclassified by multiple networks. These samples were then inspected by Aaron T. Kuan, who, in many cases, concluded that the network predictions were indeed correct, and the ground truth labels were wrong. A frequent case of truly wrong predictions was observed near the edge of the scan volumes, where the image quality deteriorates and the network could not access the whole intact cell or its extensions. Simply treating the predictions on the volume edges with less confidence or fully excluding them could already improve further analysis. Another common failure case was neuron cells where the apical dendrite had poor contrast. Such cases are also difficult for human annotators, however, one could try to mitigate this by adding more annotated low contrast cells to the training data. Alternatively, one could try using more aggressive data augmentation techniques to generate more training samples with poor contrast. Finally, the network also had a higher rate of wrong predictions for the cells in soma-dense regions. As discussed above, this happens because multiple cells occupied the same bounding box. The ideal way to avoid it would be to first segment all cells in the volume. This would allow to separate individual cells and provide the network only with the cell of interest as input.

To test whether one could acquire cell segmentation with low time investment I used ilastik tool [81], requiring a minimal amount of training data provided as scribbles (brushstrokes). However, after investing around 20 hours to manually generate training labels, proofread the resulting segmentations and apply various post-processing techniques to better separate the objects, I obtained a mediocre-quality segmentation that often failed to separate touching cells (Figure 3.4). Using this segmentation did not bring any improvements to the classification training. I concluded that either extensive manual corrections or a sophisticated segmentation pipeline might be required to generate segmentation of a sufficient quality. However, both of these options require a substantial time investment, thus, were beyond the scope of this pilot project.

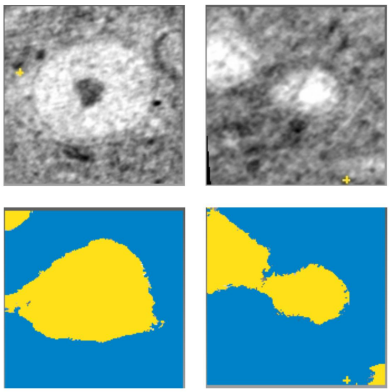


Figure 3.4: Examples of unsuccessful separation of touching cells using ilastik segmentation [81] (top row: XNH volumes, bottom row: segmentation masks).

3.3 Discussion

3.3.1 Contributions

In this project I have shown that the data quality of the lower resolution (130nm) mouse cortex XNH volume is high enough to automatically classify cell types with sufficient accuracy even without cell segmentation. This proves that, given the increased speed and scale of data acquisition in comparison to electron microscopy, the method of X-ray holographic nano-tomography would allow to efficiently localise and annotate cells in bigger brain volumes than it has been possible before.

3.3.2 Potential method improvements

The method presented in this chapter did not achieve a 100% accuracy on distinguishing neuronal cell types, but knowing the cases where the network is more prone to errors could allow for easier targeted proofreading of the automatically obtained classification results. Additionally, I would expect that adding more training data for the underrepresented cell classes will further improve the final accuracy.

Nevertheless, the most obvious and potentially most effective way to get better classification results would be to first segment all cells in the volume. This would also enable subsequent automated morphological analysis and comparison of the different cell classes in the volume. It is also important to note that, unless a corresponding EM volume is

acquired, segmentation of at least larger axons and dendrites is required for many of the subsequent analyses. Such a segmentation could also be leveraged to improve classification accuracy.

3.3.3 Future directions

The results of this pilot project have been viewed as positive by all the stakeholders and will be used for long-term planning of the brain imaging projects at EMBL Rome. The next anticipated step is to investigate whether automatic neuron type identification using aligned light microscopy data, such as spatial transcriptomics, could be used to replace the lengthy step of manual cell annotation, which is needed in order to generate the ground truth data for training a classification network.

Chapter 4

More detailed cell characterisation through combining genetic information with high resolution imaging

4.1 Background

Traditionally describing cells has been driven by a clear biological question or a hypothesis in a researcher's mind. Similar to the work mentioned in the previous chapters, researchers often focus their attention on a limited number of cell types in order to characterise their differences. Moreover, as I showed earlier, it is common to investigate specific genes expressed in these cells that one expects to be responsible for these differences. Such narrow focus certainly simplifies the analysis - it is easier to find something when you know what you are looking for. However, generating a hypothesis in the first place often requires extensive prior knowledge and heavily relies on previous observations and experiments. In this chapter I talk about hypothesis-free ways to collect large-scale observations that can be used to extensively describe variety in the data, which can in turn lead to new discoveries and new hypotheses. Moreover, I discuss ways to combine different large-scale observations to generate a more precise description of the object of interest. For this I introduce the multimodal atlas of a whole organism - a three-segmented worm of *Platynereis dumerilii* - that integrates fine morphology and gene expression on a single-cell level.

4.1.1 Analysing cells at scale

Rapid technological advancement of the methods used to acquire biological data are changing the very way we approach analysing cells. Current sequencing techniques allow us to simultaneously access all the genetic information being transcribed in each cell of com-

plex organisms. This enabled, for example, throwing light on gene regulatory programs in early-branching metazoans [41], characterising cellular diversity of the human body [82, 83, 84, 85] or showing parasite evolution inside host cells [86]. Imaging methods offer both constantly improving spatial and temporal resolution, and the possibility of imaging increasingly big samples. This has made it possible to image structures as small as a single chloroplast pyrenoid [87], organs as big as the whole *Drosophila* brain at synaptic resolution [88] or processes as extensive and dynamic as early mouse development at the single-cell level [19]. Moreover, spatially resolved mass spectrometry has recently made it possible to investigate not only proteins, but also metabolite composition of single cells [89, 90]. This, among other discoveries, has led to a better understanding of both distribution and function of various lipids in cells and tissues [91, 92, 93, 94].

All these advances finally make it possible to not be restricted to targeting specific cells or molecules within. Now a researcher can just select an organ or an organism of interest and analyse all the cells composing it simultaneously. This greatly expands the amount of raw information we can get from an experiment, since now we are not limited to confirming or rejecting a previously formulated hypothesis, but can discover and describe novel categories - new cell types, genes or metabolites. We can access all the variety in the data at once and get the most comprehensive description of this variety, not limited to the groups we already know. This is how, for example, new types of lung [95], heart [96] and intestine [97] cells were discovered in humans, or new interactions between cellular organelles and microtubules were described [98].

4.1.2 Generating all-around descriptions of cells

Although all these methods expand our exploration of cells to the whole-organ or even whole-animal scale, they nevertheless lack the ability to generate a truly comprehensive description of a cell. Having a complete transcriptome will not allow us to precisely infer which proteins and metabolites in which quantities a cell expresses. Similarly, having a full proteome and metabolome will not enable us to unambiguously describe cellular phenotype and functions, since these are heavily influenced by the environment the cell grows and functions in. Evidently, only combining these observations would create a full description of a cell, allowing to illuminate all the cellular variety in an organism.

Merging multiple types of cellular data into so-called multiomics datasets has been a topic of thorough research in the last years. Various efforts have focused on technically combining transcriptome with methylome [99, 100] or chromatin accessibility measurements in single cells [101, 102], and computationally combining diverse measurements obtained from different cells of the same population [103, 104, 105]. The analysis of such combined modalities is far less straightforward, but often brings more insight than analysing these modalities separately. For example, it allowed to identify previously unknown lymphoid subpopulations [106], or helped to show that by default mouse embryonic stem cells develop into ectoderm [107].

4.1.3 Giving cellular descriptions their spatial coordinates

Another valuable method of describing cells that receives a lot of attention recently is imaging data. Integrating multiomics data with imaging is essential, not only because the latter provides us with the ultimate phenotype of a cell, but also because it gives the whole dataset its spatial coordinates. Cellular functions in complex organisms are rarely fulfilled by single cells, but rather by ensembles of neighbouring cells. Thus, knowing cellular surroundings might be crucial to shed light on the way a cell functions.

However, such multimodal datasets including a spatial component, often referred to as multimodal atlases, are still far from being abundant. The reason for this is the difficulty of assigning the multiomics measurements to a defined cell in the imaging data. So far this has been mostly done using transgenic lines in model organisms like *Drosophila* and *C. elegans*, where such mapping is heavily assisted by extensive libraries and resources already at hand [108, 109, 110, 111, 112]. For species with less prior knowledge available, such a mapping can be done by hand, albeit for a smaller number of cells, like in a 64-cell stage of a *Phallusia* embryo [113]. Clearly, both approaches do not scale when it comes to exploring cellular diversity of new organisms that are larger in size.

The recent advances in spatial transcriptomics techniques make it a promising way to bridge spatial and non-spatial modalities with minimal amount of manual work required. While the method is limited in the number of genes whose transcription can be detected in one sample, its indisputable advantage is the access to the physical location of the studied expression at a subcellular resolution. Thus, when combined with full single-cell transcriptomics data, this generates a complete transcription atlas in space [114, 115]. Such spatially aware transcription data allows for a more precise description of cells and their functions and have, for example, assisted characterising cell types and their communication in sponges [7] or in hypothalamic preoptic region of the mouse brain [116]. Furthermore, it gives access to the tissue or cellular morphology that could as well be integrated with the gene expression data [117, 118, 50].

4.1.4 Incorporating high resolution cellular morphology

However, due to the limited resolution of light microscopy, such an approach would only allow us to analyse some basic cellular morphology, such as the shape and size of cells and nuclei. In order to get a better description of cellular interior, such as appearance and distribution of both membrane- and membraneless organelles, one would need to acquire an electron microscopy volume of the data. This could serve an ambitious goal to simultaneously characterise and possibly correlate fine cellular morphology with genes expressed in a cell. However, in order to achieve this, the additional step of assigning spatial gene expression to the EM data has to be resolved. So far this has only been done either for a limited number of genes using correlative light and electron microscopy approaches [119], or, as described above, using transgenic lines that required a considerable amount of manual work. Thus, finding a way to automate such an assignment could substantially fa-

cilitate bridging spatial gene expression data with high resolution morphology, that would, in turn, eventually allow mapping single-cell transcriptomics data onto EM volumes and characterising genotype-phenotype connection on the single-cell level.

4.1.5 Multimodal atlas of Platynereis dumerilii

To the best of my knowledge, the first work aimed at automatically bridging EM and spatial gene expression, is the multimodal atlas of a young worm of *Platynereis dumerilii* [65], that combines a fully segmented electron microscopy volume of the whole animal with gene expression maps for over 200 genes (Fig 4.1). This animal is a newly established model organism for studying evolution, development and neurobiology [120]. What makes *Platynereis* particularly suited for such types of studies is the fact that at 6 days post fertilisation (dpf) the animal possesses a multitude of almost fully differentiated cell types [121]. Still at the same time it has a modest number of total cells that considerably simplifies whole-animal analyses such as single cell transcriptomics or connectomics. Additionally, the development of *Platynereis* is highly stereotypical [122] that makes it possible to unambiguously compare multiple animals or combine data from different modalities into single-cell resolution atlases.

Such stereotypy allowed bringing together data types with greatly differing resolutions to generate a combined multimodal atlas. First, the EM volume of the animal has been acquired using serial block-face electron microscopy [123] by Benjamin Titze and Christel Genoud at a resolution of 25*10*10 nm, giving access to high resolution ultrastructure of the animal. Then all the cells and nuclei of the animal were segmented by Constantin Pape, yielding more than 11,000 of morphologically diverse cells. Additionally, many animal tissues were segmented by Constantin Pape, Hernando M. Vergara and myself. Moreover, Kimberly I. Meechan segmented chromatin to differentiate between heterochromatin (and nucleolus) and euchromatin. Finally, the gene expression maps, acquired with subcellular resolution by whole mount *in situ* hybridization [124] were registered onto the EM volume by Christian Tischer, bringing genetic information and morphology into a common space.

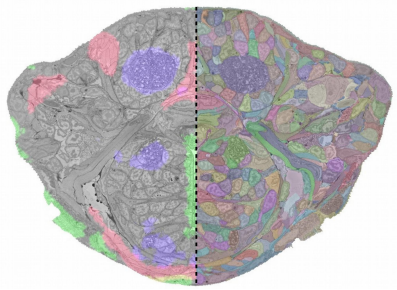


Figure 4.1: *Platynereis* data. A slice of the volume is shown, with genes *mhcl4* (red), *glt1* (blue) and *grm7* (green) overlaid on the left and cell segmentation visualised on the right.

4.1.6 Project aims and contributions

In order to investigate potential connections between gene expression and morphology of each single cell, the gene expression had to first be accurately assigned to the segmented cells - a challenging task due to substantial differences between the data modalities and their resolutions. Thus, my aim was to devise an optimal strategy to unambiguously assign each cell a vector of its gene expression. Additionally, I conceptualised an assignment validation method in order to compare multiple assignments.

The results of this chapter have been published as a part of [65] that also describes the *Platynereis* atlas and the discoveries it made possible in more detail.

4.2 Results

4.2.1 Assigning genes by overlap

The gene expression maps of the atlas were acquired at subcellular resolution. Nevertheless, these maps are unaware of cellular boundaries due to the absence of membrane staining. In theory one could expect such gene expression to accurately follow cellular boundaries. That would reduce the problem of assigning gene expression to the cells segmented in the EM volume to calculating the overlap between the segmented cell and the gene expression map.

However, in practice this expectation does not hold true for multiple reasons. Firstly, even within a single cell one can not expect a gene to be expressed uniformly, and there might be some areas truly lacking the expression of this gene. Secondly, the expression maps were created not from a single animal, but are a result of averaging the spatial gene expression of multiple animals. While at this stage of development the animal is highly stereotypical, there are still some variations in cellular locations up to one cell diameter (<4.7µm) [122]. Thirdly, due to the necessity of choosing manual thresholds that define what a true expression is in the ProSPr pipeline used for obtaining the maps [122], some gene expression volumes end up over- or undersaturated. Finally, registration of the gene maps onto the EM volume also shows variable quality in different parts of the animal, with the average registration error being around 3µm.

These factors combined lead to the fact that for many genes in various regions of the animal, the expression front is not resolved to the boundaries of single cells. For example, glutamate transporter *glt1*, whose expression is expected to be limited to the neuropil (Figure 4.2A) leaks out into the neighbouring cells. Similar leakage can be also noticed for the marker of striated muscles *mhcl4* (Figure 4.2B) and the pan-neuronal marker *elav* (Figure 4.2C). Such leakage greatly reduces the precision of a simple overlap gene expression assignment.

To illustrate this, I started with calculating the overlap between all the genes in the atlas and all the segmented cells of the EM volume. This gave me a range of values from

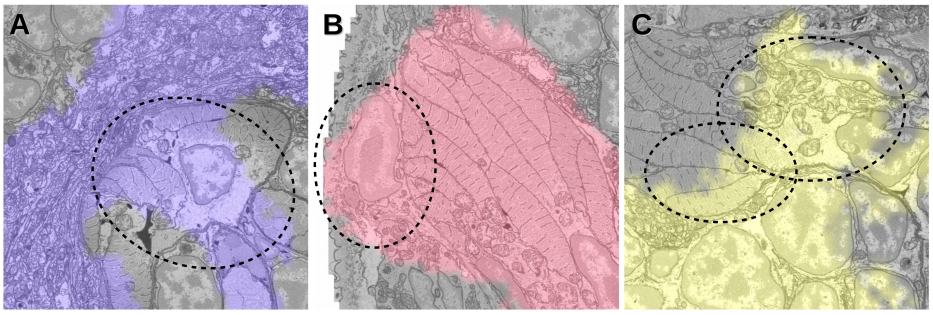


Figure 4.2: Gene leakage. Examples of gene misalignment on the border of the gene expression is shown for genes *glt1* (**A**), *mhcl4* (**B**) and *elav* (**C**). The areas where the gene should not be expressed are encircled.

0 to 1 for each cell. Since the registration quality varies in different parts of the animal, coming up with a reasonable threshold of what to consider a true gene expression of a cell is not a trivial task. That is why I converged to using an intuitive margin of 0.5 - if half or more of the cell volume expresses a gene, I considered the gene to be truly expressed. Such an assignment immediately highlighted the problem of leaking genes. For example, in the animal regions, where the boundary between muscle and neuron tissues lays, I found various cells expressing both the neuronal marker *elav* and the muscle marker *mhcl4* that should not be expressed in the same cell (Figure 4.3). Taking one step back and checking the raw, non-binarized gene overlap values in these cells did not give me an idea of what a good threshold could be that would keep the true expression and remove the false one in these cells. Thus, I aimed to find a better way to assign gene expression to the morphological cells in the EM volume.

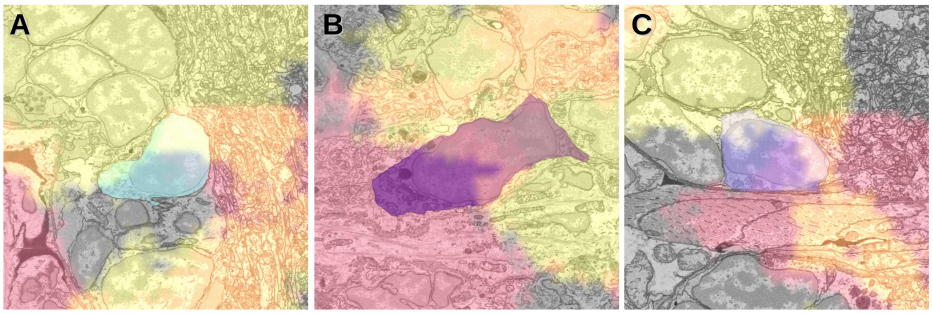


Figure 4.3: Cells with contradictory gene expression. Shown are the examples of cells expressing both the pan-neuronal marker *elav* (yellow) and the striated muscle marker *mhcl4* (red).

4.2.2 Generating Virtual Cells

The gene map imperfections, resulting from biological variability of multiple samples averaged per gene, are hard, if not impossible, to mitigate. However, I expected that the contribution of the registration error, coming from the step of aligning the gene expression maps to the EM volume, could be removed if we take a step back and make use of the unregistered spatial gene expression. More specifically, I assumed that defining gene expression profiles already before the registration step could result in profiles of higher quality. Thus, following how it was done in [122], I set out to identify 'coherent units of homogeneous gene expression' [122] referred to as Virtual Cells (VCs) (Figure 4.4, Figure S2AB). It is worth to note that, despite being called Cells, VCs do not represent gene expression profiles of single cells, but rather the profiles shared by multiple cells, mostly neighbouring or bilateral.

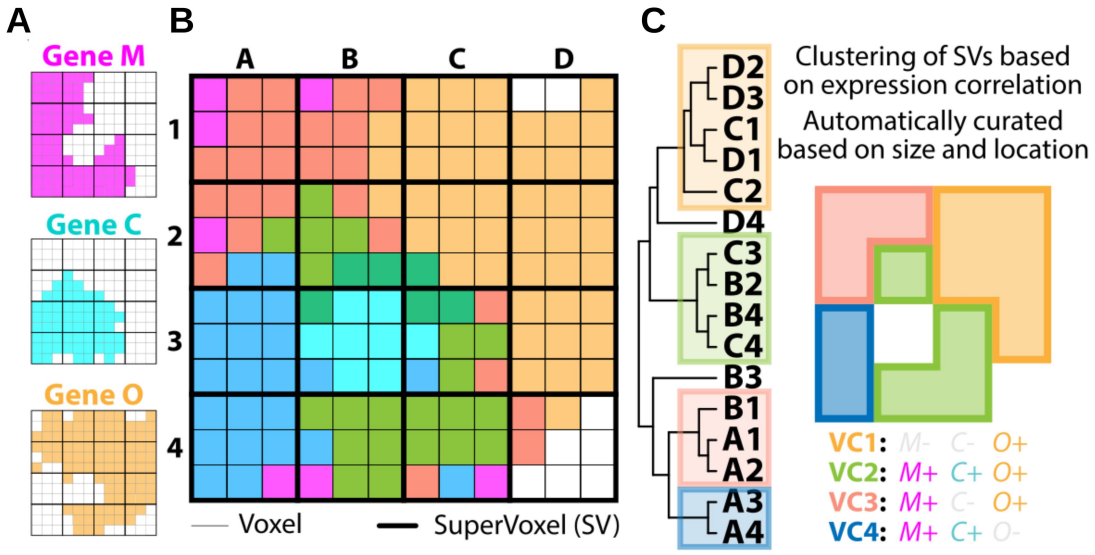


Figure 4.4: Generation of Virtual Cells. **A.** An example of a region in the animal, where genes M, C and O are expressed. **B.** An overlay of the expression of the three genes on a supervoxel grid. **C.** Hierarchical clustering brings together supervoxels with similar gene expression patterns, which make up Virtual Cells (VCs, shown in different colors). The genetic profile of a VC is calculated as the average profile of all the supervoxels that constitute the VC. Adapted from Hernando M. Vergara [65].

Assigning VCs to segmented cells instead of raw gene expression could offer multiple advantages. First, as I have mentioned before, this would remove the impact of the registration error. Second, since VCs are intended to be coherent homogeneous expression profiles shared by many cells, we would expect such profiles to contain less noise simply due to the averaging across multiple cells. Finally, an additional step of bilateral symmetry checking during VC generation could lead to even further noise reduction.

To generate such VCs I have taken the pipeline introduced by Hernando M. Vergara as a part of [122] and automated the manual steps of it. More specifically, I have introduced an automated curation that removed all the VCs smaller than a threshold defined by the level of their spatial correlation. Additionally, I have added a step that splits the VCs into connected components (CCs) and removes the CCs that are too small to represent a cell. Finally, for all the CCs around the size of one cell I introduced a quality check step that verifies bilaterality of such CCs (described in more detail in Chapter Methods 7). The updated pipeline is available at `https://github.com/mobie/prospr-vc-generation`.

4.2.3 Assigning Virtual Cells by genetic distance

Having automated the pipeline, I generated VCs using the genes available in the atlas. This resulted in 6426 VCs, i.e. distinct gene expression profiles. These VCs could be split into 12,393 CCs that roughly corresponds to the number of cells segmented in the EM volume (11,402).

Assign segmented cells to their corresponding VCs was based on the following logic. I considered the above-mentioned overlap assignment to be a reliable, yet imperfect gene expression profile, so instead of completely rewriting the assignment, I aimed at refining it using the available VCs. That is why for each cell I found a VC in its vicinity with the closest gene profile to the one assigned by overlap (Figure 4.5). Given the sample variability and registration error, I defined the vicinity to be within approximately one cell radius around the given cell (around 5µm).

Additionally, considering the highly non-uniform gene coverage of the animal, I decided to allow any cell to be assigned an empty expression vector. This was done not only in case there were no VCs found in the given radius around the cell, but also if such assignment would still be genetically closer to the cell than any of the VCs in vicinity. With this I allowed cells with minimal or completely absent gene expression to refrain from being assigned any genetic profile at all.

This approach led to 10,293 cells being assigned a VC and 1109 cells having no gene expression profile. In total, 4413 unique expression profiles got assigned.

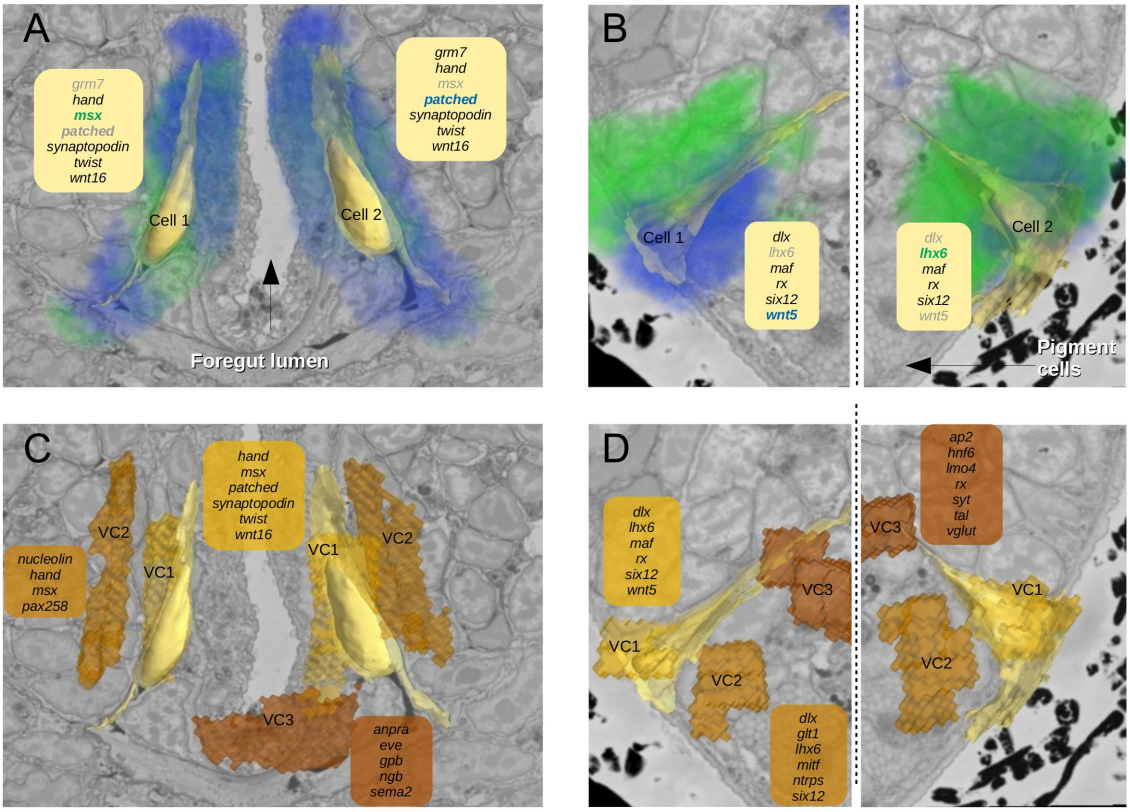


Figure 4.5: Comparing overlap and Virtual Cell assignment on bilaterally symmetric pairs of cells. **AB.** Gene expression maps are visualised in the animal volume for (**A**) *msx* (green) and *patched* (blue) and (**B**) *lhx6* (green) and *wnt5* (blue). Gene assignment by overlap is shown in yellow boxes with genes not assigned (below 50% overlap) shown in light gray, illustrating asymmetry of gene assignment by overlap. **CD.** Visualising Virtual Cell (VC) assignment for the cells in A and B. For each cell all the VCs in the neighbourhood are considered (shown in shades of brown with their genetic profiles stated in boxes of corresponding colours). Each cell gets assigned the VC with the smallest genetic difference (shown in yellow-brown) that results in a more consistent denoised genetic profile.

4.2.4 Validation via symmetric pairs

In order to compare the VC assignment to the assignment by gene overlap, I used the fact that the animal is bilaterally symmetrical, thus, almost every cell on one side of the animal has a symmetric partner on the other side. Such partners should look similar, making it possible to manually identify pairs of symmetric cells. Moreover, I would expect them to express the same genes, thus, allowing us to check the discrepancy in their gene profiles assigned by the above-mentioned methods.

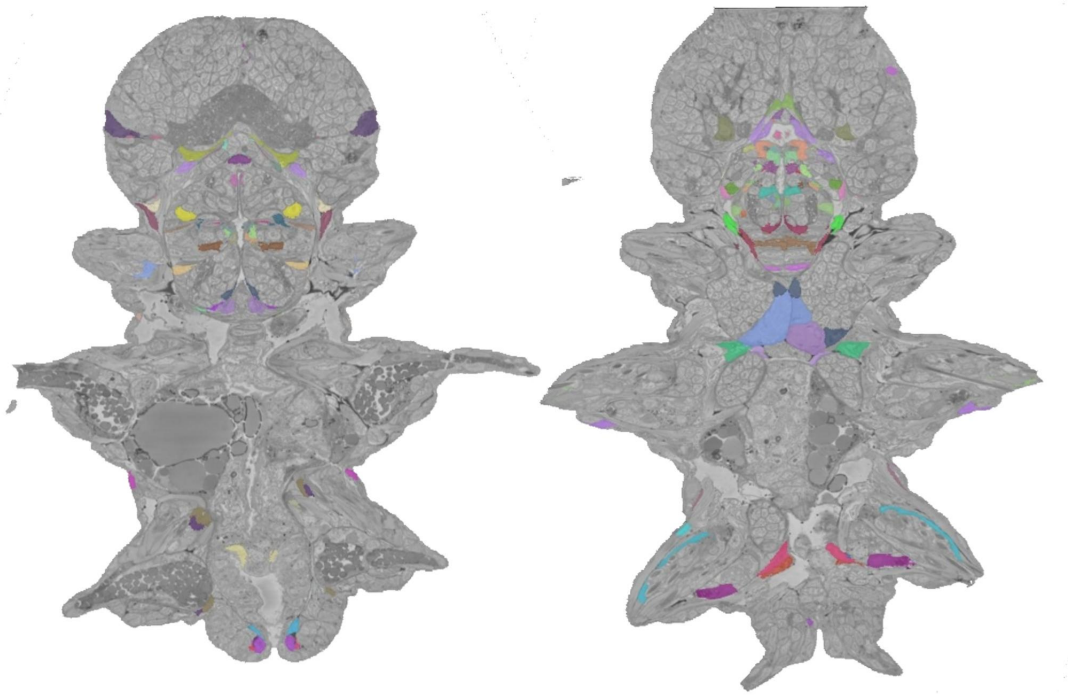


Figure 4.6: Symmetric cell pairs. Some of the annotated cell pairs with similar morphology and expectedly similar gene expression are visualised in the animal volume.

Thus, to validate the assignments, I first identified 208 symmetric cell pairs in the volume. I chose the pairs based on my ability to unambiguously identify the symmetric partner. This often implied that the pairs had distinct appearance. In order to account for the variation in registration accuracy, I tried to cover all animal regions, whenever possible (Figure 4.6).

For each symmetric cell pair I calculated the discrepancy in the assigned gene expression for both methods. The overlap and the VC assignments showed an average discrepancy of 2.8 and 2.0 genes per pair, respectively, showing the VC assignment to be around 30% more precise. The results show that the VC assignment performs better in around 70% of symmetric cell pairs (Figure 4.7A). To investigate the influence of gene coverage on the assignment quality, I related the gene assignment discrepancy in a cell to the number of genes comprising the VC expression profile assigned to this cell (Figure 4.7B). This shows that, indeed, the assignment performance is higher for more gene-rich VCs, suggesting that the quality of the gene assignment could be improved by expanding the gene expression atlas, especially in gene-poor areas, such as the parapodia or the midgut (Figure S1).

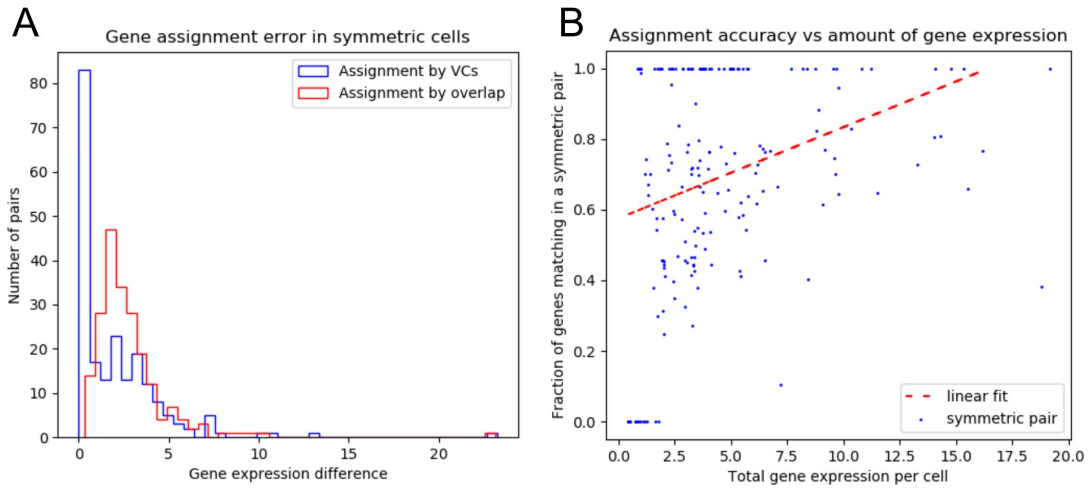


Figure 4.7: Gene expression assignment accuracy. **A.** Gene expression assignment discrepancy for the Virtual Cells and the overlap assignment in the manually annotated symmetric cell pairs. **B.** More gene expression information leads to better assignment. The gene expression assignment discrepancy is plotted against the total level of gene expression in a cell (the fractions of gene expression for each gene summed up).

4.2.5 Assigning Virtual Cells by proximity

I further explored whether the idea of refining the overlap assignment rather than rewriting it was reasonable. Despite the fact that the location of the VCs assigned to segmented cells in the EM volume was consistent with their original location before the volumes registration (Figure S2), around 60% of segmented cells got assigned to a VC that was not the spatially closest one. Thus, I investigated if assigning VCs by spatial proximity, i.e. overlap, rather than by minimising gene distance, could bring better results.

To test this for each cell I determined the VC that overlapped the most with the cell. In case no VC was overlapping with the cell, I assigned it an empty gene expression vector. Validating this assignment using the above-mentioned symmetric cell pairs resulted in a mean discrepancy value of 3.62. This showed the VCs are potentially even more misaligned to the EM volume than the raw gene expression, and the initial raw gene overlap assignment step is still required to get the best assignment accuracy.

4.2.6 Cells with contradictory gene expression

Finally, I examined the cells on the border of the neuron and muscle tissues that, as I noted before, express contradictory genes after the overlap assignment. While not all such contradictory cases were resolved by the VC assignment, many cells got assigned a gene expression profile that I could interpret as more reasonable. For example, according to the VC assignment the cell depicted in Figure 4.3A does not express myosin marker *mhcl4*, which overlaps the cell by 54%. However, the cell is assigned the expression of neuronal genes *kv33b* (overlap of 43%) and *stathmin* (overlap of 47%). Such assignment could be more justified both in terms of the genes function and their distribution in the animal volume (Figure 4.8).

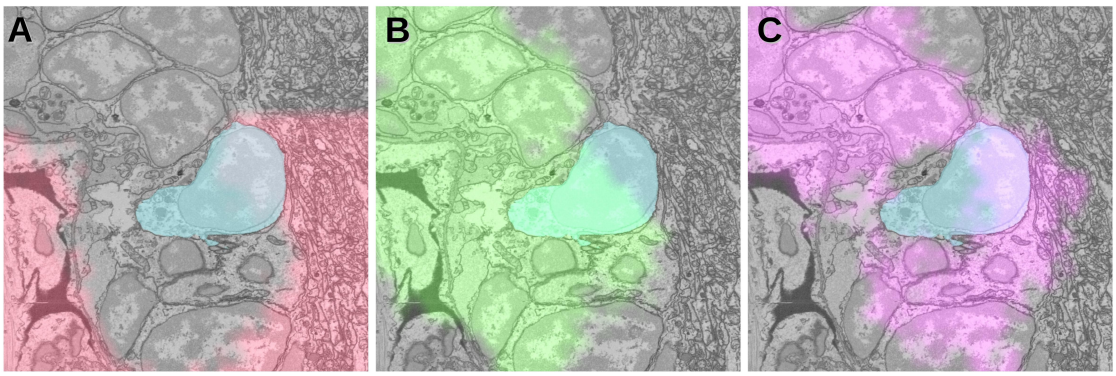


Figure 4.8: Gene expression corrected by VC assignment. Visualised in the EM volume are the gene expression maps for *mhcl4* (**A**, assigned by overlap, not assigned by VC), *kv33b* (**B**, not assigned by overlap, assigned by VC) and *stathmin* (**C**, not assigned by overlap, assigned by VC).

4.3 Discussion

4.3.1 Contributions

This chapter was centred around combining gene expression data with high resolution imaging as a way to better describe cellular diversity within a whole animal. More specifically, I focused on improving the gene expression assignment to the cells segmented in the EM volume. I have highlighted the shortcomings of the simplest assignment by overlap between a gene expression map and a segmented cell volume. Further, I have automated a previously developed way of generating coherent denoised curated gene expression profiles - Virtual Cells - and developed a pipeline to assign these profiles to the segmented cells in the EM data. Finally, I came up with a method to automatically validate such assignments by leveraging the bilateral symmetry of the animal.

4.3.2 Potential improvements

An approach that I expect to significantly improve the gene assignment is to add more genes to the gene atlas. As I have shown, the symmetric pair distance decreases with the amount of genes comprising a Virtual Cell. Therefore, creating more gene-rich VCs would result in a better assignment as well. Moreover, only a part of the atlas was generated with the aim of uniformly covering the animal tissues, while a big portion was focused on neural differentiation genes and transcription factors. Adding more genes to poorly covered regions would lead to gene assignment of a more uniform quality. I presume the best way to achieve this would be to use single-cell transcriptomics data. Having a full picture of the gene expression of the animal would allow us to define genetic cell types with corresponding specific genes. Afterwards, acquiring gene expression maps for such differentially expressed genes would ensure uniform coverage for all the defined cell types.

Further, one could focus on improving the gene map and Virtual Cell generation pipelines. There are multiple steps that could be refines, starting with optimising the hybridisation protocols, image acquisition and registration procedures. As explained in [122], each gene map is an average of multiple animals. Thus, I expect that both gene maps and Virtual Cells would greatly benefit from defining more precise ways of determining what the true expression is already on the level of a single-animal single-gene map. Likewise, finding a better way of combining gene maps from multiple animals could also be a notable improvement. Better ways of curating the VCs, as for example the mentioned above suggestion to use single-cell transcriptomics data, should also result in a better assignment.

Another interesting idea to tackle the problem of leaking genes, besides improving the gene assignment, might be to additionally calculate a confidence score for each assignment that would reflect how trustworthy the assignment of this specific gene in this specific cell is. This could be done, for example, considering the expression of the neighbours of a given cell - the more of the neighbours express a gene, the more likely it is that the expression of

this gene in this cell is a leakage from the neighbours, and vice versa. Another way of doing it would be to take cellular morphology into account - the more dissimilar the neighbours are that express this gene, the less likely it is that this gene expression is true. After such confidence scores have been calculated, they can be used to decide on a threshold of what a true expression is, i.e. the lower the confidence score is, the higher the gene overlap has to be for us to claim that this cell expresses this gene.

4.3.3 Future possibilities

I expect the coherent genetic profiles of Virtual Cells to be the closest modality to single-cell transcriptomics data. As I discussed earlier, gene expression overlap values do not necessarily correspond to the actual levels of gene expression in a specific cell. Therefore, in order to use the overlap values for matching single-cell transcriptomics profiles, one would have to account for varying assignment quality in different parts of the animal. This implies defining separate thresholds for each gene and potentially for each cell, which is clearly far from straightforward. Virtual Cells, on the other hand, do not need such thresholds, thus, are easier to match directly. The problem could be reduced even further if, as I suggested above, one directly uses the available single-cell transcriptomics data to generate and curate the VC profiles. Therefore, once such data is acquired for the 6dpf stage of *Platynereis dumerilii*, I believe, Virtual Cells could be used to unambiguously map this data to the cells in EM, creating the first multimodal atlas of such kind and resolution. This would ultimately enable linking gene expression and phenotype data on the single-cell level, allowing us to explore the interconnections between the genetic program and the function of a cell.

Chapter 5

Unsupervised comprehensive description of high resolution cellular morphology

5.1 Background

In the previous chapter I talked about combining gene expression and fine morphology information on the single-cell level. I further discussed how it could assist us in generating a comprehensive description of a cell and understanding the causation between genotype and phenotype. I focused on assigning spatial gene expression data to EM volumes, which could serve as a bridge to assign each cell its full genetic description - single-cell transcriptomics data. However, in order to explore the connection between what a cell expresses and how it looks, one also needs a full quantitative description of cellular morphology. In this chapter I justify the need to develop automatic, unbiased ways of extracting such descriptions in the most comprehensive manner. I further illustrate how these learned descriptions can assist exploratory analysis and facilitate discoveries of cell types and tissues in the whole-animal volume of *Platynereis dumerilii*.

5.1.1 Describing cellular morphology

Historically, morphological descriptions have been generated manually for a limited amount of cells by a researcher equipped with prior knowledge of cellular structures. Currently, fast advancement in volume EM allows us to image organs or small organisms composed of thousands of cells. This opened up possibilities to explore, for example, cellular morphology of the *Drosophila* brain [88, 55] or ultrastructural differences in cell types of *Platynereis* [121, 65]. The computational progress follows along, enabling us to both segment [125, 98, 78, 37] and smoothly browse these large data volumes [126, 127, 128]. However, if the goal is characterising all the morphological diversity in such imaged sam-

ples, generating manual descriptions quickly becomes infeasible.

Ideally, we would have a morphological description similar to the genetic one that we get from single-cell transcriptomics. We would want to have a set of morphological features and a way to automatically estimate the presence (or the amount) of each feature in every cell, thus generating a feature vector per cell. These features should, ideally, fully describe cellular appearance. Moreover, the created feature space should have the property that cells mapped in close proximity are visually similar to each other, and the dissimilar cells are located further apart. With this we would hope that grouping the imaged cells in such feature space results in morphological cell types that we should further be able to correlate to genetically-defined ones.

5.1.2 Automating morphological descriptions

So far automated morphological descriptions have mostly contained a limited number of well defined features of interest. Such features, for example, include shape, volume, surface area, sphericity and curvature and/or various intensity and texture features [56, 57, 129, 130]. While explicitly designed representations allow for automated large-scale comparisons, such features have to be manually chosen. This resembled how researchers were previously targeting only a couple of specific genes of interest - the approach delivered good representations given one knows what one is looking for, however, might fail to capture the full variety of morphology in a sample.

This problem has led various researchers to design complex modelling pipelines aimed at automatically extracting the full range of morphological differences in a dataset. For investigating shape variation, for example, good results were obtained using large deformation diffeomorphic metric mapping [131], spherical harmonic descriptors [132] or user-assisted classification of approximate convex patches [133]. However, such pipelines are still mostly limited to a certain aspect of cellular morphology, e.g. shape, and are often tailored to a specific type of data and/or specific task.

To be able to get a full description of cellular morphology we need broadly applicable pipelines that are not biassed neither by the pre-designed task, nor by the manual selection of features. This is especially important if we aim at fully characterising a dataset, that includes both describing already known morphological categories and discovering novel ones, such as new morphologies and new cell types. We further need such pipelines to be equally focused on various components of cellular morphology, such as shape and texture. Moreover, we need such pipelines to be broadly applicable to various types of data.

5.1.3 Using deep neural networks to describe morphology

In the last decade the rapid development of artificial deep neural networks has rendered using both manually defined features and complex feature extraction pipelines obsolete for a wide range of tasks on multiple types of data. This especially concerns analysing images, where convolutional neural networks (CNNs) have been shown to extract rich

visual features suitable for multiple image processing and analysis tasks [60, 134, 135, 136]. However, a big obstacle still hindering free and easy application of such networks to biological images is a massive amount of training data required to train a network. Such data has to be manually labelled by an expert, and, for some tasks, for example, EM segmentation that requires pixel-level annotations, generating this ground truth might take multiple months [137, 98].

To reduce the amount of human work needed, multiple approaches have been proposed and proven successful, including transferring models trained on one type of data to a different one or initially using multiple data types to train huge models that are expected to generalise well to new data (reviewed in more detail in [138, 139, 140]). Many efforts are also focused on reducing the amount of labels needed to train a model [141]. The most extreme example of this are unsupervised or self-supervised learning methods that do not require any manual annotations, but instead use training signal obtained from the data itself. Besides the obvious advantage of less manual labour, such methods do not optimise the extracted features for a specific task, such as classification, and the features they produce have been shown to be versatile and useful for diverse downstream tasks [142, 143].

Until recently the main disadvantage of self-supervised methods was their inferior performance in comparison to their fully supervised counterparts. Yet this has changed since the emergence of so-called self-supervised contrastive learning methods that try to bring closer the features of a sample and its modified versions. In the natural image domain research on how to exactly produce these modified versions allowed to bring the accuracy of contrastive learning on par with fully-supervised learning on classification tasks [144, 145]. However, many of the techniques that bridged this gap rely on peculiarities of natural images, e.g. varying scales and resolution, RGB channels, etc. Adapting the latest self-supervised training procedures to the characteristics of biological imaging data could enable researchers to use state-of-the-art deep learning to assist them in performing extensive analysis of the data without generating manual labels or even formulating a concrete analysis goal.

A handful of biological studies have already demonstrated promising results of self-supervised feature learning, generating representations useful to classify cells or analyse their morphological variability. However, many of them used a strong pretext task based on available metadata, such as the prediction of cell lines [61] or experimental conditions [146], or the classification of proteins fluorescently labelled in cells [62]. In the last example the authors explicitly show that without the pretext classification task the method performance declines more than 2-fold. Several studies used an autoencoder reconstruction loss [147] to show the possibility to refrain from using any labels on single-cell imaging for image-based profiling [148, 149]. Yet adapting contrastive learning for biological data has so far only been used on 3D electron microscopy volumes for describing local morphology of patches or subcompartments within neural cells [150, 151].

5.1.4 Project aims and contributions

Unsupervised deep learning has been shown to extract rich representations from biological data useful for various tasks. The aim of this project was to design a fully unsupervised pipeline and apply it to the cells and nuclei in the EM volume of *Platynereis dumerilii* to extract unbiased comprehensive morphological representations that would combine high-resolution ultrastructure and shape morphology of a whole cell. Such representations should allow us to access all the morphological diversity in the whole animal and to define morphological cell types, which we later aim to correlate with the gene-defined cell types in order to explore the genotype-phenotype connection.

The project was carried out together with Johannes Hugger (EMBL-EBI) and the results are available in [66]. The contributions of Johannes Hugger, which comprise the whole shape training and extraction pipeline and adding neighbourhood information, will not be described here and can be found in detail in [66]. The chapter will, thus, focus on the texture feature extraction and the analysis that was done by myself.

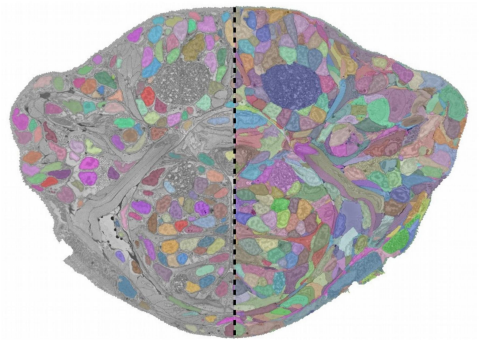


Figure 5.1: Segmentation of the *Platynereis* data (right: nuclei, left: cells).

5.2 Results

5.2.1 Feature extraction pipeline

The first step needed to extract useful features from a cell is to retrieve the cell of interest from the EM data. *Platynereis* atlas contains full segmentation of 11,402 cells and their nuclei available (Figure 5.1). I used segmentation masks to extract only the internal part of each cell in order to ensure that the extracted features represent solely the cellular morphology and not the cell surrounding (Figure 5.2). It is important to note that for neurons only cell somas were segmented, thus, morphology here does not refer to neurite morphology.

Further, I decided to split cells into three morphological components: shape, texture context (a low resolution crop around the cell nucleus) and fine texture patches (multiple crops of high resolution cellular texture) (Figure 5.2). This split was initially required by the size and complexity of the data: a whole full resolution cell will not fit into the memory of a standard GPU for processing. Moreover, it also ensured the model takes both shape and texture into account and does not focus only on the most prominent features.

Additionally, the available nuclei segmentation allowed me to further split cell cytoplasm and cell nuclei, generating in total 6 morphological components per cell. In practice,

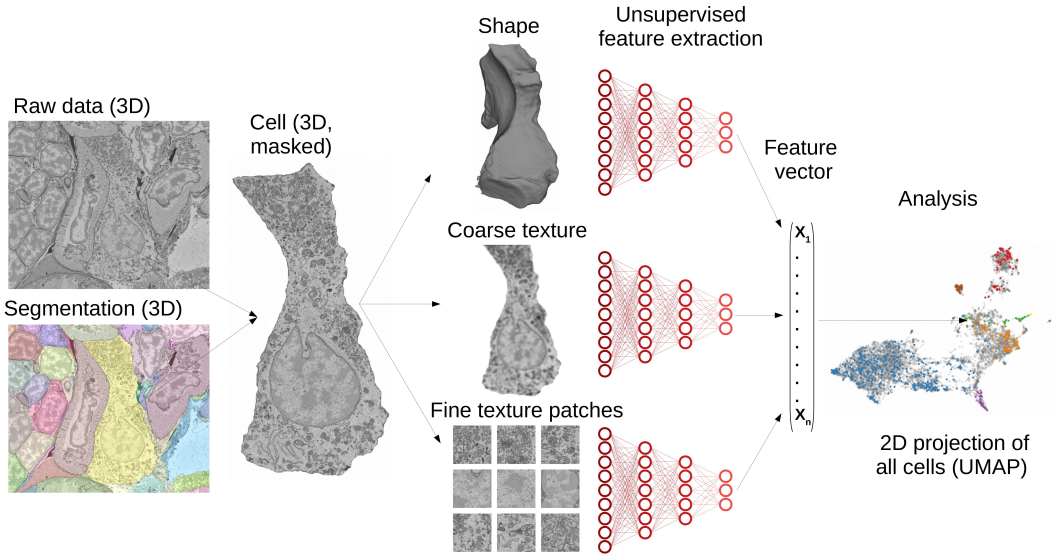


Figure 5.2: Feature prediction pipeline. A cell segmentation mask is used to extract the raw volume of a given cell. This volume is used as input for separate neural networks that yield morphology features for shape, coarse and fine texture of cytoplasm. The same procedure is applied to nuclei (not shown here). The resulting feature vectors of a cell and its nucleus together comprise one MorphoFeatures representation.

this fine-level split not only improved the quality of the features, but also made them more interpretable. This made possible estimating which of these components played an essential role in a subsequent analysis, e.g. distinguishing cell types.

After splitting the data, neural network feature extraction pipelines are run on these components separately, yielding distinct feature vectors of the same size (80 features) for fine and coarse texture and shape for cytoplasm and nucleus. These vectors get concatenated to form one comprehensive morphological representation of the cell. This feature extraction is run for each cell separately, resulting in one representation vector (480 features) per cell. A matrix of these representations is then used for subsequent analysis.

5.2.2 Training pipeline

In order to obtain high quality morphological representations for each of the texture components, I designed a suitable objective (training loss) that ensured the neural network extracts relevant features. As described before, the most commonly used self-supervised losses either predict some manually generated labels or meta-data already available for a dataset. While such objectives often show outstanding performance, they limit the applicability of a method to datasets where such labels or metadata are available in the first place. Specifically, for EM data this is rarely the case, with the *Platynereis* atlas not being

an exception.

After considering generating classification labels manually, I decided against it for the following reason: training a network to predict specific classes should lead the network to extract features that distinguish these classes, however, might have the adverse effect of discarding the intra-class variability. For example, if we train a network to distinguish muscles and neurons, it might not extract the features that would distinguish different types of neurons. Consequently, being biassed towards the predefined classes might also harm the network's ability to discover new categories in the data.

For this reason I designed a fully unsupervised, non-biassed objective that should both extract useful features and could be broadly applicable to any cell image dataset. The aim of such an objective should be twofold. Firstly, it should ensure that the most of the morphological information of a cell is present in the extracted representation. Secondly, it should map the representations of morphologically similar cells close together and dissimilar cells further apart, i.e. relative distances between cells in the representation space would reflect their morphological similarity.

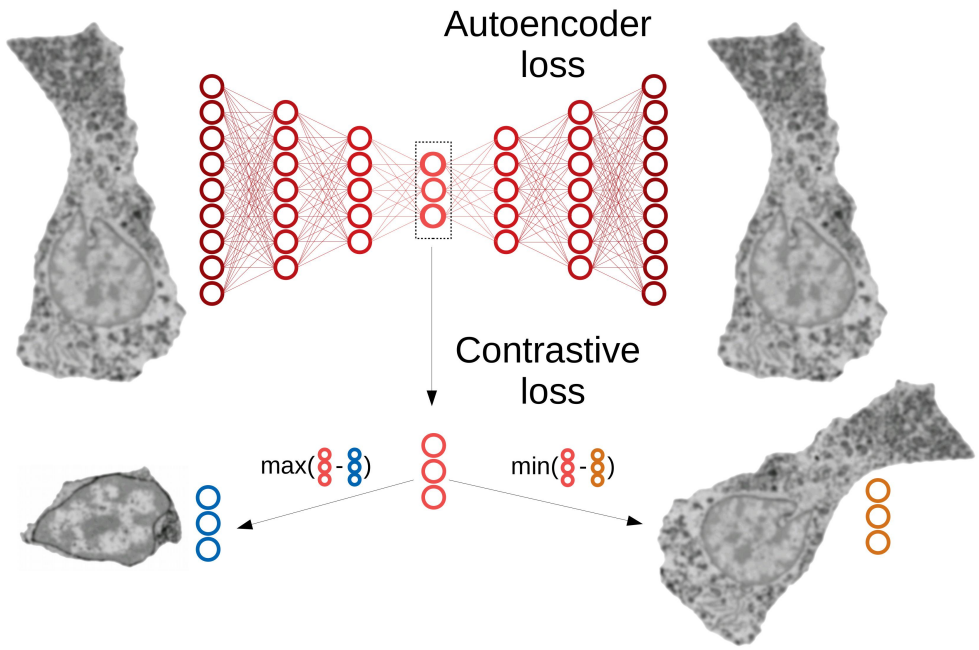


Figure 5.3: Training objectives for extracting coarse and fine texture features. An autoencoder neural network consists of an encoder that generates a feature vector from the input volume and a decoder that is trained to reconstruct the input volume from the extracted feature vector. Additionally, a contrastive loss is applied to the same feature vector that encourages the representations of modified views of the same cell to be closer to each other than to a different cell.

Running multiple experiments helped me identify two fully unsupervised training losses, which, when applied simultaneously, optimise for these two objectives (Figure 5.3). The first one is an autoencoder reconstruction loss [147] that ensures the model learns a comprehensive representation of the data. An autoencoder model normally consists of two parts: an encoder that takes a full cell as an input and compresses it into a low-dimensional representation vector, and a decoder that takes this low-dimensional vector and aims to reconstruct the complete cell from it. Thus, the training objective is to reconstruct the cell as accurately as possible. In order to do that the encoder has to learn an exhaustive, but a concise representation.

The second loss is a contrastive learning objective called NT-Xent loss [152] that enforces distances between the learned representations to be meaningful. In classical contrastive learning, for a single training run one needs three samples: an anchor sample (a cell of interest), a positive and a negative sample (a cell similar and dissimilar to the cell of interest respectively). During the training an encoder sequentially extracts a low-dimensional representation of each cell and the contrastive loss encourages the distance between the anchor and the positive cell to be closer than the distance between the anchor and the negative one. However, in a fully unsupervised setting, we do not know which cells are similar. That is why, the anchor and the positive cell can be replaced by two altered views of the same cell (orange representation in Figure 5.3). Assuming that these altered versions should still be closer to each other than to any other cell in the dataset (even if it is from the same cell type), one could just sample a random cell to be used as a negative sample (blue representation in Figure 5.3).

In the training pipeline I combined these both losses, i.e. they share one encoder. With this they also balance each other: the contrastive loss makes sure the autoencoder loss is not extracting meaningless features (e.g., angle and position), while the autoencoder loss ensures the triplet loss does not ignore features essential for reconstructing the cell (Figure 5.3).

5.2.3 Cell class annotation

While the training procedure was designed to be free of any manual labels, further analyses, including quantitative validation of feature quality, still required having some annotations. In order to generate manual labels, as a part of [65], Rachel M. Templin, Kimberly I. Meechan and myself have annotated around 1000 cells in the EM volume of *Platynereis*.

For this I first defined seven morphological categories of cells. These categories, however, do not necessarily define cell types, but rather comprise groups of morphologically distinct, easy to distinguish cells. The defined groups were neurons, muscles, epithelial cells (cells outlining the animal surface or the foregut opening), midgut cells (all cells located in the animal midgut), dark neurosecretory cells (cells having notably dark homogeneous nuclei and cytoplasm), secretory cells (any cells containing abundant organelles/inclusions that could indicate secretion, such as extensive ER and Golgi, vesicles, etc) and ciliary band cells (Figure 5.4).

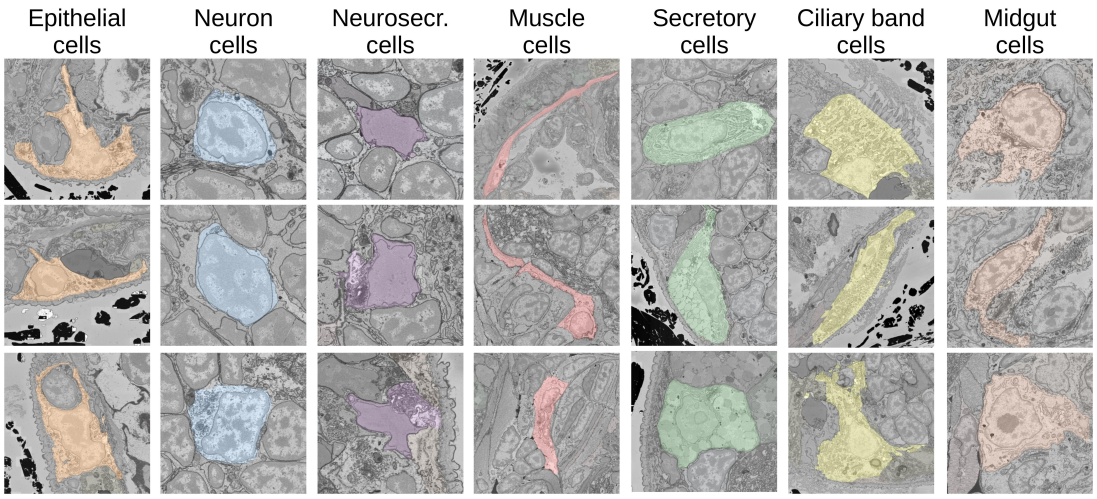


Figure 5.4: Manually annotated morphological cell classes. For each class 3 example cells are shown.

Afterwards, I sampled 10% of all the animal cells and Rachel M. Templin, Kimberly I. Meechan and myself assigned one of the 7 labels to each cell. Cells for which we were not sure about the label assignment were skipped. Even though cells were selected randomly, a dataset bias might have been created by annotating only the cells we were confident about. Additionally, the random selection led to a high class imbalance, though proportional to the biological disparity of cell types in the animal: more than half of the animal (65%) turned out to be neurons, 16% and 11% - epithelial and muscle cells, respectively, and only 3.5% - dark cells, 2.38% - midgut cells, 1.80% - secretory cells and barely 0.35% - ciliary band cells. In order to mitigate this imbalance I later added multiple cells from underrepresented classes. More specifically, I ensured that each class contained at least 60 representative cells, with the exception of ciliary band cells, since the animal only possess 30 of them at this developmental stage.

5.2.4 Predicting cell classes

The created annotations enabled me to evaluate whether the learned features were good enough to distinguish these 7 classes. In order to achieve this I trained a simple logistic regression classifier on the extracted representations, using a balanced subset of the cell classes as labels. The classifier works by assigning different weights to input features and then finding a linear separation between the given classes in the feature space. Its advantages include simplicity and possibility to train it with a few samples. Nevertheless, it can learn to ignore the features useless for a specific task by assigning them a close-to-zero weight. Therefore, one can expect a good classifier performance if at least a subset of the learned features is correlated with the annotated cell classes.

The model was trained using 5-fold cross validation and performed with an accuracy of 96%, which proves that the extracted features describe cellular morphology well enough to allow for recognition of these classes. The missclassifications made by the model (Figure 5.5, S3) indicate that, for example, the secretory class was more difficult to learn, than the others. However, this was an expected outcome, given how broad and morphologically heterogeneous it is.

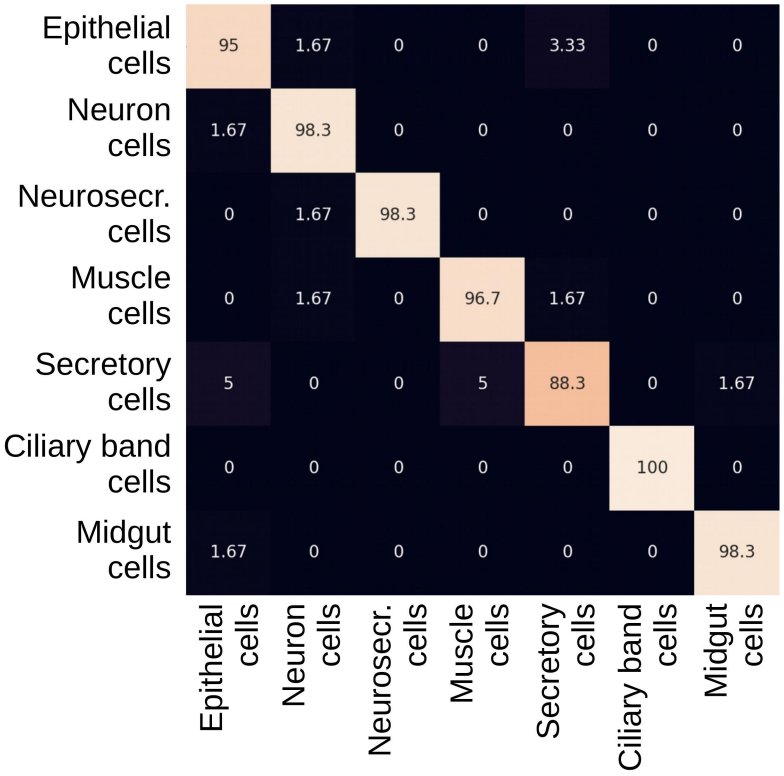


Figure 5.5: Confusion matrix for the logistic regression predictions. The rows correspond to the annotated labels, the columns show the network predictions (in %).

In order to ensure that this high performance is not just a result of a possible dataset bias, I used the trained classifier to predict the classes for all cells in the animal (Figure 5.6). The results indicate that the dataset with approximately 60 cells per class was sufficient to train a model with good generalisation. Specifically, for the class of epithelial cells I trained the classifier on the cells of the outer animal surface only, but it correctly predicted the epithelial cells of the foregut opening and the chaetae cells. Even more surprisingly, some cells in the midgut of the animal were confidently predicted as neurons. Subsequent literature research and visual examination helped me conclude that these cells might indeed be enteric neurons.

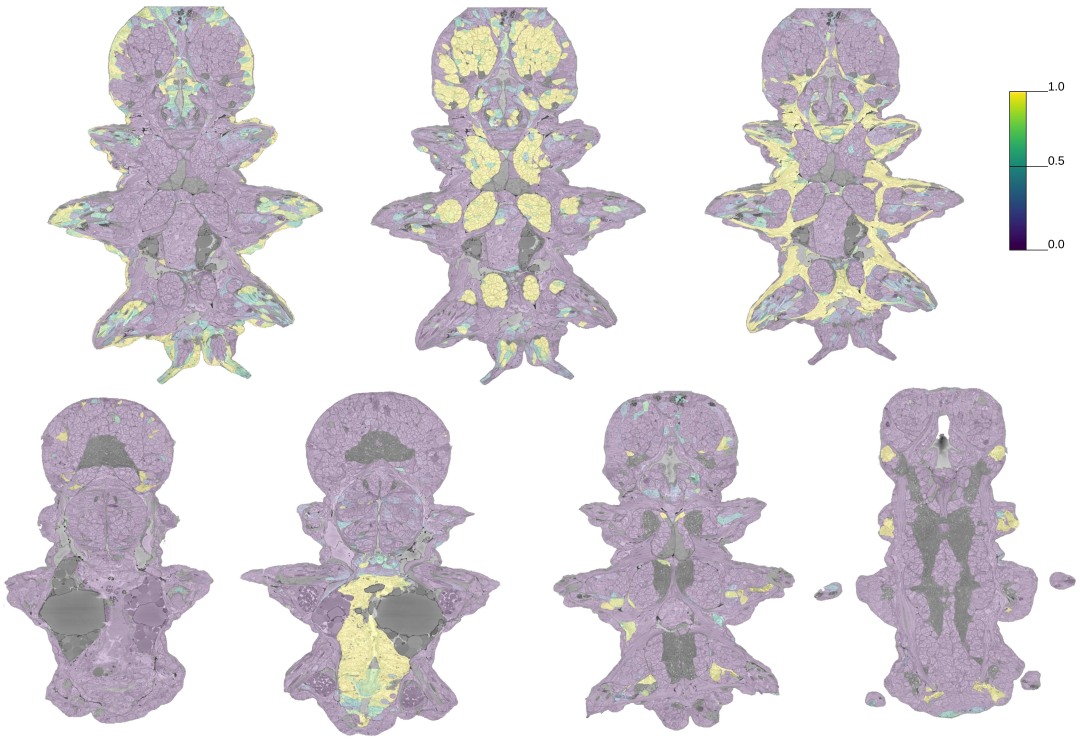


Figure 5.6: Predictions of the logistic regression model on the whole animal. From left to right: top row: epithelial cells, neurons, muscles; bottom row: neurosecretory cells, midgut cells, secretory cells, ciliary band cells.

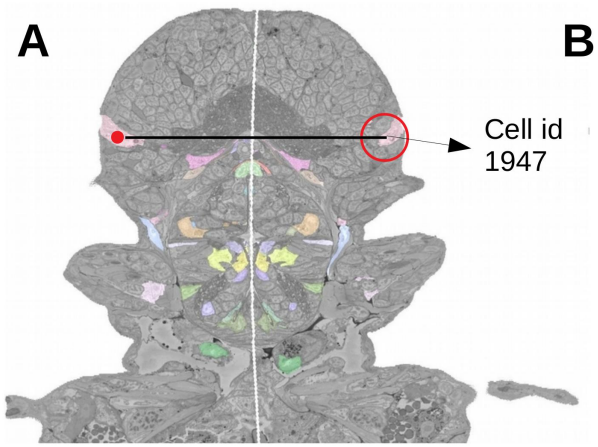
I further wanted to compare MorphoFeatures to the morphological features from [65]. These are a comprehensive set of manually defined features, including diverse shape, intensity and texture features extracted from the cell, nucleus and chromatin segmentations. Training a logistic regression on this set showed an accuracy of 94%. In general, the classifier made similar mistakes and also had a high missclassification rate for the secretory cell class. However, the explicit features showed slightly worse performance in distinguishing muscle and epithelial cells.

5.2.5 Finding bilateral neighbours

The fact that both sets of features performed so well on the task of morphological class prediction indicates that more evaluations are needed in order to better assess the different feature extraction pipelines. I further referred to the bilateral pair finding metric, designed by Kimberly I. Meechan in [65] to estimate the quality of the morphological features extracted manually.

The metric is based on the fact that *Platynereis* is bilaterally symmetric, thus for each

cell we would expect to be able to find an almost identically looking symmetric neighbour on the other side. In the previous chapter I described a manually labelled set of such pairs, based on which Kimberly I. Meechan created the metric, extending it to all the cells. For this David Puga annotated the midline of the animal, which was used to approximately estimate the location of a potential symmetric neighbour on the other side. However, since it is not possible to always find a symmetric neighbour at the mirrored location, a search radius was introduced. This radius was based on the deviation from the mirrored location for the manually annotated cell pairs.



Ranking (by distance)	Distance (feature space)	Cell id
Closest #1	0.23432	3546
Closest #2	0.74643	8654
Closest #3	1.46657	1675
Closest #4	2.74842	1947
Closest ...	...	....

Figure 5.7: Bilateral distance calculation. **A.** For a selected cell (red dot) all the cells within a certain radius of the symmetric location are considered as potential symmetric partners (red circle). **B.** All the other cells in the organism are ranked based on their feature similarity to the selected cell. From the potential symmetric partners, the one with the smallest ranking is chosen. Its rank is the bilateral distance of the selected cell.

Then for each cell the so-called bilateral distance is calculated as follows: first, one calculates the distance from a cell of interest (Figure 5.7A, red dot on the left side of the animal) to all the other cells in feature space. Based on this distance, a ranking of all the cells is calculated in relation to this cell, with the closest cell having rank 1, the second closest - rank 2, etc (Figure 5.7B). Then the location of the cell is projected to the other side of the animal. All cells within the previously mentioned search radius from this location are considered to be potentially the correct symmetric neighbour (Figure 5.7A, red circle on the right side of the animal). For each of these cells its ranking in relation to the cell of interest is checked. The cell with the minimal rank is considered to be the true symmetric neighbour, and its rank defines the bilateral distance of this cell. This procedure is repeated for all cells in the animal, and the bilateral distances are averaged to get one value for a set of features.

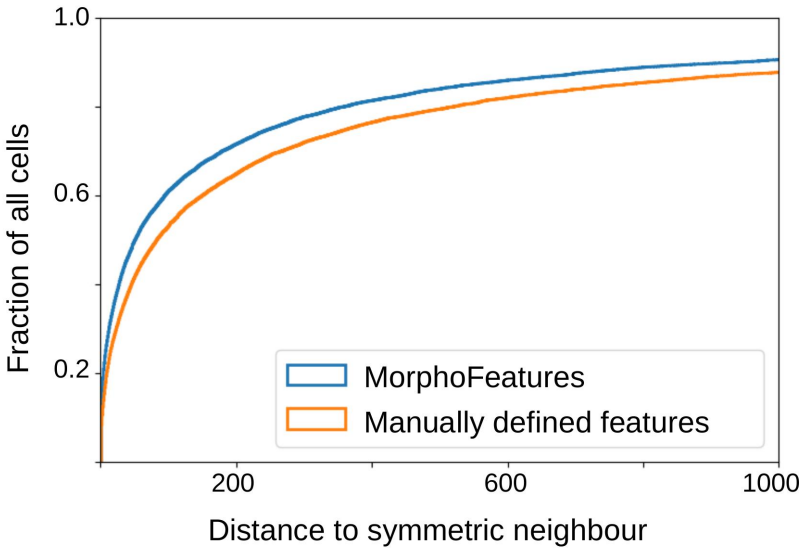


Figure 5.8: Bilateral distance distribution. For MorphoFeatures and the manually defined features from [65] fraction of cells having bilateral distance within a certain range is shown.

According to this metric, MorphoFeatures perform 38% better than the manually extracted features from [65]. The distribution of bilateral distances for all the animal cells is visualised in Figure 5.8. This shows that while both sets of features perform reasonably well on the easy task of cell classification, the task that requires more precise morphological description shows the clear superiority of the features extracted by the neural network pipeline.

5.2.6 Understanding extracted features

One of the main advantages of using neural networks as image feature extractors is the fact that a network extracts the features directly from the data and makes its own decisions about which features are important for a given task. However, this also makes it difficult to 'decode' the features afterwards. This especially hinders neural networks prevalence in the medical imaging field, where understanding what exactly led to a specific diagnosis is often not less important than the diagnosis itself. While not as vital, interpretability is still desired for biological data to be able to better characterise, for example, what exactly is different between the morphological cell types.

Thus I asked myself if it is nevertheless possible to get a grasp of which exact morphological properties were learned by the network. In order to estimate it qualitatively, I selected a random set of features and visualised which textures show the minimum and maximum value of this feature. Contrasting these textures, I presumed, could give me a hint at what the feature actually represents.

Surprisingly, many, though not all, feature dimensions seemed to be visually interpretable. For example, taking a look at the coarse texture features (Figure 5.9, upper row), for the first visualised one, one can clearly say that the feature minimum is present in elongated cells with the nucleus taking up almost the whole cell volume. The maximum of this feature is found in short, but wide muscle cells. For nuclei, one of the first coarse texture features I looked at showed its minimum in the cells with large spherical nuclei of uneven surface and its maximum in smoother, more elongated nuclei.

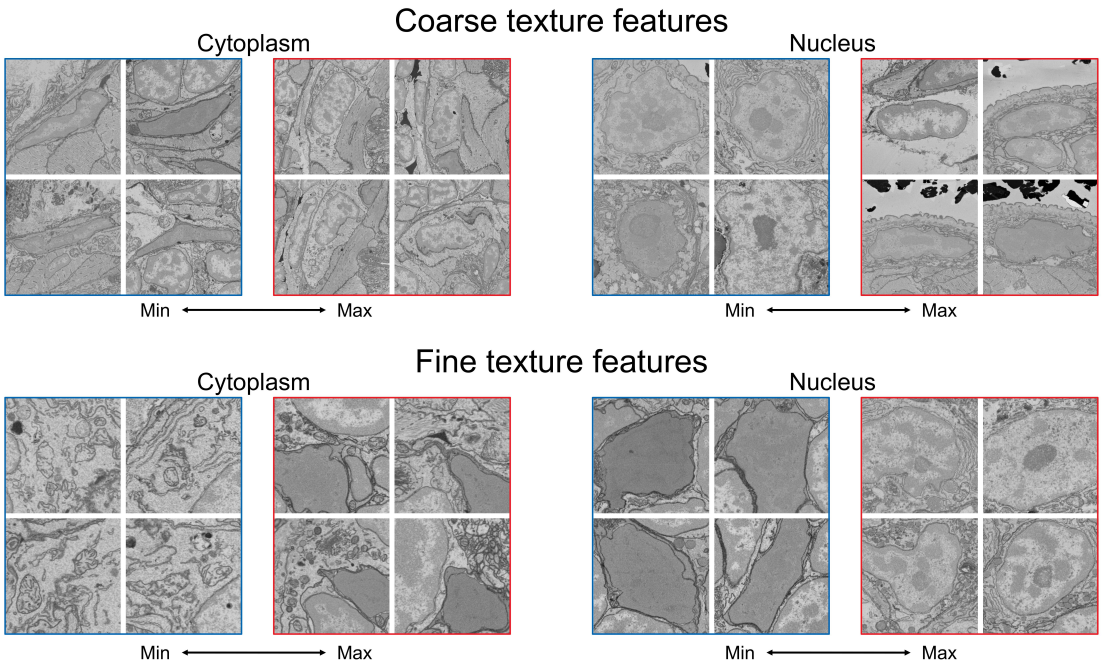


Figure 5.9: Visualisation of the learned features. For a random feature, four cells that show the minimal (blue) and the maximal (red) value of the feature are shown.

Visualising fine texture features showed that they indeed seem to describe fine ultrastructure (Figure 5.9, lower row). For example, a randomly chosen cytoplasm feature could be used to discriminate between patches densely occupied by mitochondria and Golgi cisternae and patches of evenly dark cytoplasm of dark neurosecretory cells. Similarly, one of the first nuclear fine texture features that I sampled, differentiated between evenly dark nucleoplasm and nuclear textures with distinct contrastive nucleoli.

I did not know in advance what each of the learned features correspond to. Still, a simple manual inspection could already let me correlate many of the examined features to some visually understandable morphological properties of cells and nuclei.

5.2.7 Visually analysing representations

After showing that MorphoFeatures perform well in classifying morphological cell categories and finding symmetric neighbours, I wanted to check whether the MorphoFeatures space fulfils the original objective that similarly looking cells should be grouped together. For this, I first generated a representation of the multidimensional MorphoFeatures space that could be visually examined. In order to simultaneously visualise the MorphoFeatures of all the animal cells I used UMAP [153] - a dimensionality reduction method that tries to preserve local distances between the data points. Having a two-dimensional representation, I marked the cells for which manual annotations were available (Figure 5.10).

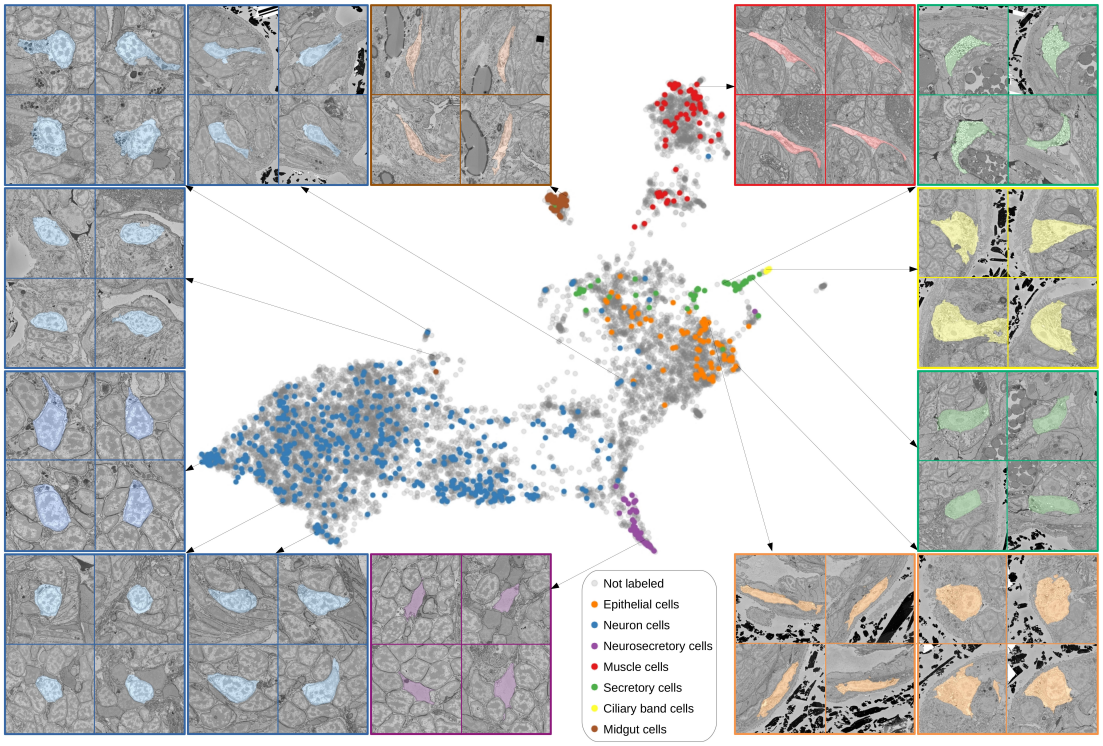


Figure 5.10: UMAP representation of all cells in the animal based on their MorphoFeatures. The manually annotated cell classes are shown in colors. Each panel shows a cell randomly sampled on the UMAP representation and its three nearest neighbors visualised in the EM volume.

One can notice that most cell classes are consistently grouped together. The biggest exception is the class of secretory cells. However, this is expected, given that the group comprises multiple types of them, which are not always morphologically similar. Cell groupings visualised in 2D also suggest that many classes can be split into multiple, po-

tentially morphologically distinct subgroups. For example, one can notice two separate groups of muscles in the upper right part of the UMAP representation. Notably, neurons also display a wide range of potential morphological types, despite the fact that their description was based on the soma morphology only.

I further wanted to investigate whether cells that are close in the representation space are also visually similar. For this, I sampled a set of random cells and visualised the three closest neighbours in the created representation space. As is shown in Figure 5.10, the corresponding neighbours look notably similar. For instance, one can admire a group of neurons with pigment granules (upper left blue panel), thin extended midgut cells (brown panel), muscles (red panel) and ciliary band cells with characteristic nucleoli and abundant mitochondria. A clear separation can be noted between the groups of thin extended and large spherical epithelial cells (orange panels). Secretory cells are also grouped according to their appearance, with two groups visualised: 'bright droplets' cells [121], rich on round bright vesicles, and parapodial spinning gland cells [121] containing plentiful ER sheets. Interestingly, there is a group of cells that was manually annotated as neurons, located closer to the epithelial cells in the representation space (upper right blue panel). Manual inspection of this group, performed by Detlev Arendt, suggested that these are sensory neurons, which are both evolutionarily and morphologically close to epithelial cells.

This simple visual investigation has shown that the distances between cells in the MorphoFeature space indeed reflect their morphological similarity. Such analysis could prove extremely useful for investigating new EM volumes, e.g. in case a researcher finds a cell with a particular morphology and is interested in locating similar cells across the animal volume. Furthermore, such fine groupings can potentially be used to separate and describe morphological cell types of an animal.

5.2.8 Clustering the MorphoFeatures space

Having shown that local distances in the MorphoFeatures space indeed reflect visual similarity, I was interested whether the feature space can be further partitioned into groups of cells with consistent morphology. These smaller partitions might then represent the true morphological cell types. In order to do that I used the Leiden algorithm for community detection [154], which is widely used in the field of single-cell transcriptomics. It is important to mention that in order to cluster the data one has to manually set the desired clustering resolution level that influences the number of detected clusters. I first aimed at generating rather big general clusters, making it easier to visually inspect if clustering makes sense. However, afterwards, whenever needed, I could increase the clustering resolution to get finer clusters.

This approach initially resulted in 16 general clusters (Figure 5.11). One of them I discarded after manual inspection, since it only contained cells with split segmentation errors. Out of the remaining 15 clusters, the biggest proportion (clusters 1-8) was taken by neurons, reflecting their prevalence in the animal at this stage. Some of the neuron clusters showed remarkably consistent location in the animal. For example, the location

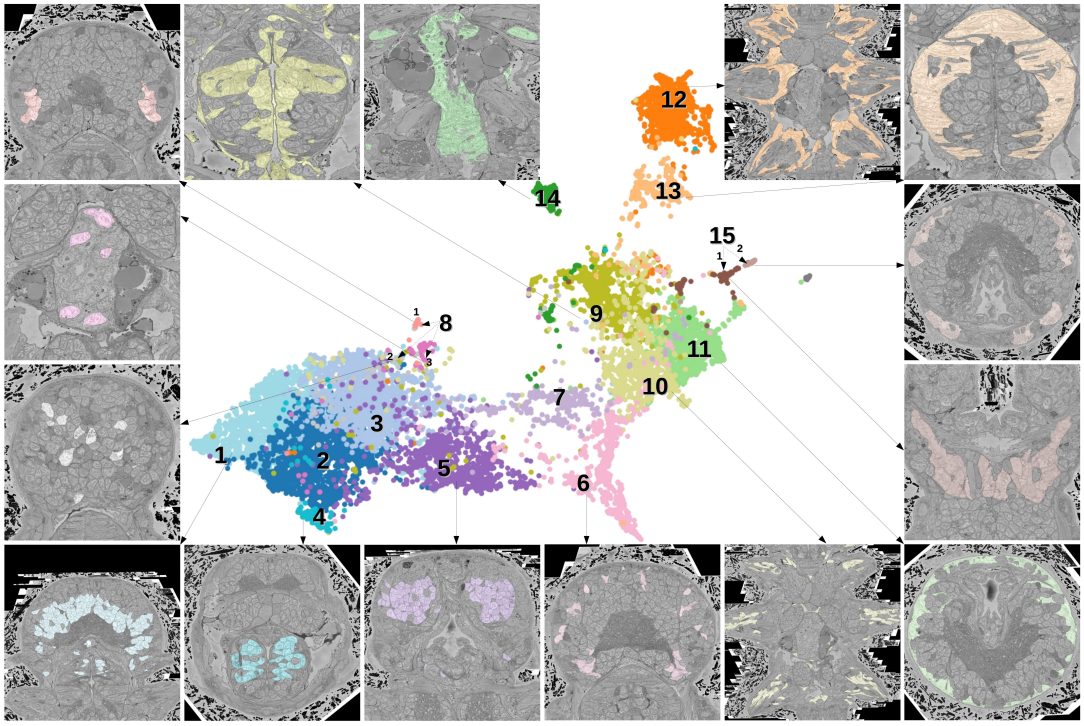


Figure 5.11: Morphological clusters plotted on the UMAP representation from Figure 5.10. The panels show the location of some clusters in the EM volume.

of cluster 4 (visualised in Figure 5.11) indicated it comprises foregut neurons.

Cluster 8 showed high morphological heterogeneity, indicating it might be composed of multiple cell types, which were grouped together because each type contains only a few cells. A finer clustering resulted in three subclusters. The first of them (8.1) consisted of arch-shaped cells with big black inclusions that I identified as rhabdomeric photoreceptors of the adult eye (Figure 5.12A). The second subcluster (8.2) contained various neuron-shaped cells with smaller black pigment granules. The third one (8.3) comprised cells, whose shape and cytoplasm resemble neurons (Figure 5.12B), however, their location in the animal was restricted to the developing midgut (Figure 5.12C). This let me speculate that these cells were the enteric neurons previously confirmed to autonomously regulate digestion in *Platynereis* [8].

The clusters 9-11 were epithelial cells, with cluster 9 located predominantly in the foregut, 10 - in the chaetae and 11 - on the outside surface of the animal. Cluster 12 was composed of most of the muscle cells with the exception of foregut muscles that made up cluster 13. Cluster 14 included all midgut cells of the animal, and cluster 15 grouped together secretory and ciliated cells, that could be separated by finer clustering.

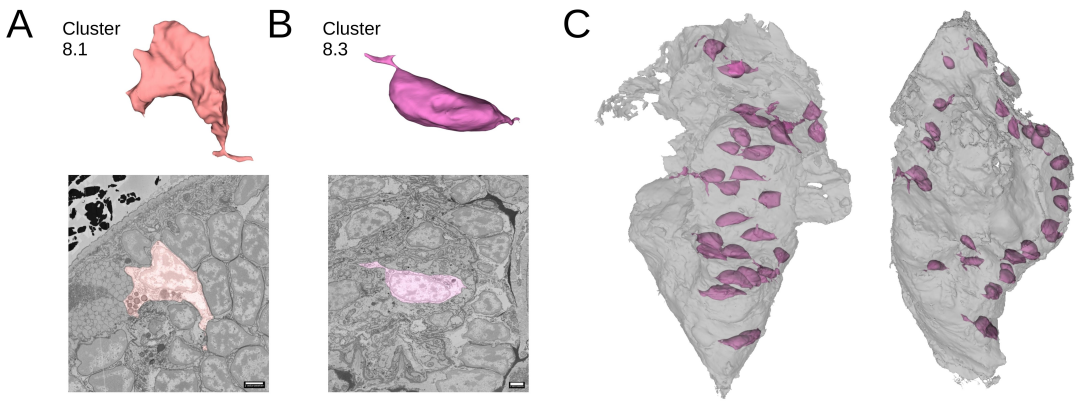


Figure 5.12: Analysing subgroups of cluster 8. **AB.** Shape and ultrastructure of an average cell of subcluster 8.1 (rhabdomeric photoreceptors, **A**) and subcluster 8.3 (enteric neurons, **B**). **C.** Location of the enteric neurons subcluster in the animal midgut.

This cluster analysis showed that the MorphoFeatures space can be meaningfully partitioned into morphologically coherent groups that mostly display consistent location in the animal volume. I expected that such groups could represent morphological cell types that should be relatable to genetic cell types.

5.2.9 Characterising morphological types genetically

To investigate whether the morphological grouping described above correlates with gene expression, I used the expression maps from [65]. More specifically, I was interested to examine if the morphological clusters also have differentially expressed genes, which in the perfect case would be expressed in all cells of the cluster and nowhere else in the animal.

It is important to note that not all clusters would have such genes for the following reasons. Firstly, though the gene expression atlas in the *Platynereis* resource is constantly expanding, it is only available for a limited number of genes that cover the animal nonuniformly, so some of the truly differentially expressed genes might not be present in the resource. Secondly, we would expect some morphological differences between cells to arise from their location and neighbourhood, thus, not being manifested in their gene expression. Finally, due to the fact that neuron segmentation was only available for their cell bodies, our morphological analysis is not taking into account the neurite morphology that plays a crucial role in defining neural cell types.

Nevertheless, many clusters displayed a clear genetic signature (Figure 5.13, S4), showing that the morphological types arising from the MorphoFeatures space might indeed correspond to genetic cell types. For example, the rhabdomeric photoreceptors of the adult eye (cluster 8.1) were shown to specifically express the bHLH transcription factor *mitf* and *globin-like* - a cytoglobin (oxygen transferring molecule) that in *Platynereis*

might play a role in protecting cells from oxidative stress [155]. Cluster 5 expresses he bHLH transcription factor *asci* and TRV-channels *pkd2*, *trpV4* and *trpV5*, which indicates that the cluster contains sensory neurons. Cluster 6, which is composed of dark nurosecretory cells, can be described by the expression of prohormone convertase *phc2* and atrial natriuretic peptide receptor *anpra*, which are both neurosecretion markers.

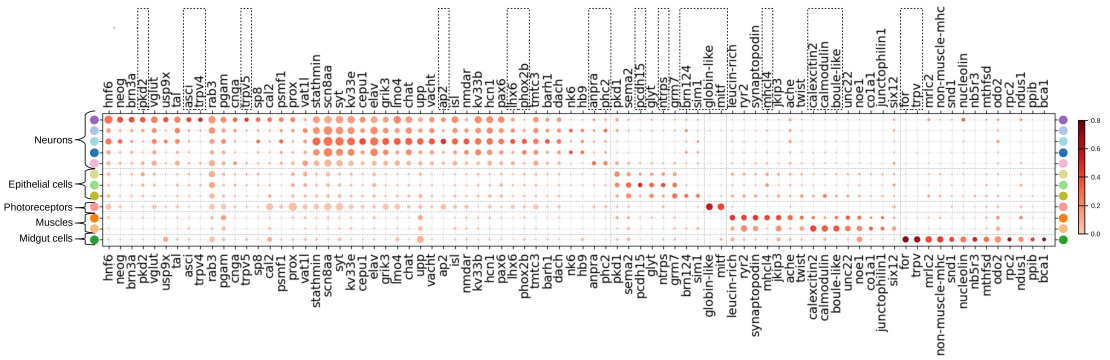


Figure 5.13: Gene expression analysis of the morphological clusters (dot plot). The dot size reflects the fraction of the cluster expressing a gene; the dot colour reflects the gene specificity to a cluster (see Methods).

The 3 clusters of epithelial cells could as well be differentiated by the genes they express. For example, cluster 11, comprised of surface cells, shows specific expression of protocadherin (*pcdh15*) and neurotrypsin (*ntrps*). It differs from cluster 9 of cells outlining the foregut opening, that expresses the homeodomain factor *pou3* (*brn124*), shown to participate in stomodeal ectoderm in sea urchin [156] and bHLH transcription factor *sim1*. The cluster 10 of chaetae cells shows a higher expression of the receptor *pkd1*.

While inspecting genes that can be used to distinguish the two groups of muscles, I noted that the bigger cluster 12 specifically expresses such genes as *leucin-rich*, *synaptopodin*, *ache* and, most importantly, myosin heavy chain *mhcl4*, which is specific to skeletal (striated) muscles. Interestingly, the foregut muscles (cluster 11) develop as smooth muscles, but transform to striated at a later stage of development [8]. Thus, the absence of *mhcl4* in this cluster indicates that this switch is yet to occur. However, this cluster of foregut muscles showed specific gene expression of the *calexcitin2* (calexcitin-like sarcoplasmic calcium-binding protein *Scp2*) [157], *calmodulin* and of the Boule homolog *Boll* (*boule-like*), which indicates that this group of muscles might have developed a calcium excitation mechanism different from striated muscles.

Another cluster with prominent specific gene expression was the developing midgut (cluster 14) which showed highly unique expression for multiple genes, including *for*, *trpv*, *mrlc2*, *non-muscle-mhc*, *snd1*, *nb5r3*, etc. Looking at gene expression plotted on the UMAP representations (Figure S4) I suspected that this cluster might display high genetic variability, thus, aimed to explore it further.

5.2.10 Exploring midgut cell types

In order to characterise the heterogeneous midgut cells, I clustered this group with a higher resolution, which resulted in three subclusters (Figure 5.14A). One of them appeared to be located outside of the midgut (Figure 5.14B), however, the cells comprising it looked visually similar to the midgut cells (Figure 5.14C). This subcluster did not have any specific gene expression, so I further focused on the remaining two subclusters.

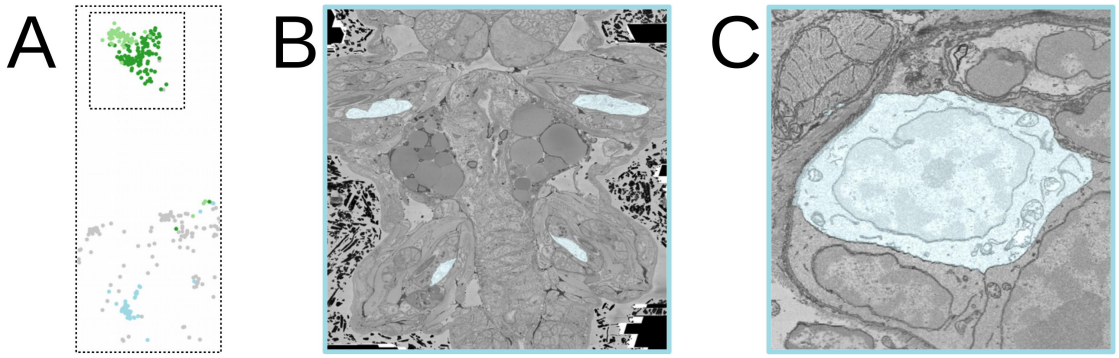


Figure 5.14: Analysing subgroups of the cluster 14. **A.** Finer clustering of the midgut cluster results in three subgroups. **BC.** Location and appearance of the subgroup located outside of the midgut.

To characterise the other two subclusters, I calculated their gene specificity within the midgut cluster (Figure 5.15). This confirmed that they, indeed, have different gene expression. Analysing the specific genes, I noticed that one of the subclusters expressed the smooth muscle markers *non-muscle-mhc* and *mrlc2*. The other one showed specific expression of a chitinase related to chitotriosidase (*nov2*) and the zinc finger transcription factor *mecom* (*Prdm16*), which was shown to regulate homeostasis in epithelial cells of the gut [158]. I hypothesised that the first subcluster could later differentiate into the muscle cells and the second one into the epithelial cells of the *Platynereis* midgut.

In order to further confirm this hypothesis, I visualised both clusters and the above-mentioned genes in the animal volume (Figure S5). The subclusters were proven to occupy separate territories, consistent with the expression of the genes. This showed that such type of exploratory morphological analysis can indeed reveal novel distinct morphological groups with differing gene expression, which would not be apparent during browsing the data, unless one knows in advance which groups to look for.

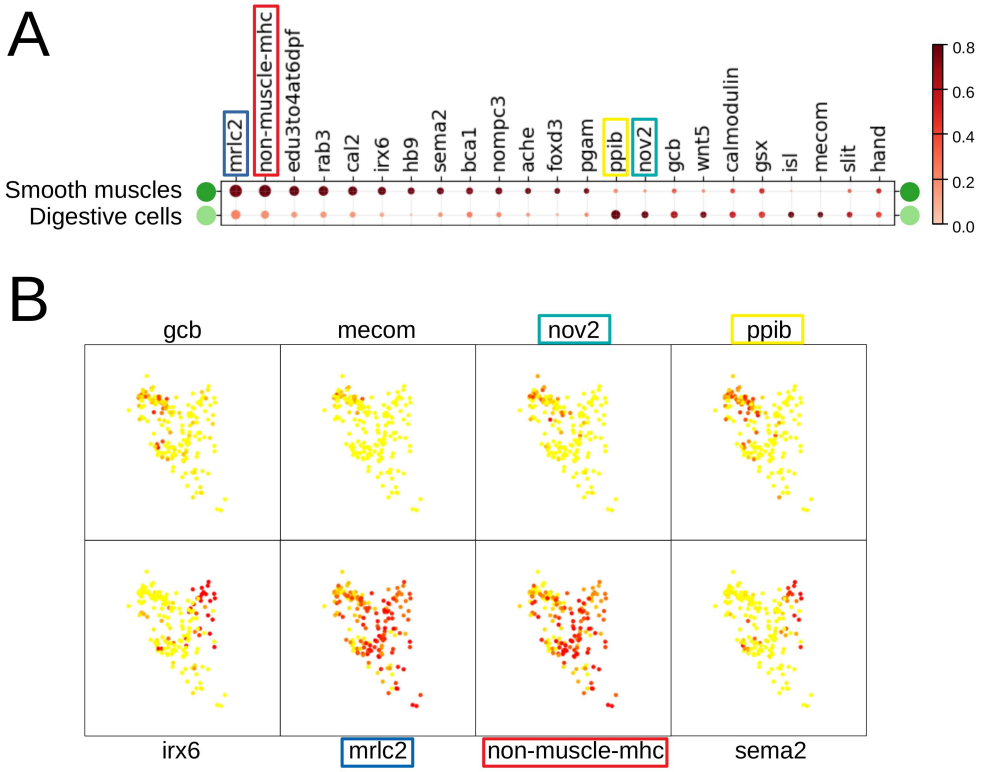


Figure 5.15: Gene expression analysis of the midgut cell types. **A.** Gene expression dot plot. The dot size reflects the fraction of the cluster expressing a gene; the dot colour reflects the gene specificity to a cluster (see Methods). **B.** Some of the genes from **A** plotted on the UMAP representation of the midgut region.

5.2.11 Investigating nuclear features separately

The fact that the pipeline explicitly combines different essential components of cellular morphology makes it possible to investigate these components separately. In order to use this opportunity, Detlev Arendt wondered how descriptive the pure nuclear features are. To investigate this, I separated the nuclear features (shape, fine and coarse texture) and performed a similar analysis on them. First, I generated a UMAP representation of the combined nuclear features (Figure S6) that showed that the major morphological categories were separated surprisingly well, given that no cytoplasmic features were taken into account. For example, the ciliary band cells still form a coherent group in this representation. Even more surprising is that muscle and midgut cells are distinguishable from all the other cells based purely on their nuclei morphology.

I was further interested in how the nuclear features compare to the complete MorphoFeatures set on the quantitative tasks of cell class prediction and symmetric neighbour detection. On the first task they scored 92% accuracy (in comparison to 96% for MorphoFeatures and 94% for the manually defined ones), slightly losing accuracy in detecting epithelial and secretory cells. The latter was expected, given that the whole class was defined based on the cytoplasm appearance. On the symmetric neighbour detection task the performance of the nuclear features was 73% worse than of the full MorphoFeature set, but only 6% worse than the set of manually defined features from [65].

I was also curious if visualising separate components of nuclear morphology (coarse and fine texture and shape) could reveal anything peculiar about the data. So I plotted UMAP representations of these components separately; and by visual examination I noticed that only in the nuclear fine texture representation plot there were some muscle cells located closer to neurons than to other muscle cells (Figure 5.16). I visualised this group in the animal volume and noticed that all of them are the antennal muscles (shown on Figure 5.18). To see how justified it is to group these muscles with neurons, rather than with other muscles, I visualised an example of nuclear texture from (1) an antennal muscle, (2) a random oblique muscle and (3) a neuron close to the antennal muscles in the representation space (Figure 5.16). I noticed that indeed the nucleus of an antennal muscle cell looked more similar to the neuron nucleus, containing much less euchromatin and more consistent patches of heterochromatin. This illustrated how a simple visual examination of an extensive set of morphological features can lead to interesting observations.

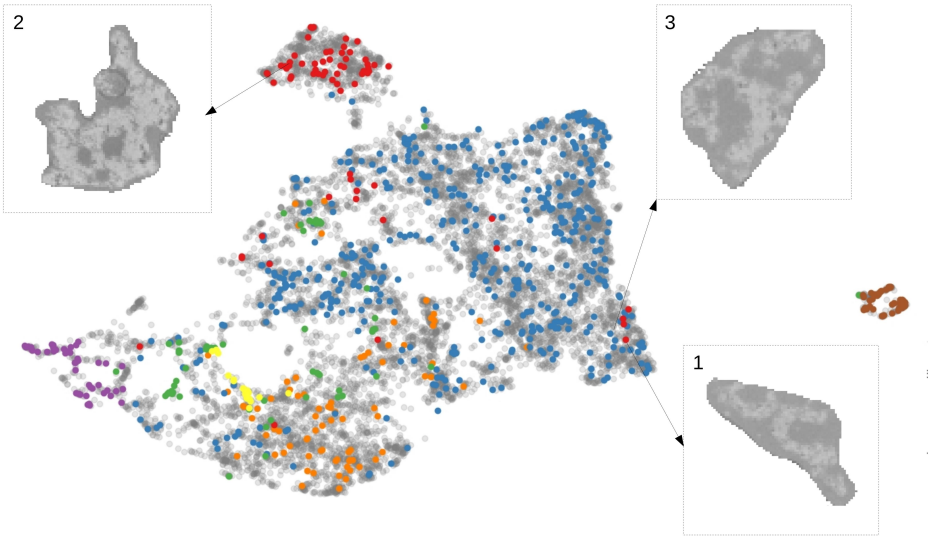


Figure 5.16: UMAP representation of all cells in the animal based on their nuclear fine texture features. The manually annotated cell classes are shown in colors. The panels show nuclei of an antennal muscle cell (1), an oblique muscle cell (2) and a neuron (3).

5.2.12 Comparing MorphoFeatures to explicitly defined features

I further wondered how MorphoFeatures compared to explicitly defined features in terms of resolving morphological groups and types. As Kimberly I. Meechan showed in [65], the set of explicit features showed good performance on separating all the manually labelled cells classes (e.g., neurons, muscles, midgut cells), except for the secretory cells, which, as mentioned before, one would not expect to be grouped together (Figure 5.17A).

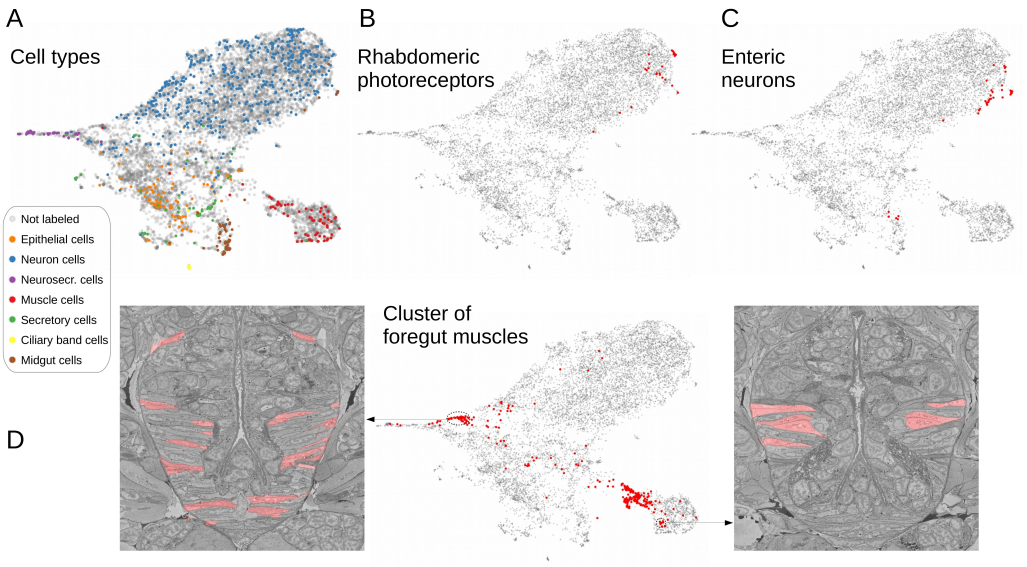


Figure 5.17: Comparing MorphoFeatures to the manually defined features from [65]. **A.** UMAP representation of the manually defined features. **BCD.** The morphological clusters of rhabdomeric photoreceptors (A), enteric neurons (B) and foregut muscles (D) plotted on the UMAP representation from **A**. The manually defined features split the group of foregut muscles into two subgroups. The location of these subgroups in the animal volume is visualised.

However, I wanted to inspect whether such features would also be able to define some of the clear morphological types that were discovered by clustering MorphoFeatures. For this I took the clusters with clear morphology that I was sure correspond to morphological cell types - rhabdomeric photoreceptors and enteric neurons - and plotted them on the UMAP representation of the explicitly defined features. In the representation of the explicit features both groups did not seem to form coherent clusters, hinting that their resolution might not be sufficient to distinguish these groups (Figure 5.17BC).

I further noticed that the explicit feature representation contained only one evident cluster of muscles. So I also plotted the cluster of the foregut muscles on this representation to see if these cells are grouped together (Figure 5.17D). While most of these muscles indeed showed consistent location close to the striated muscles, a big portion of them was

found in-between the dark neurosecretory cells and general neuron group. I investigated the specific features that made these two subgroups so different from each other and found that the subgroup close to neurons highly differed in their nuclear intensity features from all the other muscle cells.

This comparison showed that, while both sets of features capture the differences between broad morphological classes well, MorphoFeatures achieved a higher resolution on more precisely defined groups of cells. Moreover, the drastic split of the foregut muscles hinted that the manually defined features might be too sensitive to specific intensity characteristics.

5.2.13 Adding neighbourhood information

As previously shown in this thesis, cellular phenotype is mostly formed under the influence of a combination of a cell's internal genetic program and the environmental forces. In the case of multicellular organisms, such environmental forces are often coming from neighbouring cells. Cells in cellular neighbourhoods often influence each other to coordinate their development and to be able work together in order to fulfil a specific function. Thus, to describe a functional phenotype, it might be useful to describe the whole neighbourhood, rather than a single cell.

To generate a morphological description of cellular neighbourhoods in *Platynereis*, Johannes Hugger first investigated different ways of incorporating morphological information from neighbouring cells, using the symmetric neighbour detection task to evaluate how good a certain method is. He showed that the optimal way of integrating information from cellular neighbourhood is simply taking an average of the neighbours MorphoFeature vectors and appending it to the cell's own MorphoFeature vector. I further refer to the resulting representation as the MorphoContextFeature vector. This way of describing a cell together with its immediate surrounding improved the symmetric neighbour distance by 71%, showing, as expected, that local neighbourhood's morphology is more indicative of location than the morphology of single cells.

To estimate how well this description captures similar neighbourhoods, I generated a UMAP representation of the MorphoContextFeatures. Visualising the coherent groups of cells on this representation in the animal volume, I noticed that while MorphoFeatures separated morphological cell types, MorphoContextFeatures distinguished the whole tissues and organs of the animal (Figure 5.18). For example, the lower left part of the 2D representation corresponds to the animal foregut, that consists of the foregut epithelium (left orange panel), foregut muscles (lower red panel), foregut neurons (lower left blue panel) and a layer of neuron-like cells surrounding the foregut muscles (lower central blue panel). The representation also captured a group of muscles surrounding the foregut (left upper red panel), secretory (green panel) and epithelial (upper orange panel) parapodial cells and a group of antennal muscles (upper red panel), which participate in moving the chemosensory antenna on the animal's head. I also noticed that neurons got grouped into ganglia (Figure 5.18, the upper right part) and I decided to analyse this grouping further.

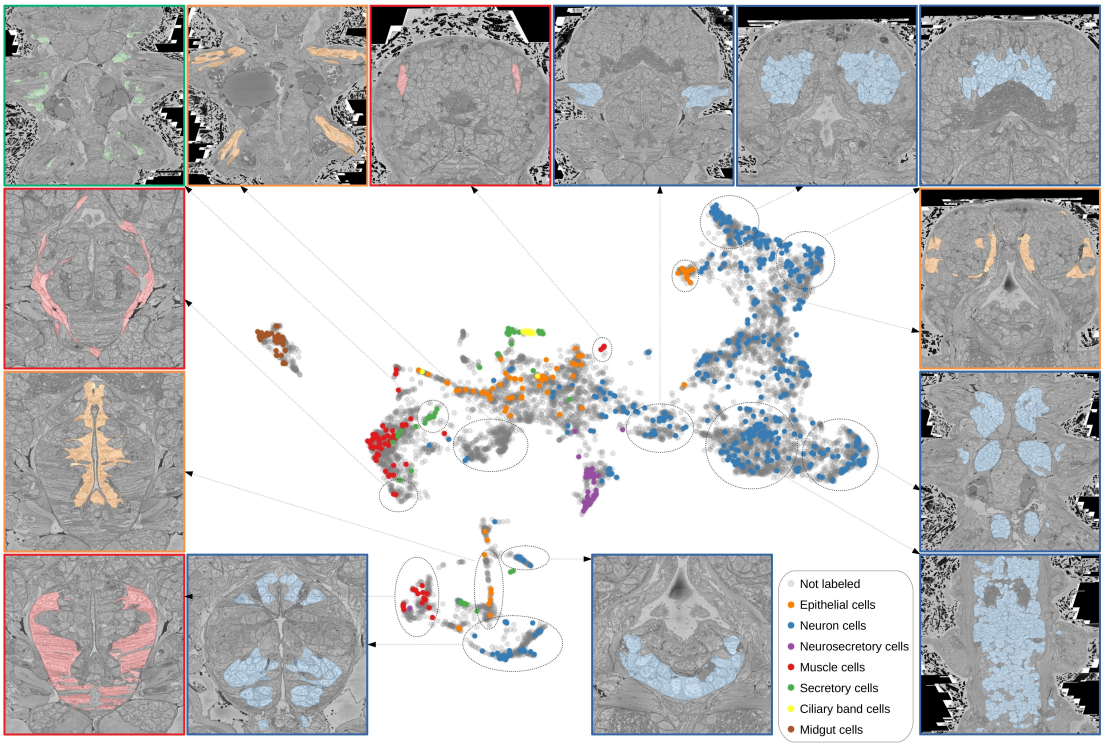


Figure 5.18: UMAP representation of the MorphoContextFeatures. The location of some groups of cells in the animal is visualized in panels. Panel colors represent classes cells belong to.

5.2.14 Distinguishing ganglia

To be able to better evaluate and characterise the ganglia grouping, I performed community detection on MorphoContextFeatures of all cells, and selected the clusters corresponding to the ganglia part of the UMAP representation. To annotate these groups, I relied on the manual ganglia segmentation based on tissue boundaries done by Hernando M. Vergara in [65]. In cases when MorphoContextFeatures clustering produced finer separation than the manual segmentation, I consulted Hernando M. Vergara and Detlev Arendt to refine the annotations. This helped me identify the groups of cells in the ganglia part of the UMAP representation as cirral, palpal and dorsal ganglia of the head neurons (upper blue panels from left to right), epithelial-sensory circumpalpal ganglion (orange panel), the peripheral ganglia and the ventral nerve cord (VNC) (two lower right blue panels).

I further compared the results of the MorphoContextFeatures clustering to the manual ganglia segmentation and to automatic ganglia detection by clustering of the available gene expression [65]. In general, all three methods yielded comparable grouping (Figure 5.19). However, in cases when no visible boundary was available, the manual separation was not possible, as for example can be seen for the dorso-posterior and dorsal-anterior ganglionic

nuclei and for the VNC and the peripheral ganglia. These groups are nevertheless well separated by both gene expression and MorphoContextFeatures clustering, with the latter method showing even higher resolution clustering of the dorsal ganglion. Similarly, as expected, the gene expression clustering showed worse performance on the regions with less gene expression available in the atlas. For example, the epithelial-sensory circumpalpal ganglion, defined both by manual segmentation and MorphoContextFeatures clustering, got merged with the neighbouring ganglia due to the lack of consistent gene expression.

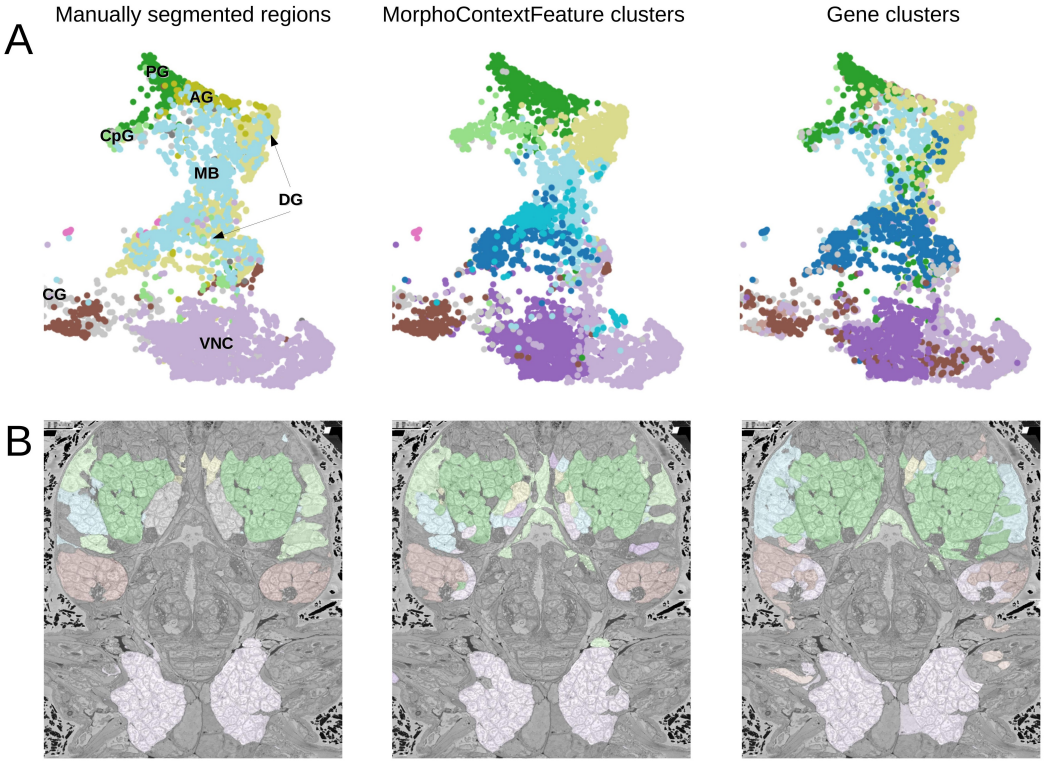


Figure 5.19: Automatic ganglia detection by MorphoContextFeatures. **A.** Ganglia defined by manual segmentation [65], MorphoContextFeature clustering and gene clustering are visualised on the ganglia part of the MorphoContextFeatures UMAP from Figure 5.18. **B.** The abovementioned ganglia are visualised in the animal volume.

To see whether the clusters defined by MorphoContextFeatures also display consistent expression profiles, I examined their specific gene expression (Figure S7). Interestingly, the above-mentioned circumpalpal ganglion, which was not detected by gene clustering, still shows high expression levels of epithelial genes not present in other ganglia, such as, for example, glycine neurotransmitter transporter *glyt*, semaphorin *sema2* and glutamate metabotropic receptor *grm7*. Despite the lack of clear tissue boundary, the VNC and the peripheral ganglia can be clearly distinguished by their gene expression, with the ventral ganglionic mass specifically expressing the homeobox transcription factors *nk6*, *hb9* and *pitxb* and the dorsal one having a higher *phox2b* expression.

5.2.15 Characterising foregut

To further explore the clustering of MorphoContextFeatures I shifted my attention to the foregut region (Figure 5.18, the lower left part). The clusters in this region included the foregut epithelium, foregut muscles, foregut neurons and the layer of neuron-like cells (Figure 5.20). To identify the latter group, I visualised its shape and location in the animal volume (Figure 5.20, in green). This structure is leaf-shaped and positioned below the brain neuropil between the posterior pair of adult eyes (Figure S8B), forming a collar around the foregut opening, and is surrounded by layers of muscle cells on both sides (Figure 5.20, in red and beige, Figure S8A). This position made Detlev Arendt assume that these neuron-like cells might be in fact neurosecretory, forming the infracerebral gland, supposedly regulating glucose levels in the animal blood [159, 160, 161, 162]. The latter might also hint that the cavity above the structure (Figure S8C) is a developing blood vessel. I identified specific genes for the foregut clusters (Figure S9), and noticed that there is a genetic difference between the foregut neurons and the infracerebral gland cells, with the former expressing sodium channel *scn8aa*, and the latter being distinct in expressing the homeobox transcription factor *hnf6*. I also observed that in comparison to the foregut neurons, the infracerebral gland cells are positive for proliferative EdU stainings between 4 and 5 days post fertilisation, indicating that these cells might be less differentiated.

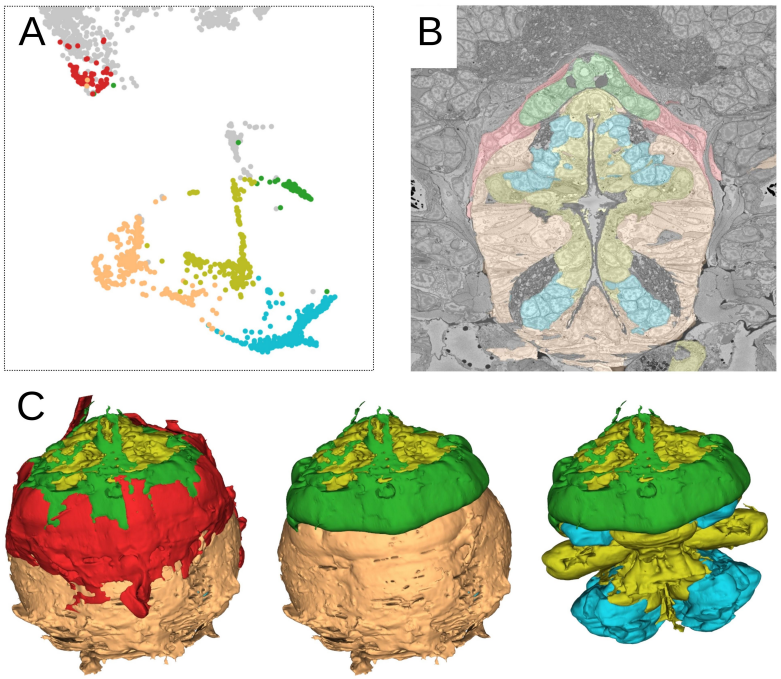


Figure 5.20: Foregut characterisation using MorphoContextFeatures. **AB.** Clustering of the foregut visualised on the MorphoContextFeatures UMAP from Figure 5.18 (A) and in the animal volume (B). **C.** 3D visualisation of the tissues defined by the foregut clustering. 3D volumes were generated by Johannes Hugger (adapted from [66]).

I also got interested in a group of muscle cells located on the outer surface of the infracerebral gland and surrounding the animal foregut (Figure 5.20, in red). To identify them I referred to their specific gene expression (Figure S9), however, they seemed to mostly display gene expression similar to the one of the foregut muscles. I suspected it might be so because this layer of muscles is relatively thin, thus, one could expect a high level of gene leakage from and to the neighbouring foregut muscles. Visualising some of the genes in the animal volume proved this suspicion to be true, and I identified the muscle junctional gene *junctophilin1* and the type I collagen gene *col1a1* as specific to the unknown group of muscles only (Figure 5.21). The type I collagen has been shown to be secreted by the axocord in *Platynereis* [163], indicating that this group of muscles could be the anterior extension of the axochord. This hypothesis was further supported by the location of the group - "between central nervous system and axial blood vessel" as described in [163].

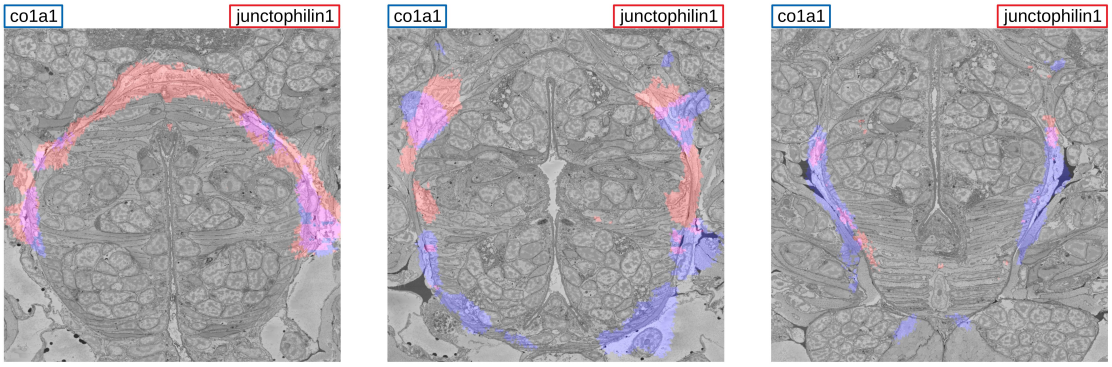


Figure 5.21: Gene expression characterising the axocord muscles. Collagen gene *co1a1* and muscle junctional gene *junctophilin1* are visualised in three different EM sections.

In general, the analysis of MorphoContextFeatures showed that combining the morphological information of a cell and its immediate neighbours reliably identified tissues, ganglia and organs. Further exploration of the groups revealed by these morphological descriptions helped me characterise structures that were not yet noticed in the EM volume of the animal.

5.3 Discussion

5.3.1 Contributions

In this chapter, I developed a method for automatic unbiased extraction of texture features from electron microscopy data using deep neural networks. The designed pipeline can be trained in a fully unsupervised way, i.e. does not require any annotations or metadata available. For this I combined two training objectives - contrastive learning and

autoencoder reconstruction - that ensured that the extracted features are representative of cellular morphology and capture the morphological distances between the cells. Then I applied the developed pipeline to a whole-animal segmented 3D EM volume of a marine annelid *Platynereis dumerilii* that contains multiple morphologically distinct cell types. This produced high quality features that perform better than a set of extensive manually designed features [65] on the proxy tasks of cell classification and symmetric neighbour detection. I also showed that many dimensions of the learned features can be interpreted through visualisation of the inputs corresponding to feature minima and maxima.

I further analysed the extracted features and demonstrated that they not only capture fine morphological peculiarities, but can also be used to recognise morphological cell types, which can be further described genetically. For example, I characterised two types of muscles, three types of epithelial cells and multiple types of neurons. Furthermore, I detected rare cell types, such as rhabdomeric photoreceptors, enteric neurons and two populations (epithelial and muscular) of the developing midgut cells. Additionally, I analysed the combined morphological representations of cellular neighbourhoods, generated by Johannes Hugger, and showed that these correspond not to cell types, but rather to tissues and organs, such as foregut or ganglia. I used these representations to detect structures previously unnoticed in the volume, such as infracerebral gland and axocord muscles, and illustrated how one could categorise such structures.

5.3.2 Improving the pipeline

The developed pipeline showed outstanding performance on extracting comprehensive morphological features from EM data. Still, a number of potential improvements could further increase the pipeline resolution in capturing visual similarities and differences. First, the pipeline was based on the state-of-the-art methods in the field of self-supervised deep learning. However, the field is relatively new and extremely fast-developing - every month there are new improvements hugely advancing the training procedures. Incorporating the latest progress of the field would definitely enhance the MorphoFeatures extraction pipeline. For example, one could adapt pixel-level contrastive learning [164, 165] or extend the definition of positive samples beyond the augmented versions of the anchor samples [166].

Additionally, the pipeline could benefit from devising better strategies to incorporate coarse and fine texture. Right now the method only takes into account two scales of ultrastructure resolution. However, one could also adapt a truly multiscale approach and integrate multiple resolutions into one feature vector adapting, for example, a feature pyramid approach [167]. Beside this, one could also focus on improving the calculation of the morphological neighbourhood. For example, instead of simple averaging, one could assign each neighbour a weight based on the area of contact between two cells. Furthermore, the neighbourhood description might benefit from using novel methods for self-supervised learning on graphs [168].

Finally, it would also be of high importance to focus on improving the interpretability

of the extracted features. Even though, as I showed, many of the features learned by the neural network can be traced back to human-understandable morphological properties, it would be beneficial to be able to understand what each of the features means. For this, one could focus on encouraging the learned features to be more interpretable in the first place, for example, by enforcing them to be non-negative and sparse [169]. Alternatively, one could integrate better mechanisms for tracing features back to the raw data characteristics, for instance, Oclusion [170] or Grad-CAM [171] methods.

5.3.3 New datasets

Besides enabling a fast automated exploration of whole animal EM volumes, I expect the pipeline to be of great use when comparing multiple EM volumes of the same species. This could, for example, be necessary if one would want to compare the animal morphology under different conditions or environments, or describe cellular morphology changes across the course of the organism development. In this case the learned morphological features could be used to both establish correspondence between similar cells in different volumes and estimate the morphological difference between the corresponding cells.

Still, in order to apply the designed pipeline to multiple samples that were imaged independently, one would have to adapt it for being trained on multiple datasets. More specifically, the pipeline would have to be adjusted to ignore all non-biological differences between the samples. For example, it is widely known that neural networks are intrinsically intensity-sensitive, thus, one would have to make sure that the dataset intensities reflect the true differences in the sample ultrastructure, rather than the imaging conditions. A potential way to achieve this sample independence would be to use reverse gradients [172]. For this, one would train the network to additionally distinguish which sample a cell came from, but 'penalise' it every time the network guesses correctly.

Another interesting application of the feature extraction pipeline would be a newly acquired volume of the same species that is not yet segmented. In this case one should, nevertheless, be able to use the fine texture part of MorphoFeatures to find corresponding regions with similar morphology in the new volume. For example, if a researcher wants to find a cell of interest in the new volume, they could just describe the cell as a set of (non-averaged) fine texture feature vectors and look for similar vectors that are co-localised in the new volume. Furthermore, fine texture features could also be used as a biological prior to assist segmentation of new volumes, since the already segmented volume could allow us to define which textures can co-exist in a cell and which ones are mutually exclusive. For instance, one could (automatically) prohibit merging patches with muscle fibres and secretory vesicles, since they never co-occur in a cell. This could, for example, be implemented as negative potentials in the graph partitioning step of an EM segmentation pipeline [173].

5.3.4 Other types of data

One more important direction yet to be explored is adjusting the pipeline to other imaging data modalities. Electron microscopy currently offers the richest description of cellular morphology, however, it is still time-consuming and expensive. This is one of the reasons why up until now there has been only a modest amount of whole-organism EM datasets, and the absolute majority of data of such scale is taken by lower-resolution methods. Therefore, in order to ensure the broad applicability of the MorphoFeatures pipeline, it is vital to make sure that the pipeline produces reasonable results on the lower-resolution imaging data modalities as well.

One could start with adapting the proposed method for feature extraction to intermediate resolution imaging techniques, as, for example, X-ray holographic nano-tomography (XNH) [64, 174]. As I showed in Chapter 3, despite having lower resolution, this technique can produce volumes with cellular morphology descriptive enough to distinguish cell types. For this reason I expect adapting the texture part of the MorphoFeatures pipeline to this type of data to be a fairly straightforward process. Additionally, if the dataset is segmented, one could also directly apply the shape part of the pipeline. However, the data resolution provided by XNH is not high enough to trace smaller neurites, which are an important part of neuron morphology. That is why, in order to to fully differentiate neuron classe one would have to devise a way to supplement soma morphology description from XNH data with neuronal branching features that could be extracted, for instance, from EM data.

If the dataset of interest was acquired using light (fluorescence) microscopy and has been segmented, one could directly apply the shape part of the MorphoFeatures pipeline. However, the outcomes of applying the texture part would highly depend on which structures are labelled inside the cells. For example, a similar approach used on images of a multitude of fluorescently labelled proteins in single cells [62] has been shown to extract rich descriptions of protein subcellular localisation. Nevertheless, as the authors show, the pipeline performs much worse when applied without the strong pre-text task of protein classification. Since such descriptive classification labels are not always available even for LM datasets, removing the dependence on such a pre-text task would be an important step towards general usage of contrastive learning methods for training feature extractor networks on light microscopy data.

5.3.5 Data integration

As I described in the previous chapter, in order to get a truly complete description of an object of interest, e.g. a cell, one might have to combine multiple modalities of data into one multi-sided representation. For instance, MorphoFeatures representation already explicitly combines shape and texture. However, other types of data can additionally be added to further improve the completeness of such descriptions.

One of the applications where such an integration would be highly needed is describ-

ing neuronal morphology. While MorphoFeatures already show a good performance on subgrouping neurons based on their soma morphology, in order to accurately distinguish neuron types one needs to take into account their neurites as well. For example, concurrent to our work, Dorkenwald et al. [175] apply a method to the segmented cells in mouse and human cerebral cortex that is similar to the texture contrastive learning part of the MorphoFeatures pipeline. Instead of aggregating the fine texture features across the whole cell, like we do, the authors experiment with limiting the patch aggregation to a smaller radius, which allows them to classify cell parts, such as somas, axons, dendrites and astrocytic processes. Moreover, they show that even a small aggregation radius can still produce features descriptive of the main glia types (astrocytes, microglia, etc) and general neuron classes (pyramidal, thalamic, basket, etc). While they successfully extract texture features from axons and dendrites, what is still missing to precisely describe the neural types is their branching patterns. Thus, to supplement pure texture features, developing a similar automated unbiased method to extract neuron skeletons features would be of utter importance for unsupervised exploration of neuron types in brain EM volumes.

Another important direction, as discussed above, would be to combine morphological descriptions with multiomics data. In this Chapter I showed how making use of integrated gene expression data, even when limited to several hundreds of genes, could assist in describing new cell types and tissues that I discovered during the analysis of MorphoFeatures in the EM data of *Platynereis*. I expect that extending such a gene atlas to thousands of genes would allow us to automate the genetic characterisation of all the morphological diversity in the data. This could be achieved by acquiring single-cell transcriptomics data and mapping it to the EM volume via Virtual Cells, as proposed in Chapter 4. Such data could also be combined with epigenome profiling, which is currently possible to acquire in parallel with single-cell transcriptomics [99, 100, 101, 102]. In order to further enrich the cellular descriptions one could also acquire spatial proteomics and metabolomics data [89, 90] and align it with the EM volume in a similar manner to how the spatial gene expression maps were aligned. This would add an additional intermediate state between the gene expression and the final phenotype of cellular morphology, allowing to better investigate how the former is translated into the latter.

Finally, an equally important task would be to investigate ways to integrate such multimodal data. In this project I was mostly interested in finding and describing groups of cells using the morphological representations of the dataset. However, given the full gene expression data and possibly additional data modalities, one could start exploring the variability of the combined data. For example, one could use the method proposed by Argelaguet et al. [105] to find the major axes of heterogeneity across all the available data views. This method could locate not only specific groups of cells, but also continuous gradients of existing variations. Exploring all the heterogeneity in the data in an automatic unbiased and explainable way could enable us to get a full picture of all the cell types in an organism and understand how they develop and function.

Chapter 6

Concluding remarks

6.1 Thesis summary

In brief, with this thesis I aimed to design better automatic ways of describing cells from microscopy imaging data. I used various, mainly deep learning based computational methods to analyse cell differences in diverse imaging data modalities and to integrate information from complementary modalities. Specifically, my contributions can be summarised as follows:

- In chapter 2 I analysed a fluorescent light-sheet microscopy movie of a developing *Drosophila* embryo. Using a neural network based pipeline I showed that there are no obvious differences in myosin patterns between two distinct populations of mesoderm cells. I further showed that the shape difference of these populations arise from the combined influence of their own and their neighbourhood myosin.

- In chapter 3 I investigated a newly developed X-ray holographic nano-tomography imaging method. Using a neural network classification pipeline I showed that the method resolution is sufficient to automatically recognise cell types in the mouse brain cortex.

- In chapter 4 I focused on integrating image-based gene expression data with EM morphology on the whole-animal single-cell level. More specifically, I devised a way to automatically assign coherent denoised gene expression profiles to segmented cells in the EM volume of a three-segmented annelid worm *Platynereis dumerilii*.

- In chapter 5 I aimed at designing unbiased comprehensive descriptors of cellular morphology in EM data. I designed a fully unsupervised neural network pipeline that extracts extensive morphological representations, grouping cells compatible with their gene expression. I further showed how such representations could be used to discover and describe morphological cell types, tissues and organs.

The work I have done only begins to address the fundamental problem of automating generation of exhaustive image-based cell representations. In the following paragraphs I will outline the two broad directions that I believe are necessary to pursue in order to bring the field closer to this challenging goal.

6.2 General methods for extracting morphological information from imaging data

The ever increasing amount and complexity of biological imaging data require efficient automatic methods for unbiased analysis. This thesis, among multiple other works, illustrates how computational methods can facilitate describing cell differences. However, as I discussed in Introduction, current pipelines are mostly tailored to a specific dataset and/or specific research question. Therefore, an essential direction of future research is developing generally applicable methods for characterising cells in any imaging modalities.

The first step in this direction has been made by deep learning. There already exist standard approaches to common analysis problems in the fields of natural, medical and biological imaging that with a high probability deliver a successful solution. For example, for classifying mouse brain cells (Chapter 3) I used a ResNet network [79], which is a commonly used architecture for classifying images. I have experimented with multiple, more modern architectures. However, given the modest amount of training data, which is mostly the case in biological studies, the classical ResNet still showed the best performance. A similar picture can be seen for various image-to-image problems that include, for example, image segmentation, restoration and denoising. Despite many method advances, the base network architecture used for these tasks - a Unet architecture [60] - remains unchanged for years. If no training data is available, self-supervised approaches based on contrastive learning are highly likely to extract useful features for multiple tasks, including classification, segmentation or general exploratory data analysis, as I have shown in Chapter 5. Thus, for most common cell analysis tasks there often exists a 'classical' deep learning approach to tackle it.

However, as humans, we normally do not approach such tasks separately. For example, to segment individual cells we do not focus on separate boundaries, but rather on the whole view of a cell that includes both the cell membrane and its inner morphology. Furthermore, our view of a cell is sufficiently data-agnostic - we would recognise a nucleus in both fluorescent and electron microscopy as the same structure. A challenging but promising future step would, thus, be developing one model operating on any cell imaging modality and universal for all cell-related image analysis tasks. Similar to our perception, such a model should learn the very concept of a cell, therefore be able to properly describe morphology, classify and segment cells in any kind of imaging data.

A recent focus in the field of natural image and language processing are so-called foundation models [176, 177, 178] that are supposed to learn a holistic view of the data, useful for a broad range of applications and modalities. To account for this, the architec-

ture of such models, currently based on the transformer architecture [179], usually consists of three major modules:

- (1) modality-specific encoders that process different types of input, i.e. different data modalities, to generate modality-agnostic representations;
- (2) a general encoder receiving input from (1) and extracting fundamental data representation, useful for multiple tasks;
- (3) small task-specific decoders that use the representation from (2) to generate the desired output, for example segmentation masks or classification labels.

To promote such models to extract general-purpose, not task-specific representations, the training of the general encoder (2) is done in a self-supervised manner, which encourages the model to rather learn the relationship between parts of the data, than some limited set of labels. Afterwards, a minimal set of labels is normally used to train the task-specific decoders.

I believe that the next step to automate morphological description of cells would be adjusting such foundation models to biological data. At the moment the main limitation of applying big transformer-based foundation models to biological data is the immense amount of training data and computational resources required to train them. This, I believe, could, on one hand, be mitigated by introducing smaller light-weight models, as it is currently done in the medical imaging fields [180]. On the other hand, the ever increasing amount of imaging data and better availability of computational resources would eventually allow to train much bigger models on the biological imaging data as well.

The expected exponential increase of imaging data will also render it infeasible to generate manual labels, thus, another direction of primary importance is designing proper self-supervised training objectives that should be based on the profound understanding of the underlying data. Developing such objectives has been shown to considerably contribute to the success of self-supervised learning on natural images [144]. However, many of these criteria are based on peculiarities of this domain (e.g. complementary RGB channels, fixed orientation or object-centred images), and, therefore, are not directly transferable to biological data. Multiple self-supervised objectives have been proposed for various cell imaging datasets [146, 61, 62]. However, they are largely specific to a certain imaging modality and/or type of experiment. Therefore, designing generally applicable, unbiased training objectives for cell imaging data is the next essential step to be able to extract cell descriptors from massive amounts of imaging data of different modalities.

6.3 Generating exhaustive cell representations

Concurrent to the development of general methods for extracting holistic cell representations from any imaging modality, another essential direction is integration of the representations from other modalities. Ideally, for a full cell type description we would want to

bring together all the molecular and anatomical characteristics of a cell on the whole-organ or even whole-organism level. This would include genomics, epigenomics, transcriptomics, proteomics and metabolomics, as well as morphology and physiological properties, such as mechanical and electrophysiological properties, spatial distribution and connectivity of individual cells. Combining all these views of a cell is essential to get a full understanding of both how cell identity is established and how this identity leads to cellular function.

Several studies have already shown the possibility of successful automatic integration of multiple modalities on the single-cell level. For instance, the *Platynereis* atlas [65], discussed in Chapter 4, is, to the best of my knowledge, the first effort to automatically integrate gene expression with EM morphology on the whole-organism single-cell level. Another prominent example is a recent study generating a single-cell atlas of the mammalian primary motor cortex that combines transcriptomics, epigenomics, light microscopy and electrophysiological properties [181]. Not only do such studies show the importance of integrating multiple data sources to generate better cell type descriptions, they also lay the methodological foundation for creating a multi-view representation of a cell.

Generalising from such multi-modal atlases, I expect single-cell transcriptomics to be the cornerstone modality required for bringing together multiple data views. Various studies have shown the possibility of acquiring transcription data together with epigenetic [99, 100, 101, 102], electrophysiological [182, 183] or connectivity [184] profiles on the single-cell level. These modalities could be used to specify cell types present in the data and identify the genes defining these types. Such differentially expressed genes could be further targeted by spatial transcriptomics, adding the information about cell location and neighbourhood. To get a precise snapshot of cellular morphology, EM volumes could be acquired and aligned with spatial transcriptomics data using some morphological landmarks [185], for example, cell nuclei [65]. In case such alignment is hindered by a low level of stereotypy of the studied organ/animal, the volumes could even be co-acquired using the latest technology for spatial transcriptomics-correlated EM [186]. Afterwards, other necessary modalities could be integrated computationally. For example, spatial proteomics data could be matched with spatial transcriptomics by using both cell location and expression features. Finally, functional descriptions of cells could be obtained using genetic tools, such as transgenic lines.

The possibilities of analysing cell types with such united complete representations are virtually limitless. In my opinion, one of the most interesting riddles such representations could tackle is the genotype-phenotype link. For example, the access to the gene expression, protein content and morphology of every cell in an organism could be used to disentangle which phenotypic features are caused by the gene expression program and which are the result of environmental influences. One could further analyse cellular neighbourhoods to examine how cells influence each other. The next exciting step would be to add another dimension in order to investigate phenotype formation during the course of the organism development. To achieve this, one could further acquire single-cell lineage information [187, 188, 189] and integrate spatial transcriptomics and EM data at different developmental stages with fluorescent movies of a developing embryo [190, 185]. Observing

and correlating gene expression and morphology changes over time would result in a much better understanding of how genetically identical cells acquire their functional specialisation over time. Finally, an ambitious, but highly promising direction would be to compare such comprehensive cell type representations across different species, that has already been attempted for single-cell transcriptomics data [191, 192]. Integrated knowledge of how the genotype-phenotype relationship is established not only during the development of a single organism, but also in the course of evolution would be invaluable for understanding the emergence of cellular diversity.

Chapter 7

Methods

7.1 Characterising shape changes of mesodermal cells during Drosophila gastrulation

7.1.1 Neural network classification

The data and its generation is described in detail in [63]. For training I used two types of input data: cell segmentation masks generated by Sourabh Bhide (as negative control) and myosin channel of the acquired movies. In the first case each input cell was loaded as a binary segmentation mask, while for myosin training cell segmentation masks were used to extract myosin expression of each cell that was further normalised. Images were cropped/padded to 96*96 pixel image, and the following augmentations were applied: random flips, random rotations and random transposing.

I used the row numbers (3-8) assigned by Sourabh Bhide as classification labels, renormalised to range [0, 1]. To test the network setup, I first tested row number prediction on the cells of the same timeframe for timeframes 10, 15 and 20 (250s, 375s and 500s). To further check for differences in myosin patterns between cells from different rows with the same myosin concentration, for each cell I selected a timeframe where its myosin concentration was 5 (+-30%). The value 5 was taken as the highest myosin concentration still present in the sufficient number of cells (at least 10 in each row).

The data was input into a small convolutional neural network, consisting of 4 or 5 convolutional layers (for training on the cells from the same timeframe both showed similar performance, however for training on the cells with the same myosin concentration, the 5-layer networks performed better). The initial number of feature maps was 8, increasing two-fold with every convolutional layer. Afterwards, there was a global average pooling layer, one fully connected layer with 20 neurons and the last fully connected layer with one neuron and a sigmoid activation that output a value in range [0, 1].

For training I used L1 loss, Adam optimizer [193] with a learning rate of 0.0001,

weight decay of 0.00005 and mixed precision training with O1 optimisation level. I used a batch size of 5 for both training and validation. I evaluated the model every 100 iterations and the best result was saved with a patience of 0.4. The learning rate was reduced by 0.98 whenever the validation score showed no improvement with a patience of 0.4. The training was done for 20 epochs on a single laptop NVIDIA GeForce GTX 1050 Ti Mobile GPU.

During prediction, the output values (in range [0, 1]) were renormalised back into the range [3, 8] and rounded up to the closest integer to get a categorical row number. However, since in the end I wanted to estimate if the network is able to distinguish between constricting and expanding cells, rather than the row numbers, to calculate the final accuracy I calculated the percentage of constricting cells predicted as any row from 3-6 and expanding cells as either row 7 or 8, and then took an average of these two values.

7.1.2 Investigating influence of myosin concentration

To understand whether cell shape change is connected to myosin concentrations inside and outside the mesoderm cells I did the following. I first calculated myosin intensity inside a cell (Figure 2.4A, inner circle) as a sum intensity of all pixels in a cell, and then calculated myosin concentration by dividing the myosin intensity by the cell area. To calculate myosin concentration in the neighbourhood I did the same calculation for the area of 70 pixels around the cell boundary (Figure 2.4A, outer circle). Then for each cell I calculated its relative shape change by dividing the cell area of the following time frame by the cell area in the current one. For every cell I calculated a vector of each value over the 25 time points, and further smoothed it along the time axis to reduce the noise using a 1D Gaussian filter (sigma=3) [194].

7.1.3 Calculating myosin offset

To check the hypothesis that myosin in the neighbourhood of constricting/expanding cells is located closer/further from the cell membrane respectively, I calculated the myosin offset in the neighbourhood as follows: I multiplied the value of myosin intensity of each pixel in a radius of 70 pixels from the cell membrane (Figure 2.4A, outer circle) by the distance of this pixel to the membrane (Figure 2.4B). I summed up the resulting weighted intensity and divided it by the total myosin intensity (sum of the outer circle on Figure 2.4A).

The obtained value shows the average distance of surrounding myosin from a cell, with lower values indicating cells where surrounding myosin is located closer to the cell and vice versa.

To calculate the myosin offset inside a cell I calculated the centre of mass of the cell and the centre of mass of the myosin expression within this cell and calculated the euclidean distance between the two values.

7.1.4 Software and code

I trained the networks using Python packages pytorch [195] and personally modified repositories of inferno [196] and neurofire [197]. I stored and analysed the data using h5py [198], numpy [199], pandas [200], scipy [194] and scikit-image [201] and visualised it using matplotlib [202] and napari [203]. The code for the analysis can be found at: `https://github.com/vzinche/drosophila_emryo_cells`

7.2 Classifying cells imaged by a new intermediate resolution modality

7.2.1 Neural network classification

To train a classification neural network I used two XNH volumes of the mouse cortex of the resolution 130nm, acquired as described in [64]. To extract cells from the volume I I cropped a region of 160 pixel diameter around each cell soma annotation seed. For the training and validation I further reduced the crop size to 128 and normalised the data to be in the range of [0, 1]. I additionally applied the following augmentations for the training only: random flips, random noise (sigma=0.01), random elastic transforms (alpha=100, sigma=50). I only used the labels 'pyramidal', 'not pyramidal' and 'non neuronal'.

This input data is further fed into a 3D ResNet18 [79] with 3 output classes. I used soft cross-entropy as loss with 0.05 softening [80], Adam optimizer [193] with a learning rate of 0.0001 and weight decay of 0.00005, mixed precision training with O1 optimisation level and batch size of 20 for both training and validation. I evaluated the model every 100 iterations and the best result was saved with a patience of 0.4. The learning rate was reduced by 0.98 whenever the validation score showed no improvement with a patience of 0.4. The training was done for 20 epochs on a single NVIDIA GeForce RTX 2080 Ti GPU.

7.2.2 Software and code

To process the data, train the network and analyse the results I used Python packages pymaid [204], numpy [199], pandas [200], z5py [205], scipy [194], scikit-image [201], scikit-learn [206], pytorch [195] and personally modified repositories of inferno [196] and neurofire [197]. The code for data fetching, training and analysis can be found at `https://github.com/vzinche/brain_cells`.

7.3 More detailed cell characterisation through combining genetic information with high resolution imaging

7.3.1 Automating Virtual Cell generation

The detailed description of how the VCs were generated can be found in [122]. I improved the pipeline by automating the previously required steps of manual curation. More specifically, in the automated pipeline I firstly removed all the VCs composed of less than 16 supervoxels. According to Hernando Vergara, this was the threshold below which VCs showed low spatial correlation, indicating that these VCs might be an artefact of mixed gene expression on the cell borders. Afterwards I split all the VCs into spatially connected components (CCs) that potentially represent single cells or groups of neighbouring cells. I further removed CCs composed of less than 5 supervoxels, since these were much smaller than we expected cells to be. All CCs above the size of 8 supervoxels were considered big enough not to be considered artefacts, but the ones in the range between 5 and 8 supervoxels were kept only if there was a CC with the same genetic profile symmetrically located on the other side of the animal. Symmetric location was defined by mirroring the coordinates of the CC centre to the other side of the animal and defining a radius of 4 supervoxels around this location. Next, all the CC that were retained after this curation were assembled back into VCs and their average expression was used as the VC expression profile. The full pipeline that includes the VC generation and curation is available on GitHub at `https://github.com/mobie/prospr-vc-generation`.

7.3.2 Virtual cell assignment

In the first step of the assignment I calculated the overlap between each gene expression map and each segmented cell. This is further referred to as overlap assignment. Afterwards, for each segmented cell I located all the VCs in the radius of 5µm from the boundaries of the cell. This radius was chosen based on the estimated registration error and biological variability in cell locations. Next, for each segmented cell I compared the expression profile assigned by overlap to the profiles of all the VCs found within the radius, and assigned the VC that was the closest genetically. However, I expected that for some cells there might be no matching VC, thus, if not assigning any gene expression (assigning a profile with zero expression of all genes) was genetically closer than any VC in the proximity, I assigned zero expression instead. I used version 1.0.0 of the segmentation data and performed the assignment only for the cells that had nuclei assigned (column cells in the table `https://github.com/mobie/platybrowser-project/blob/main/data/1.0.0/tables/sbem-6dpf-1-whole-segmented-cells/default.tsv`). The resulting assignment can be accessed at `https://github.com/mobie/platybrowser-project/blob/main/data/1.0.0/tables/sbem-6dpf-1-whole-segmented-cells/vc_assignments.tsv`.

7.3.3 Assignment validation

To validate the assignment I manually identified 208 symmetric cell pairs in the animal, aiming to cover most of the animal regions. I defined the cells as symmetric pairs if they showed similar morphological appearance and were symmetrically located on two sides of the animal. To estimate the assignment accuracy for each pair I calculated the difference between the expression profiles assigned to the cells comprising the pair and averaged this difference across the 208 pairs.

7.3.4 Software and code

For the assignment I used Python packages numpy [199], elf [207], pybdv [208] and vigra [209]. For data visualisation I used matplotlib [202], mayavi [210] and MoBIE [128]. The code can be found at `https://github.com/mobie/platybrowser-project/blob/main/mmpb/extension/attributes/vc_assignments_impl.py`.

7.4 Unsupervised comprehensive description of high resolution cellular morphology

7.4.1 Data, training and feature extraction

The data used is described in detail in [66]. I only used the cells with assigned nuclei and filtered out the obvious merge errors. I also excluded cells at the top of the animals head and the lower part of pygidium, since the algorithm used for intensity correction of the EM volume performed unreliably on the boundaries of the volume. The excluded cells, negatively affected by intensity correction, were taken from `https://github.com/mobie/platybrowser-project/blob/main/data/1.0.0/tables/sbem-6dpf-1-whole-segmented-cells/extrapolated_intensity_correction.tsv`.

The network to extract features from the data consists of six encoders (Figure 5.3) which are extracting shape, coarse and fine-grained texture of cells and nuclei. The shape extraction network was designed by Johannes Hugger and is described in detail in [66].

I did the following to extract the texture features: "Due to the limitations of the GPU memory, for the coarse texture a central crop is taken from a downsampled version of the data (80*80*100nm). The centre of mass of the nucleus is determined, and a bounding box of size 144*144*144 pixels (cytoplasm) or 104*104*104 (nuclei) is taken around it. During prediction the bounding box is increased to 320*320*320 (cytoplasm) and 160*160*160 (nuclei) to fit more information. Fine texture patches have a higher resolution (20*20*25nm) and are taken separately from cytoplasm and nucleus. For this the bounding box surrounding cytoplasm or nucleus is split into cubes of 32*32*32 pixels, and only those cubes are considered, which are less than 50% empty. For both coarse and fine texture the data is normalised to the range of [0, 1] and for cytoplasm crops the nuclei

are masked out and vice versa.

The encoding of fine texture details and coarser contextual information is done using the encoder of a U-Net [60]. The fully convolutional network reduces spatial information gradually by employing 3 blocks that consist of 3D convolutional layers with ELU activations [211] separated by 3DMaxPooling layers. The number of feature maps starts with 64, increases twofold in each block, and afterwards gets reduced to 80 by an additional convolutional layer followed by an ELU activation. In the end a global average pooling operation is applied to reduce feature dimensionality from 3D to 2D. To ensure that empty regions of the data are not affecting the final feature vector, they are excluded for this final pooling" [66].

The training is done in a fully unsupervised way using a combination of contrastive and autoencoder losses as the training objective. As described in [66], "First, we randomly sample a mini-batch of cells or nuclei. Two augmented views are then created for each sample in the batch and concatenated to form an augmented batch of twice the size. To create positive samples for the contrastive training of texture we used random flips and rotations and elastic transformations of the data.The augmented batch is passed through the neural network encoder to compute representations. Those representations that share the same underlying sample are considered as positive pairs whereas the others are considered as negative pairs. The NT-Xent loss [144] is then used to score the representations based on their similarity, and, therefore, encourages the network to embed positive pairs closer to each other than negative pairs.

In order to learn representations of texture, we combine the NT-Xent loss with an Autoencoder reconstruction objective. The latent representations, taken before the dimensionality reduction step of global average pooling, are further processed by a symmetric decoder part of the UNet network with only 1 block, which aims to reconstruct the network inputs. The reconstructions are then compared to the original inputs and scored by a mean squared error loss. Additionally, an L2-Norm loss was applied to flattened bottleneck features to restrict the range of possible values. The final loss is thus a weighted sum of the NT-Xent, the mean squared error loss and the L2-Norm loss. The batch size was set to 12 and 16 for cell and nuclei coarse texture and to 32 for fine texture. For training we used Adam [193] as optimizer with a learning rate of 0.0001 and a weight decay of 0.00005. The model was evaluated every 100 iterations and the best result was saved with a patience of 0.95. The learning rate was reduced by 0.98 whenever the validation score did not show improvement with a patience of 0.95.

In order to compute the MorphoFeatures representation for the whole animal, we processed each cells cellular and nuclear shape and texture, as described in previous sections, with our trained networks. For the fine texture patches to get one feature vector per cell the feature vectors of all the patches from a cell were averaged. Each cell is represented by a 480-dimensional vector that consists of a nuclear and cellular part. Each of these parts consists of a shape, coarse and fine-grained texture 80-dimensional feature vector".

To get a representation of cellular neighbourhood morphology - MorphoContextFea-

tures - Johannes Hugger combined MorphoFeatures of each cell and its immediate neighbours as described in [66].

7.4.2 Evaluating MorphoFeatures

To quantify how good MorphoFeatures can predict cell classes, I used broad cell categories, annotated as described in [65] and Chapter 5. I then extracted MorphoFeatures for every cell in the animal with a nucleus assigned (except cells with substantial merge segmentation errors) and standardised these representations across the whole dataset. Afterwards I trained a logistic regression model with stratified 5-fold cross-validation [206] (C=1, solver=lbfgs). Since the initial number of features was higher than the number of training samples, to prevent overfitting I reduced the amount of correlated features using feature agglomeration [206] (n_clusters=100).

To further quantify how well MorphoFeatures can detect symmetric neighbour cells I used the bilateral pair evaluation, introduced by Kimberly Meechan in [65].

To compare MorphoFeatures to explicitly defined features I used the set of morphological features from [65] manually designed by Kimberly Meechan. The cellular features can be accessed at `https://github.com/mobie/platybrowser-project/blob/main/data/1.0.1/tables/sbem-6dpf-1-whole-segmented-cells/morphology.tsv` and the nuclear and chromatin features at `https://github.com/mobie/platybrowser-project/blob/main/data/1.0.1/tables/sbem-6dpf-1-whole-segmented-nuclei/morphology.tsv`. The UMAP of these features can be found at `https://github.com/mobie/platybrowser-project/blob/main/data/1.0.1/tables/sbem-6dpf-1-whole-segmented-cells/morphology_umap.tsv`.

7.4.3 Umap and clustering

As I described in [66], "Visualisation was done using the UMAP package [153] with the following parameters: euclidean metric, 15 neighbours and 0 minimal distance. Cell metadata, such as annotated types, animal regions or gene expression were plotted on the resulting UMAP.

Cell clustering was done according to the following procedure. First the features were standardised to zero mean and unit variance. Then, following [212] a weighted k-neighbour graph was constructed from MorphoFeatures representations of all the animal cells using the UMAP algorithm [153] with the following arguments: n_neighbors=20, metric=euclidean, min_dist=0. Afterwards the Leiden algorithm [154] for community detection was used to partition the resulting graph, using the CPMVertexPartition method and a resolution parameter of 0.004. This resulted in 16 clusters, one of which was discarded, since it contained only cells with split segmentation errors. For more fine-grained clustering the resolution was increased to 0.1 to split apart secretory and ciliary band cells, 0.2 to split photoreceptors and enteric neurons, 0.07 to split the midgut cluster, 0.005 for muscle neighbourhood cluster (n_neighbors for the UMAP clustering was adjusted to 10)

and 0.01 for splitting foregut neurons and epithelial cells. To visualise average shape and texture of a cluster, a median feature vector was computed across the cluster and the cell with the feature vector closest to the median one was taken as the average cell".

7.4.4 Gene expression analysis

As I describe in [66], "to produce the cluster gene expression 'dot plots' the gene expression values for each cell were taken from [65] at `https://github.com/mobie/platybrowser-project/blob/main/data/1.0.1/tables/sbem-6dpf-1-whole-segmented-cells/genes.tsv`. Similar to [65], for each cluster and for each gene we calculated three values: (A) the mean expression in the cluster, (B) the fraction of the total animal/region expression within this cluster (the sum of expression in this cluster divided by the sum of the expression in the whole animal or the selected region) and (C) the specificity defined as C = 2AB(A + B). Intuitively, A indicates how much of the cells in a given cluster express a given gene, B shows how selective this gene for this cluster is and C is a harmonic mean of the two values. To create a dot plot only the genes with C value above 0.15 were used. For the neuron clusters 1, 3 and 5 the threshold was increased to 0.2 due to a high number of specifically expressed genes available. For the midgut clusters to show only the genes differing between the two clusters, the genes were removed, where the specificity was above 0.15 in both clusters, but differed by less than 0.4. The size of dots in the plots reflects A and the colour corresponds to B. The cluster with the most gene expression was determined and the remaining clusters were sorted by their similarity to it. To illustrate some of the genes showing differential expression in these groups, the corresponding regions of the UMAP representation were cut out and these genes were plotted on top".

7.4.5 Ganglionic nuclei analysis

To compare ganglionic nuclei I used manual segmentation from `https://github.com/mobie/platybrowser-project/blob/main/data/1.0.1/tables/sbem-6dpf-1-whole-segmented-cells/ganglia_ids.tsv` and `https://github.com/mobie/platybrowser-project/blob/main/data/1.0.1/tables/sbem-6dpf-1-whole-segmented-cells/regions.tsv` and gene clusters from `https://github.com/mobie/platybrowser-project/blob/main/data/1.0.1/tables/sbem-6dpf-1-whole-segmented-cells/gene_clusters.tsv` [65].

7.4.6 Software and code

I trained the networks using Python packages pytorch [195] and personally modified repositories of inferno [196] and neurofire [197]. I stored and analysed the data using z5py [205], h5py [198], pybdv [208], numpy [199], pandas [200], scipy [194], scikit-learn [206], scikit-image [201], networkx [213], igraph [214], leidenalg [154] and umap [153]. I visualised the data using matplotlib [202], seaborn [215] and MoBIE [128]. The code for the training and analysis can be found at `https://github.com/kreshuklab/MorphoFeatures`.

Bibliography

[1] Merriam-Webster, "Life." `https://www.merriam-webster.com/dictionary/life`, 2022.

[2] Britannica, "Life on earth." `https://www.britannica.com/science/life/Life-on-Earth`, 2022.

[3] M. Herron, P. Conlin, and W. Ratcliff, *The Evolution of Multicellularity*. Evolutionary Cell Biology, CRC Press, 2022.

[4] T. Brunet and N. King, "The origin of animal multicellularity and cell differentiation," *Developmental cell*, vol. 43, no. 2, pp. 124–140, 2017.

[5] A. Sebé-Pedrós, B. M. Degnan, and I. Ruiz-Trillo, "The origin of metazoa: a unicellular perspective," *Nature Reviews Genetics*, vol. 18, no. 8, pp. 498–512, 2017.

[6] T. Brunet, M. Albert, W. Roman, M. C. Coyle, D. C. Spitzer, and N. King, "A flagellate-to-amoeboid switch in the closest living relatives of animals," *Elife*, vol. 10, p. e61037, 2021.

[7] J. M. Musser, K. J. Schippers, M. Nickel, G. Mizzon, A. B. Kohn, C. Pape, P. Ronchi, N. Papadopoulos, A. J. Tarashansky, J. U. Hammel, *et al.*, "Profiling cellular diversity in sponges informs animal cell type and nervous system evolution," *Science*, vol. 374, no. 6568, pp. 717–723, 2021.

[8] T. Brunet, A. H. Fischer, P. R. Steinmetz, A. Lauri, P. Bertucci, and D. Arendt, "The evolutionary origin of bilaterian smooth and striated myocytes," *Elife*, vol. 5, p. e19607, 2016.

[9] K. Thorn, "A quick guide to light microscopy in cell biology," *Molecular biology of the cell*, vol. 27, no. 2, pp. 219–222, 2016.

[10] S. L. French, P. Vijey, K. W. Karhohs, A. R. Wilkie, L. J. Horin, A. Ray, B. Posorske, A. E. Carpenter, K. R. Machlus, and J. E. Italiano Jr, "High-content, label-free analysis of proplatelet production from megakaryocytes," *Journal of Thrombosis and Haemostasis*, vol. 18, no. 10, pp. 2701–2711, 2020.

[11] M. Buckup, J. M. Kaneda, A. M. Birk, E. Glockner, R. Venook, A. Jain, S. Sharma, C. Wong, and K. Sutha, "Utilising low-cost, easy-to-use microscopy techniques for early peritonitis infection screening in peritoneal dialysis patients," *Scientific reports*, vol. 12, no. 1, pp. 1–15, 2022.

[12] J. Lad, S. Serra, F. Quereshy, M. Khorasani, and A. Vitkin, "Polarimetric biomarkers of peri-tumoral stroma can correlate with 5-year survival in patients with left-sided colorectal cancer," *Scientific Reports*, vol. 12, no. 1, pp. 1–14, 2022.

[13] P. J. Thul, L. Åkesson, M. Wiking, D. Mahdessian, A. Geladaki, H. Ait Blal, T. Alm, A. Asplund, L. Björk, L. M. Breckels, *et al.*, "A subcellular map of the human proteome," *Science*, vol. 356, no. 6340, p. eaal3321, 2017.

[14] Y. Cai, M. J. Hossain, J.-K. Hériché, A. Z. Politi, N. Walther, B. Koch, M. Wachsmuth, B. Nijmeijer, M. Kueblbeck, M. Martinic-Kavur, *et al.*, "Experimental and computational framework for a dynamic protein atlas of human cell division," *Nature*, vol. 561, no. 7723, pp. 411–415, 2018.

[15] N. H. Cho, K. C. Cheveralls, A.-D. Brunner, K. Kim, A. C. Michaelis, P. Raghavan, H. Kobayashi, L. Savy, J. Y. Li, H. Canaj, *et al.*, "Opencell: Endogenous tagging for the cartography of human cellular organization," *Science*, vol. 375, no. 6585, p. eabi6983, 2022.

[16] P. J. Keller, A. D. Schmidt, J. Wittbrodt, and E. H. Stelzer, "Reconstruction of zebrafish early embryonic development by scanned light sheet microscopy," *science*, vol. 322, no. 5904, pp. 1065–1069, 2008.

[17] A. Maizel, D. von Wangenheim, F. Federici, J. Haseloff, and E. H. Stelzer, "High-resolution live imaging of plant growth in near physiological bright conditions using light sheet fluorescence microscopy," *The Plant Journal*, vol. 68, no. 2, pp. 377–385, 2011.

[18] K. Khairy, W. C. Lemon, F. Amat, and P. J. Keller, "Light sheet-based imaging and analysis of early embryogenesis in the fruit fly," in *Tissue Morphogenesis*, pp. 79–97, Springer, 2015.

[19] K. McDole, L. Guignard, F. Amat, A. Berger, G. Malandain, L. A. Royer, S. C. Turaga, K. Branson, and P. J. Keller, "In toto imaging and reconstruction of post-implantation mouse development at the single-cell level," *Cell*, vol. 175, no. 3, pp. 859–876, 2018.

[20] S. Valuchova, P. Mikulkova, J. Pecinkova, J. Klimova, M. Krumnikl, P. Bainar, S. Heckmann, P. Tomancak, and K. Riha, "Imaging plant germline differentiation within arabidopsis flowers by light sheet microscopy," *Elife*, vol. 9, p. e52546, 2020.

[21] S. W. Hell and J. Wichmann, "Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy," *Optics letters*, vol. 19, no. 11, pp. 780–782, 1994.

[22] J. M. Guerra, "Super-resolution through illumination by diffraction-born evanescent waves," *Applied physics letters*, vol. 66, no. 26, pp. 3555–3557, 1995.

[23] M. G. Gustafsson, "Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy," *Journal of microscopy*, vol. 198, no. 2, pp. 82–87, 2000.

[24] E. Betzig, G. H. Patterson, R. Sougrat, O. W. Lindwasser, S. Olenych, J. S. Bonifacino, M. W. Davidson, J. Lippincott-Schwartz, and H. F. Hess, "Imaging intracellular fluorescent proteins at nanometer resolution," *science*, vol. 313, no. 5793, pp. 1642–1645, 2006.

[25] M. J. Rust, M. Bates, and X. Zhuang, "Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (storm)," *Nature methods*, vol. 3, no. 10, pp. 793–796, 2006.

[26] J. Otterstrom, A. Castells-Garcia, C. Vicario, P. A. Gomez-Garcia, M. P. Cosma, and M. Lakadamyali, "Super-resolution microscopy reveals how histone tail acetylation affects dna compaction within nucleosomes in vivo," *Nucleic acids research*, vol. 47, no. 16, pp. 8470–8484, 2019.

[27] T. T. Yang, W. M. Chong, W.-J. Wang, G. Mazo, B. Tanos, Z. Chen, T. M. N. Tran, Y.-D. Chen, R. R. Weng, C.-E. Huang, *et al.*, "Super-resolution architecture of mammalian centriole distal appendages reveals distinct blade and matrix functional components," *Nature communications*, vol. 9, no. 1, pp. 1–11, 2018.

[28] V. J. Sabinina, M. J. Hossain, J.-K. Hériché, P. Hoess, B. Nijmeijer, S. Mosalaganti, M. Kueblbeck, A. Callegari, A. Szymborska, M. Beck, *et al.*, "Three-dimensional superresolution fluorescence microscopy maps the variable molecular architecture of the nuclear pore complex," *Molecular Biology of the Cell*, vol. 32, no. 17, pp. 1523–1533, 2021.

[29] D. C. Pease, *Histological techniques for electron microscopy*. Academic Press, 1960.

[30] G. A. Meek, *Practical electron microscopy for biologists*. Wiley-Interscience, 1970.

[31] F. I. Seymour and M. Benmosche, "Magnification of spermatozoa by means of the electron microscope: Preliminary report," *Journal of the American Medical Association*, vol. 116, no. 22, pp. 2489–2490, 1941.

[32] H. S. Bennett and K. R. Porter, "An electron microscope study of sectioned breast muscle of the domestic fowl," *American Journal of Anatomy*, vol. 93, no. 1, pp. 61–105, 1953.

[33] C. Golgi, "On the structure of nerve cells," *Journal of microscopy*, vol. 155, no. 1, pp. 3–7, 1989.

[34] C. J. Peddie, C. Genoud, A. Kreshuk, K. Meechan, K. D. Micheva, K. Narayan, C. Pape, R. G. Parton, N. L. Schieber, Y. Schwab, *et al.*, "Volume electron microscopy," *Nature Reviews Methods Primers*, vol. 2, no. 1, pp. 1–23, 2022.

[35] Y. Wu, C. Whiteus, C. S. Xu, K. J. Hayworth, R. J. Weinberg, H. F. Hess, and P. De Camilli, "Contacts between the endoplasmic reticulum and other membranes in neurons," *Proceedings of the National Academy of Sciences*, vol. 114, no. 24, pp. E4859–E4867, 2017.

[36] C. S. Xu, S. Pang, G. Shtengel, A. Müller, A. T. Ritter, H. K. Hoffman, S.-y. Takemura, Z. Lu, H. A. Pasolli, N. Iyer, *et al.*, "An open-access volume electron microscopy atlas of whole cells and tissues," *Nature*, vol. 599, no. 7883, pp. 147–151, 2021.

[37] A. Müller, D. Schmidt, C. S. Xu, S. Pang, J. V. DCosta, S. Kretschmar, C. Münster, T. Kurth, F. Jug, M. Weigert, *et al.*, "3d fib-sem reconstruction of microtubule–organelle interaction in whole primary mouse β cells," *Journal of Cell Biology*, vol. 220, no. 2, 2021.

[38] C. Trapnell, "Defining cell types and states with single-cell genomics," *Genome research*, vol. 25, no. 10, pp. 1491–1498, 2015.

[39] A. Zeisel, A. B. Muñoz-Manchado, S. Codeluppi, P. Lönnerberg, G. La Manno, A. Juréus, S. Marques, H. Munguba, L. He, C. Betsholtz, *et al.*, "Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq," *Science*, vol. 347, no. 6226, pp. 1138–1142, 2015.

[40] D. Usoskin, A. Furlan, S. Islam, H. Abdo, P. Lönnerberg, D. Lou, J. Hjerling-Leffler, J. Haeggström, O. Kharchenko, P. V. Kharchenko, *et al.*, "Unbiased classification of sensory neuron types by large-scale single-cell rna sequencing," *Nature neuroscience*, vol. 18, no. 1, pp. 145–153, 2015.

[41] A. Sebé-Pedrós, E. Chomsky, K. Pang, D. Lara-Astiaso, F. Gaiti, Z. Mukamel, I. Amit, A. Hejnol, B. M. Degnan, and A. Tanay, "Early metazoan cell type diversity and the evolution of multicellular gene regulation," *Nature ecology & evolution*, vol. 2, no. 7, pp. 1176–1188, 2018.

[42] C. Cao, L. A. Lemaire, W. Wang, P. H. Yoon, Y. A. Choi, L. R. Parsons, J. C. Matese, M. Levine, and K. Chen, "Comprehensive single-cell transcriptome lineages of a proto-vertebrate," *Nature*, vol. 571, no. 7765, pp. 349–354, 2019.

[43] D. Arendt, J. M. Musser, C. V. Baker, A. Bergman, C. Cepko, D. H. Erwin, M. Pavlicev, G. Schlosser, S. Widder, M. D. Laubichler, *et al.*, "The origin and evolution of cell types," *Nature Reviews Genetics*, vol. 17, no. 12, pp. 744–757, 2016.

[44] H. Clevers, S. Rafelski, M. Elowitz, A. Klein, J. Shendure, C. Trapnell, E. Lein, E. Lundberg, M. Uhlen, A. Martinez-Arias, *et al.*, "What is your conceptual definition of cell type in the context of a mature organism?," *Cell Systems*, vol. 4, no. 3, pp. 255–259, 2017.

[45] M. Sachkova and P. Burkhardt, "Exciting times to study the identity and evolution of cell types," 2019.

[46] B. Xia and I. Yanai, "A periodic table of cell types," *Development*, vol. 146, no. 12, p. dev169854, 2019.

[47] S. A. Morris, "The evolving concept of cell identity in the single cell era," *Development*, vol. 146, no. 12, p. dev169748, 2019.

[48] H. Zeng, "What is a cell type and how to define it?," *Cell*, vol. 185, no. 15, pp. 2739–2755, 2022.

[49] U. Mayr, D. Serra, and P. Liberali, "Exploring single cells in space and time during tissue development, homeostasis and regeneration," *Development*, vol. 146, no. 12, p. dev176727, 2019.

[50] A. Rao, D. Barkley, G. S. França, and I. Yanai, "Exploring tissue architecture using spatial transcriptomics," *Nature*, vol. 596, no. 7871, pp. 211–220, 2021.

[51] C. G. Williams, H. J. Lee, T. Asatsuma, R. Vento-Tormo, and A. Haque, "An introduction to spatial transcriptomics for biomedical research," *Genome Medicine*, vol. 14, no. 1, pp. 1–18, 2022.

[52] C.-H. L. Eng, M. Lawson, Q. Zhu, R. Dries, N. Koulena, Y. Takei, J. Yun, C. Cronin, C. Karp, G.-C. Yuan, *et al.*, "Transcriptome-scale super-resolved imaging in tissues by rna seqfish+," *Nature*, vol. 568, no. 7751, pp. 235–239, 2019.

[53] S. G. Rodriques, R. R. Stickels, A. Goeva, C. A. Martin, E. Murray, C. R. Vanderburg, J. Welch, L. M. Chen, F. Chen, and E. Z. Macosko, "Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution," *Science*, vol. 363, no. 6434, pp. 1463–1467, 2019.

[54] J. G. Gall and M. L. Pardue, "Formation and detection of rna-dna hybrid molecules in cytological preparations," *Proceedings of the National Academy of Sciences*, vol. 63, no. 2, pp. 378–383, 1969.

[55] L. K. Scheffer, C. S. Xu, M. Januszewski, Z. Lu, S.-y. Takemura, K. J. Hayworth, G. B. Huang, K. Shinomiya, J. Maitlin-Shepard, S. Berg, *et al.*, "A connectome and analysis of the adult drosophila central brain," *Elife*, vol. 9, 2020.

[56] P.-H. Wu, D. M. Gilkes, J. M. Phillip, A. Narkar, T. W.-T. Cheng, J. Marchand, M.-H. Lee, R. Li, and D. Wirtz, "Single-cell morphology encodes metastatic potential," *Science Advances*, vol. 6, no. 4, p. eaaw6938, 2020.

[57] C. Uwizeye, J. Decelle, P.-H. Jouneau, S. Flori, B. Gallet, J.-B. Keck, D. D. Bo, C. Moriscot, C. Seydoux, F. Chevalier, *et al.*, "Morphological bases of phytoplankton energy management and physiological responses unveiled by 3d subcellular imaging," *Nature communications*, vol. 12, no. 1, pp. 1–12, 2021.

[58] T. E. Abdel-Maguid and D. Bowsher, "Classification of neurons by dendritic branching pattern. a categorisation based on golgi impregnation of spinal and cranial somatic and visceral afferent and efferent cells in the adult human.," *Journal of anatomy*, vol. 138, no. Pt 4, p. 689, 1984.

[59] M. Rowe and J. Stone, "Naming of neurones. classification and naming of cat retinal ganglion cells.," *Brain, behavior and evolution*, vol. 14, no. 3, pp. 185–216, 1977.

[60] T. Falk, D. Mai, R. Bensch, Ö. Çiçek, A. Abdulkadir, Y. Marrakchi, A. Böhm, J. Deubner, Z. Jäckel, K. Seiwald, *et al.*, "U-net: deep learning for cell counting, detection, and morphometry," *Nature methods*, vol. 16, no. 1, pp. 67–70, 2019.

[61] K. Yao, N. D. Rochman, and S. X. Sun, "Cell type classification and unsupervised morphological phenotyping from low-resolution images using deep learning," *Scientific reports*, vol. 9, no. 1, pp. 1–13, 2019.

[62] H. Kobayashi, K. C. Cheveralls, M. D. Leonetti, and L. A. Royer, "Self-supervised deep learning encodes high-resolution features of protein subcellular localization," *Nature methods*, vol. 19, no. 8, pp. 995–1003, 2022.

[63] S. Bhide, D. Gombalova, G. Mönke, J. Stegmaier, V. Zinchenko, A. Kreshuk, J. M. Belmonte, and M. Leptin, "Mechanical competition alters the cellular interpretation of an endogenous genetic program," *Journal of Cell Biology*, vol. 220, no. 11, p. e202104107, 2021.

[64] A. T. Kuan, J. S. Phelps, L. A. Thomas, T. M. Nguyen, J. Han, C.-L. Chen, A. W. Azevedo, J. C. Tuthill, J. Funke, P. Cloetens, *et al.*, "Dense neuronal reconstruction through x-ray holographic nano-tomography," *Nature neuroscience*, vol. 23, no. 12, pp. 1637–1643, 2020.

[65] H. M. Vergara, C. Pape, K. I. Meechan, V. Zinchenko, C. Genoud, A. A. Wanner, K. N. Mutemi, B. Titze, R. M. Templin, P. Y. Bertucci, *et al.*, "Whole-body integration of gene expression and single-cell morphology," *Cell*, vol. 184, no. 18, pp. 4819–4837, 2021.

[66] V. Zinchenko, J. Hugger, V. Uhlmann, D. Arendt, and A. Kreshuk, "Morphofeatures: unsupervised exploration of cell types, tissues and organs in volume electron microscopy," *bioRxiv*, 2022.

[67] S. Grainger and K. Willert, "Mechanisms of wnt signaling and control," *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, vol. 10, no. 5, p. e1422, 2018.

[68] M. Leptin and B. Grunewald, "Cell shape changes during gastrulation in drosophila," *Development*, vol. 110, no. 1, pp. 73–84, 1990.

[69] D. Sweeton, S. Parks, M. Costa, and E. Wieschaus, "Gastrulation in drosophila: the formation of the ventral furrow and posterior midgut invaginations," *Development*, vol. 112, no. 3, pp. 775–789, 1991.

[70] R. E. Dawes-Hoang, K. M. Parmar, A. E. Christiansen, C. B. Phelps, A. H. Brand, and E. F. Wieschaus, "Folded gastrulation, cell shape change and the control of myosin localization," *Development*, 2005.

[71] A. C. Martin, M. Kaschube, and E. F. Wieschaus, "Pulsed contractions of an actin–myosin network drive apical constriction," *Nature*, vol. 457, no. 7228, pp. 495–499, 2009.

[72] B. Lim, M. Levine, and Y. Yamazaki, "Transcriptional pre-patterning of drosophila gastrulation," *Current Biology*, vol. 27, no. 2, pp. 286–290, 2017.

[73] A. C. Martin, M. Gelbart, R. Fernandez-Gonzalez, M. Kaschube, and E. F. Wieschaus, "Integration of contractile forces during tissue invagination," *Journal of Cell Biology*, vol. 188, no. 5, pp. 735–749, 2010.

[74] P. J. Withers, C. Bouman, S. Carmignato, V. Cnudde, D. Grimaldi, C. K. Hagen, E. Maire, M. Manley, A. Du Plessis, and S. R. Stock, "X-ray computed tomography," *Nature Reviews Methods Primers*, vol. 1, no. 1, pp. 1–21, 2021.

[75] S. D. Rawson, J. Maksimcuka, P. J. Withers, and S. H. Cartmell, "X-ray computed tomography in life sciences," *BMC biology*, vol. 18, no. 1, pp. 1–15, 2020.

[76] M. A. Karreman, L. Mercier, N. L. Schieber, G. Solecki, G. Allio, F. Winkler, B. Ruthensteiner, J. G. Goetz, and Y. Schwab, "Fast and precise targeting of single tumor cells in vivo by multimodal correlative microscopy," *Journal of Cell Science*, vol. 129, no. 2, pp. 444–456, 2016.

[77] B. K. Hulse, H. Haberkern, R. Franconville, D. Turner-Evans, S.-y. Takemura, T. Wolff, M. Noorman, M. Dreher, C. Dan, R. Parekh, A. M. Hermundstad, G. M. Rubin, and V. Jayaraman, "A connectome of the *Drosophila* central complex reveals network motifs suitable for flexible navigation and context-dependent action selection," *eLife*, vol. 10, p. e66039, oct 2021.

[78] N. L. Turner, T. Macrina, J. A. Bae, R. Yang, A. M. Wilson, C. Schneider-Mizell, K. Lee, R. Lu, J. Wu, A. L. Bodor, *et al.*, "Reconstruction of neocortex: Organelles, compartments, cells, circuits, and activity," *Cell*, vol. 185, no. 6, pp. 1082–1100, 2022.

[79] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[80] C. Thiel, "Classification on soft labels is robust against label noise," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pp. 65–73, Springer, 2008.

[81] S. Berg, D. Kutra, T. Kroeger, C. N. Straehle, B. X. Kausler, C. Haubold, M. Schiegg, J. Ales, T. Beier, M. Rudy, *et al.*, "Ilastik: interactive machine learning for (bio) image analysis," *Nature methods*, vol. 16, no. 12, pp. 1226–1232, 2019.

[82] S. He, L.-H. Wang, Y. Liu, Y.-Q. Li, H.-T. Chen, J.-H. Xu, W. Peng, G.-W. Lin, P.-P. Wei, B. Li, *et al.*, "Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs," *Genome biology*, vol. 21, no. 1, pp. 1–34, 2020.

[83] X. Han, Z. Zhou, L. Fei, H. Sun, R. Wang, Y. Chen, H. Chen, J. Wang, H. Tang, W. Ge, *et al.*, "Construction of a human cell landscape at single-cell level," *Nature*, vol. 581, no. 7808, pp. 303–309, 2020.

[84] T. S. Consortium*, R. C. Jones, J. Karkanias, M. A. Krasnow, A. O. Pisco, S. R. Quake, J. Salzman, N. Yosef, B. Bulthaup, P. Brown, *et al.*, "The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans," *Science*, vol. 376, no. 6594, p. eabl4896, 2022.

[85] D. Osumi-Sutherland, C. Xu, M. Keays, A. P. Levine, P. V. Kharchenko, A. Regev, E. Lein, and S. A. Teichmann, "Cell type ontologies of the human cell atlas," *Nature Cell Biology*, vol. 23, no. 11, pp. 1129–1135, 2021.

[86] A. Dia, C. Jett, S. G. Trevino, C. S. Chu, K. Sriprawat, T. J. Anderson, F. Nosten, and I. H. Cheeseman, "Single-genome sequencing reveals within-host evolution of human malaria parasites," *Cell Host & Microbe*, vol. 29, no. 10, pp. 1496–1506, 2021.

[87] C. S. Xu, K. J. Hayworth, Z. Lu, P. Grob, A. M. Hassan, J. G. García-Cerdán, K. K. Niyogi, E. Nogales, R. J. Weinberg, and H. F. Hess, "Enhanced fib-sem systems for large-volume 3d imaging," *Elife*, vol. 6, 2017.

[88] Z. Zheng, J. S. Lauritzen, E. Perlman, C. G. Robinson, M. Nichols, D. Milkie, O. Torrens, J. Price, C. B. Fisher, N. Sharifi, *et al.*, "A complete electron microscopy volume of the brain of adult drosophila melanogaster," *Cell*, vol. 174, no. 3, pp. 730–743, 2018.

[89] M. J. Taylor, J. K. Lukowski, and C. R. Anderton, "Spatially resolved mass spectrometry at the single cell: recent innovations in proteomics and metabolomics," *Journal of the American Society for Mass Spectrometry*, vol. 32, no. 4, pp. 872–894, 2021.

[90] A. Mund, A.-D. Brunner, and M. Mann, "Unbiased spatial proteomics with single-cell resolution in tissues," *Molecular Cell*, vol. 82, no. 12, pp. 2335–2349, 2022.

[91] P. Bourceau, D. Michellod, B. Geier, and M. Liebeke, "Spatial metabolomics shows contrasting phosphonolipid distributions in tissues of marine bivalves," *PeerJ Analytical Chemistry*, vol. 4, p. e21, 2022.

[92] L. Capolupo, I. Khven, A. R. Lederer, L. Mazzeo, G. Glousker, S. Ho, F. Russo, J. P. Montoya, D. R. Bhandari, A. P. Bowman, *et al.*, "Sphingolipids control dermal fibroblast heterogeneity," *Science*, vol. 376, no. 6590, p. eabh1623, 2022.

[93] S. G. Snowden, H. J. Fernandes, J. Kent, S. Foskolou, P. Tate, S. F. Field, E. Metzakopian, and A. Koulman, "Development and application of high-throughput single cell lipid profiling: a study of snca-a53t human dopamine neurons," *Iscience*, vol. 23, no. 11, p. 101703, 2020.

[94] Z. Li, S. Cheng, Q. Lin, W. Cao, J. Yang, M. Zhang, A. Shen, W. Zhang, Y. Xia, X. Ma, *et al.*, "Single-cell lipidomics with high structural specificity by mass spectrometry," *Nature communications*, vol. 12, no. 1, pp. 1–10, 2021.

[95] A. Gillich, F. Zhang, C. G. Farmer, K. J. Travaglini, S. Y. Tan, M. Gu, B. Zhou, J. A. Feinstein, M. A. Krasnow, and R. J. Metzger, "Capillary cell-type specialization in the alveolus," *Nature*, vol. 586, no. 7831, pp. 785–789, 2020.

[96] M. Litviňuková, C. Talavera-López, H. Maatz, D. Reichart, C. L. Worth, E. L. Lindberg, M. Kanda, K. Polanski, M. Heinig, M. Lee, *et al.*, "Cells of the adult human heart," *Nature*, vol. 588, no. 7838, pp. 466–472, 2020.

[97] R. Elmentaite, N. Kumasaka, K. Roberts, A. Fleming, E. Dann, H. W. King, V. Kleshchevnikov, M. Dabrowska, S. Pritchard, L. Bolt, *et al.*, "Cells of the human intestinal tract mapped across space and time," *Nature*, vol. 597, no. 7875, pp. 250–255, 2021.

[98] L. Heinrich, D. Bennett, D. Ackerman, W. Park, J. Bogovic, N. Eckstein, A. Petruncio, J. Clements, S. Pang, C. S. Xu, *et al.*, "Whole-cell organelle segmentation in volume electron microscopy," *Nature*, vol. 599, no. 7883, pp. 141–146, 2021.

[99] C. Angermueller, S. J. Clark, H. J. Lee, I. C. Macaulay, M. J. Teng, T. X. Hu, F. Krueger, S. A. Smallwood, C. P. Ponting, T. Voet, *et al.*, "Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity," *Nature methods*, vol. 13, no. 3, pp. 229–232, 2016.

[100] L. F. Cheow, E. T. Courtois, Y. Tan, R. Viswanathan, Q. Xing, R. Z. Tan, D. S. Tan, P. Robson, Y.-H. Loh, S. R. Quake, *et al.*, "Single-cell multimodal profiling reveals cellular epigenetic heterogeneity," *Nature methods*, vol. 13, no. 10, pp. 833–836, 2016.

[101] S. Chen, B. B. Lake, and K. Zhang, "High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell," *Nature biotechnology*, vol. 37, no. 12, pp. 1452–1457, 2019.

[102] S. Ma, B. Zhang, L. M. LaFave, A. S. Earl, Z. Chiang, Y. Hu, J. Ding, A. Brack, V. K. Kartha, T. Tay, *et al.*, "Chromatin potential identified by shared single-cell profiling of rna and chromatin," *Cell*, vol. 183, no. 4, pp. 1103–1116, 2020.

[103] J. Liu, Y. Huang, R. Singh, J.-P. Vert, and W. S. Noble, "Jointly embedding multiple single-cell omics measurements," in *Algorithms in bioinformatics:... International Workshop, WABI..., proceedings. WABI (Workshop)*, vol. 143, NIH Public Access, 2019.

[104] K. D. Yang, A. Belyaeva, S. Venkatachalapathy, K. Damodaran, A. Katcoff, A. Radhakrishnan, G. Shivashankar, and C. Uhler, "Multi-domain translation between single-cell imaging and sequencing data using autoencoders," *Nature communications*, vol. 12, no. 1, pp. 1–10, 2021.

[105] R. Argelaguet, B. Velten, D. Arnol, S. Dietrich, T. Zenz, J. C. Marioni, F. Buettner, W. Huber, and O. Stegle, "Multi-omics factor analysisa framework for unsupervised integration of multi-omics data sets," *Molecular systems biology*, vol. 14, no. 6, p. e8124, 2018.

[106] Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck III, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, *et al.*, "Integrated analysis of multimodal single-cell data," *Cell*, vol. 184, no. 13, pp. 3573–3587, 2021.

[107] R. Argelaguet, S. J. Clark, H. Mohammed, L. C. Stapel, C. Krueger, C.-A. Kapourani, I. Imaz-Rosshandler, T. Lohoff, Y. Xiang, C. W. Hanna, *et al.*, "Multi-omics profiling of mouse gastrulation at single-cell resolution," *Nature*, vol. 576, no. 7787, pp. 487–491, 2019.

[108] N. Milyaev, D. Osumi-Sutherland, S. Reeve, N. Burton, R. A. Baldock, and J. D. Armstrong, "The virtual fly brain browser and query interface," *Bioinformatics*, vol. 28, no. 3, pp. 411–415, 2012.

[109] A. Jenett, G. M. Rubin, T.-T. Ngo, D. Shepherd, C. Murphy, H. Dionne, B. D. Pfeiffer, A. Cavallaro, D. Hall, J. Jeter, *et al.*, "A gal4-driver line resource for drosophila neurobiology," *Cell reports*, vol. 2, no. 4, pp. 991–1001, 2012.

[110] A. S. Bates, J. Janssens, G. S. Jefferis, and S. Aerts, "Neuronal cell types in the fly: single-cell anatomy meets single-cell genomics," *Current opinion in neurobiology*, vol. 56, pp. 125–134, 2019.

[111] J. Clements, C. Goina, P. M. Hubbard, T. Kawase, D. J. Olbris, H. Otsuna, R. Svirskas, and K. Rokicki, "Neuronbridge: an intuitive web application for neuronal morphology search across large data sets," *bioRxiv*, 2022.

[112] Z. Altun, L. Herndon, C. Wolkow, C. Crocker, R. Lints, and D. Hall, "Wormatlas." `http://www.wormatlas.org`, 2002.

[113] H. L. Sladitschek, U.-M. Fiuza, D. Pavlinic, V. Benes, L. Hufnagel, and P. A. Neveu, "Morphoseq: full single-cell transcriptome dynamics up to gastrulation in a chordate," *Cell*, vol. 181, no. 4, pp. 922–935, 2020.

[114] K. Achim, J.-B. Pettit, L. R. Saraiva, D. Gavriouchkina, T. Larsson, D. Arendt, and J. C. Marioni, "High-throughput spatial mapping of single-cell rna-seq data to tissue of origin," *Nature biotechnology*, vol. 33, no. 5, pp. 503–509, 2015.

[115] S. K. Longo, M. G. Guo, A. L. Ji, and P. A. Khavari, "Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics," *Nature Reviews Genetics*, vol. 22, no. 10, pp. 627–644, 2021.

[116] J. R. Moffitt, D. Bambah-Mukku, S. W. Eichhorn, E. Vaughn, K. Shekhar, J. D. Perez, N. D. Rubinstein, J. Hao, A. Regev, C. Dulac, *et al.*, "Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region," *Science*, vol. 362, no. 6416, p. eaau5324, 2018.

[117] X. Tan, A. Su, M. Tran, and Q. Nguyen, "Spacell: integrating tissue morphology and spatial gene expression to predict disease cells," *Bioinformatics*, vol. 36, no. 7, pp. 2293–2294, 2019.

[118] F. Bao, Y. Deng, S. Wan, S. Q. Shen, B. Wang, Q. Dai, S. J. Altschuler, and L. F. Wu, "Integrative spatial analysis of cell morphologies and transcriptional states with muse," *Nature biotechnology*, pp. 1–10, 2022.

[119] P. De Boer, J. P. Hoogenboom, and B. N. Giepmans, "Correlated light and electron microscopy: ultrastructure lights up!," *Nature methods*, vol. 12, no. 6, pp. 503–513, 2015.

[120] B. D. Özpolat, N. Randel, E. A. Williams, L. A. Bezares-Calderón, G. Andreatta, G. Balavoine, P. Y. Bertucci, D. E. Ferrier, M. C. Gambi, E. Gazave, *et al.*, "The nereid on the rise: Platynereis as a model system," *EvoDevo*, vol. 12, no. 1, pp. 1–22, 2021.

[121] C. Verasztó, S. Jasek, M. Gühmann, R. Shahidi, N. Ueda, J. D. Beard, S. Mendes, K. Heinz, L. A. Bezares-Calderón, E. Williams, *et al.*, "Whole-animal connectome and cell-type complement of the three-segmented platynereis dumerilii larva," *BioRxiv*, 2020.

[122] H. M. Vergara, P. Y. Bertucci, P. Hantz, M. A. Tosches, K. Achim, P. Vopalensky, and D. Arendt, "Whole-organism cellular gene-expression atlas reveals conserved cell types in the ventral nerve cord of platynereis dumerilii," *Proceedings of the National Academy of Sciences*, vol. 114, no. 23, pp. 5878–5885, 2017.

[123] B. Titze, C. Genoud, and R. W. Friedrich, "Sbemimage: versatile acquisition control software for serial block-face electron microscopy," *Frontiers in neural circuits*, vol. 12, p. 54, 2018.

[124] K. Tessmar-Raible, P. R. Steinmetz, H. Snyman, M. Hassel, and D. Arendt, "Fluorescent two-color whole mount in situ hybridization in platynereis dumerilii (polychaeta, annelida), an emerging marine molecular model for evolution and development," *Biotechniques*, vol. 39, no. 4, pp. 460–464, 2005.

[125] C. Pape, T. Beier, P. Li, V. Jain, D. D. Bock, and A. Kreshuk, "Solving large multicut problems for connectomics via domain decomposition," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 1–10, 2017.

[126] T. Pietzsch, S. Saalfeld, S. Preibisch, and P. Tomancak, "Bigdataviewer: visualization and processing for large image data sets," *Nature methods*, vol. 12, no. 6, pp. 481–483, 2015.

[127] J. Maitin-Shepard, A. Baden, W. Silversmith, E. Perlman, F. Collman, T. Blakely, J. Funke, C. Jordan, B. Falk, N. Kemnitz, tingzhao, C. Roat, M. Castro, S. Jagannathan, moenigin, J. Clements, A. Hoag, B. Katz, D. Parsons, J. Wu, L. Kamentsky, P. Chervakov, P. Hubbard, S. Berg, J. Hoffer, A. Halageri, C. Machacek, K. Mader, L. Roeder, and P. H. Li, "google/neuroglancer:," Oct. 2021.

[128] C. Pape, K. Meechan, E. Moreva, M. Schorb, N. Chiaruttini, V. Zinchenko, H. Vergara, G. Mizzon, J. Moore, D. Arendt, *et al.*, "Mobie: A fiji plugin for sharing and exploration of multi-modal cloud-hosted big image data," *bioRxiv*, 2022.

[129] P. Vallotton, R. Lagerstrom, C. Sun, M. Buckley, D. Wang, M. De Silva, S.-S. Tan, and J. M. Gunnersen, "Automated analysis of neurite branching in cultured cortical neurons using hca-vision," *Cytometry Part A: The Journal of the International Society for Analytical Cytology*, vol. 71, no. 10, pp. 889–895, 2007.

[130] B. A. Barad, M. Medina, D. Fuentes, R. L. Wiseman, and D. A. Grotjahn, "A surface morphometrics toolkit to quantify organellar membrane ultrastructure using cryo-electron tomography," *bioRxiv*, 2022.

[131] G. R. Johnson, T. E. Buck, D. P. Sullivan, G. K. Rohde, and R. F. Murphy, "Joint modeling of cell and nuclear shape variation," *Molecular Biology of the Cell*, vol. 26, no. 22, pp. 4046–4056, 2015.

[132] X. Ruan, G. R. Johnson, I. Bierschenk, R. Nitschke, M. Boerries, H. Busch, and R. F. Murphy, "Image-derived models of cell organization changes during differentiation and drug treatments," *Molecular biology of the cell*, vol. 31, no. 7, pp. 655–666, 2020.

[133] M. K. Driscoll, E. S. Welf, A. R. Jamieson, K. M. Dean, T. Isogai, R. Fiolka, and G. Danuser, "Robust and automated detection of subcellular morphological motifs in 3d microscopy images," *Nature methods*, vol. 16, no. 10, pp. 1037–1044, 2019.

[134] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[135] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[136] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

[137] A. Motta, M. Berning, K. M. Boergens, B. Staffler, M. Beining, S. Loomba, P. Hennig, H. Wissler, and M. Helmstaedter, "Dense connectomic reconstruction in layer 4 of the somatosensory cortex," *Science*, vol. 366, no. 6469, p. eaay3134, 2019.

[138] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[139] A. Farahani, B. Pourshojae, K. Rasheed, and H. R. Arabnia, "A concise review of transfer learning," in *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 344–351, IEEE, 2020.

[140] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[141] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM computing surveys (csur)*, vol. 53, no. 3, pp. 1–34, 2020.

[142] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

[143] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[144] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[145] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21271–21284, 2020.

[146] J. C. Caicedo, C. McQuin, A. Goodman, S. Singh, and A. E. Carpenter, "Weakly supervised learning of single-cell feature embeddings," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9309–9318, 2018.

[147] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," tech. rep., California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[148] A. X. Lu, O. Z. Kraus, S. Cooper, and A. M. Moses, "Learning unsupervised feature representations for single cell microscopy images with paired cell inpainting," *PLoS computational biology*, vol. 15, no. 9, p. e1007348, 2019.

[149] M. W. Lafarge, J. C. Caicedo, A. E. Carpenter, J. P. Pluim, S. Singh, and M. Veta, "Capturing single-cell phenotypic variation via unsupervised representation learning," in *International Conference on Medical Imaging with Deep Learning*, pp. 315–325, PMLR, 2019.

[150] P. J. Schubert, S. Dorkenwald, M. Januszewski, V. Jain, and J. Kornfeld, "Learning cellular morphology with neural networks," *Nature communications*, vol. 10, no. 1, pp. 1–12, 2019.

[151] G. B. Huang, H.-F. Yang, S.-y. Takemura, P. Rivlin, and S. M. Plaza, "Latent feature representation via unsupervised learning for pattern discovery in massive electron microscopy image volumes," *arXiv preprint arXiv:2012.12175*, 2020.

[152] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," *Advances in neural information processing systems*, vol. 29, 2016.

[153] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[154] V. A. Traag, L. Waltman, and N. J. Van Eck, "From louvain to leiden: guaranteeing well-connected communities," *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.

[155] S. Song, V. Starunov, X. Bailly, C. Ruta, P. Kerner, A. J. Cornelissen, and G. Balavoine, "Globins in the marine annelid platynereis dumerilii shed new light on hemoglobin evolution in bilaterians," *BMC evolutionary biology*, vol. 20, no. 1, pp. 1–23, 2020.

[156] A. Cole and M. Arnone, "Fluorescent in situ hybridization reveals multiple expression domains for spbrn1/2/4 and identifies a unique ectodermal cell type that co-expresses the parahox gene splox," *Gene Expression Patterns*, vol. 9, no. 5, pp. 324–328, 2009.

[157] A. J. White, M. J. Northcutt, S. E. Rohrback, R. O. Carpenter, M. M. Niehaus-Sauter, Y. Gao, M. G. Wheatly, and C. M. Gillen, "Characterization of sarcoplasmic calcium binding protein (scp) variants from freshwater crayfish procambarus clarkii," *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, vol. 160, no. 1, pp. 8–14, 2011.

[158] R. R. Stine, A. P. Sakers, T. TeSlaa, M. Kissig, Z. E. Stine, C. W. Kwon, L. Cheng, H.-W. Lim, K. H. Kaestner, J. D. Rabinowitz, *et al.*, "Prdm16 maintains homeostasis of the intestinal epithelium by controlling region-specific metabolism," *Cell Stem Cell*, vol. 25, no. 6, pp. 830–845, 2019.

[159] D. Golding, "The infracerebral gland in nephtysa possible neuroendocrine complex," *General and Comparative Endocrinology*, vol. 14, no. 1, pp. 114–126, 1970.

[160] D. G. Baskin, "Further observations on the fine structure and development of the infracerebral complex (infracerebral gland) of nereis limnicola (annelida, polychaeta)," *Cell and Tissue Research*, vol. 154, no. 4, pp. 519–531, 1974.

[161] D. Hofmann, "Regeneration and endocrinology in the polychaeteplatynereis dumerilii," *Wilhelm Roux's archives of developmental biology*, vol. 180, no. 1, pp. 47–71, 1976.

[162] B. Backfisch, *Regulatory tools and the characterization of insulinergic cells in the annelid Platynereis dumerilii*. na, 2013.

[163] A. Lauri, T. Brunet, M. Handberg-Thorsager, A. H. Fischer, O. Simakov, P. R. Steinmetz, R. Tomer, P. J. Keller, and D. Arendt, "Development of the annelid axochord: insights into notochord evolution," *Science*, vol. 345, no. 6202, pp. 1365–1368, 2014.

[164] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, "Dense contrastive learning for self-supervised visual pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3024–3033, 2021.

[165] Z. Xie, Y. Lin, Z. Zhang, Y. Cao, S. Lin, and H. Hu, "Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16684–16693, 2021.

[166] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman, "With a little help from my friends: Nearest-neighbor contrastive learning of visual representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9588–9597, 2021.

[167] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

[168] K. Hassani and A. H. Khasahmadi, "Contrastive multi-view representation learning on graphs," in *International Conference on Machine Learning*, pp. 4116–4126, PMLR, 2020.

[169] A. Subramanian, D. Pruthi, H. Jhamtani, T. Berg-Kirkpatrick, and E. Hovy, "Spine: Sparse interpretable neural embeddings," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[170] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, pp. 818–833, Springer, 2014.

[171] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

[172] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*, pp. 1180–1189, PMLR, 2015.

[173] C. Pape, A. Matskevych, A. Wolny, J. Hennies, G. Mizzon, M. Louveaux, J. Musser, A. Maizel, D. Arendt, and A. Kreshuk, "Leveraging domain knowledge to improve microscopy image segmentation with lifted multicuts," *Frontiers in Computer Science*, p. 6, 2019.

[174] C. Bosch, T. Ackels, A. Pacureanu, Y. Zhang, C. J. Peddie, M. Berning, N. Rzepka, M.-C. Zdora, I. Whiteley, M. Storm, *et al.*, "Functional and multiscale 3d structural investigation of brain tissue through correlative in vivo physiology, synchrotron microtomography and volume electron microscopy," *Nature communications*, vol. 13, no. 1, pp. 1–16, 2022.

[175] S. Dorkenwald, P. Li, M. Januszewski, D. R. Berger, J. Maitin-Shepard, A. L. Bodor, F. Collman, C. M. Schneider-Mizell, N. M. da Costa, J. W. Lichtman, *et al.*, "Multi-layered maps of neuropil with segmentation-guided contrastive learning," *bioRxiv*, 2022.

[176] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li, *et al.*, "Florence: A new foundation model for computer vision," *arXiv preprint arXiv:2111.11432*, 2021.

[177] Q. Li, B. Gong, Y. Cui, D. Kondratyuk, X. Du, M.-H. Yang, and M. Brown, "Towards a unified foundation model: Jointly pre-training transformers on unpaired images and text," *arXiv preprint arXiv:2112.07074*, 2021.

[178] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[179] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[180] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *arXiv preprint arXiv:2201.09873*, 2022.

[181] P. manuscript editors, A. coordination, I. data analysis Armand Ethan 42 Yao Zizhen 5, A. seq data generation, processing Fang Rongxin 45 Hou Xiaomeng 10 Lucero Jacinta D. 18 Osteen Julia K. 18 Pinto-Duarte Antonio 18 Poirion Olivier

10 Preissl Sebastian 10 Wang Xinxin 10 97, E. retro-seq data generation, processing Dominguez Bertha 53 Ito-Cole Tony 1 Jacobs Matthew 1 Jin Xin 54 99 100 Lee Cheng-Ta 53 Lee Kuo-Fen 53 Miyazaki Paula Assakura 1 Pang Yan 1 Rashid Mohammad 1 Smith Jared B. 54 Vu Minh 1 Williams Elora 54, *et al.*, "A multimodal cell census and atlas of the mammalian primary motor cortex," *Nature*, vol. 598, no. 7879, pp. 86–102, 2021.

[182] C. R. Cadwell, A. Palasantza, X. Jiang, P. Berens, Q. Deng, M. Yilmaz, J. Reimer, S. Shen, M. Bethge, K. F. Tolias, *et al.*, "Electrophysiological, transcriptomic and morphologic profiling of single neurons using patch-seq," *Nature biotechnology*, vol. 34, no. 2, pp. 199–203, 2016.

[183] J. Fuzik, A. Zeisel, Z. Máté, D. Calvigioni, Y. Yanagawa, G. Szabó, S. Linnarsson, and T. Harkany, "Integration of electrophysiological recordings with single-cell rna-seq data identifies neuronal subtypes," *Nature biotechnology*, vol. 34, no. 2, pp. 175–183, 2016.

[184] B. Tasic, Z. Yao, L. T. Graybuck, K. A. Smith, T. N. Nguyen, D. Bertagnolli, J. Goldy, E. Garren, M. N. Economo, S. Viswanathan, *et al.*, "Shared and distinct transcriptomic cell types across neocortical areas," *Nature*, vol. 563, no. 7729, pp. 72–78, 2018.

[185] A. Santella, I. Kolotuev, C. Kizilyaprak, and Z. Bao, "Cross-modality synthesis of em time series and live fluorescence imaging," *eLife*, vol. 11, p. e77918, 2022.

[186] P. Androvic, M. Schifferer, K. P. Anderson, L. Cantuti-Castelvetri, H. Ji, L. Liu, S. Besson-Girard, J. Knoferle, M. Simons, and O. Gokce, "Spatial transcriptomics-correlated electron microscopy," *bioRxiv*, 2022.

[187] B. Spanjaard, B. Hu, N. Mitic, P. Olivares-Chauvet, S. Janjuha, N. Ninov, and J. P. Junker, "Simultaneous lineage tracing and cell-type identification using crispr–cas9-induced genetic scars," *Nature biotechnology*, vol. 36, no. 5, pp. 469–473, 2018.

[188] B. Raj, D. E. Wagner, A. McKenna, S. Pandey, A. M. Klein, J. Shendure, J. A. Gagnon, and A. F. Schier, "Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain," *Nature biotechnology*, vol. 36, no. 5, pp. 442–450, 2018.

[189] A. Alemany, M. Florescu, C. S. Baron, J. Peterson-Maduro, and A. Van Oudenaarden, "Whole-organism clone tracing using single-cell sequencing," *Nature*, vol. 556, no. 7699, pp. 108–112, 2018.

[190] M. Lalit, M. Handberg-Thorsager, Y.-W. Hsieh, F. Jug, and P. Tomancak, "Registration of multi-modal volumetric images by establishing cell correspondence," in *European Conference on Computer Vision*, pp. 458–473, Springer, 2020.

[191] A. Tanay and A. Sebé-Pedrós, "Evolutionary cell type mapping with single-cell genomics," *Trends in genetics*, vol. 37, no. 10, pp. 919–932, 2021.

[192] Y. Song, Z. Miao, A. Brazma, and I. Papatheodorou, "Benchmarking strategies for cross-species integration of single-cell rna sequencing data," *bioRxiv*, 2022.

[193] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[194] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, *et al.*, "Scipy 1.0: fundamental algorithms for scientific computing in python," *Nature methods*, vol. 17, no. 3, pp. 261–272, 2020.

[195] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[196] N. Rahaman, T. Beier, C. Pape, and S. Wolf, "inferno." `https://github.com/vzinche/inferno`, 2021.

[197] C. Pape and N. Rahaman, "neurofire." `https://github.com/vzinche/neurofire`, 2021.

[198] A. Collette, T. Kluyver, T. A. Caswell, J. Tocknell, J. Kieffer, A. Scopatz, D. Dale, Chen, A. Jelenak, payno, juliagarriga, T. VINCENT, P. Sciarelli, V. Valls, S. Ghosh, U. K. Pedersen, jakirkham, M. Raspaud, A. Parsons, H. Abbasi, J. Readey, A. Paramonov, L. Chan, V. A. Sol, jialin, C. Danilevski, Y. Feng, G. A. Vaillant, M. Teichmann, and M. Brucher, "h5py/h5py: 3.2.1," Mar. 2021.

[199] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, pp. 357–362, Sept. 2020.

[200] T. pandas development team, "pandas-dev/pandas: Pandas," Feb. 2020.

[201] S. Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu, "scikit-image: image processing in python," *PeerJ*, vol. 2, p. e453, 2014.

[202] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in science & engineering*, vol. 9, no. 03, pp. 90–95, 2007.

[203] N. Sofroniew, T. Lambert, K. Evans, J. Nunez-Iglesias, G. Bokota, P. Winston, G. Pea-Castellanos, K. Yamauchi, M. Bussonnier, D. Doncila Pop, A. Can Solak, Z. Liu, P. Wadhwa, A. Burt, G. Buckley, A. Sweet, L. Migas, V. Hilsenstein,

L. Gaifas, J. Bragantini, J. Rodrguez-Guerra, H. Muoz, J. Freeman, P. Boone, A. Lowe, C. Gohlke, L. Royer, A. PIERR, H. Har-Gil, and A. McGovern, "napari: a multi-dimensional image viewer for Python," Oct. 2022.

[204] P. Schlegel, "pymaid." `https://github.com/navis-org/pymaid`, 2017.

[205] C. Pape, C. Barnes, paulhfu, and D. P. Orfanos, "constantinpape/z5," Oct. 2022.

[206] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[207] C. Pape, "elf." `https://github.com/constantinpape/elf`, 2021.

[208] C. Pape and B. Mohar, "pybdv." `https://github.com/constantinpape/pybdv`, 2021.

[209] U. Köthe, "Reusable software in computer vision," *Handbook of computer vision and applications*, vol. 3, pp. 103–132, 1999.

[210] P. Ramachandran and G. Varoquaux, "Mayavi: 3D Visualization of Scientific Data," *Computing in Science & Engineering*, vol. 13, no. 2, pp. 40–51, 2011.

[211] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[212] G. Dubourg-Felonneau, S. Shams, E. Akiva, and L. Lee, "Protein organization with manifold exploration and spectral clustering," *bioRxiv*, 2021.

[213] A. Hagberg, P. Swart, and D. S Chult, "Exploring network structure, dynamics, and function using networkx," tech. rep., Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[214] G. Csrdi, T. Nepusz, S. Horvt, V. Traag, F. Zanini, and D. Noom, "igraph," Oct. 2022.

[215] M. L. Waskom, "seaborn: statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021.

Appendix

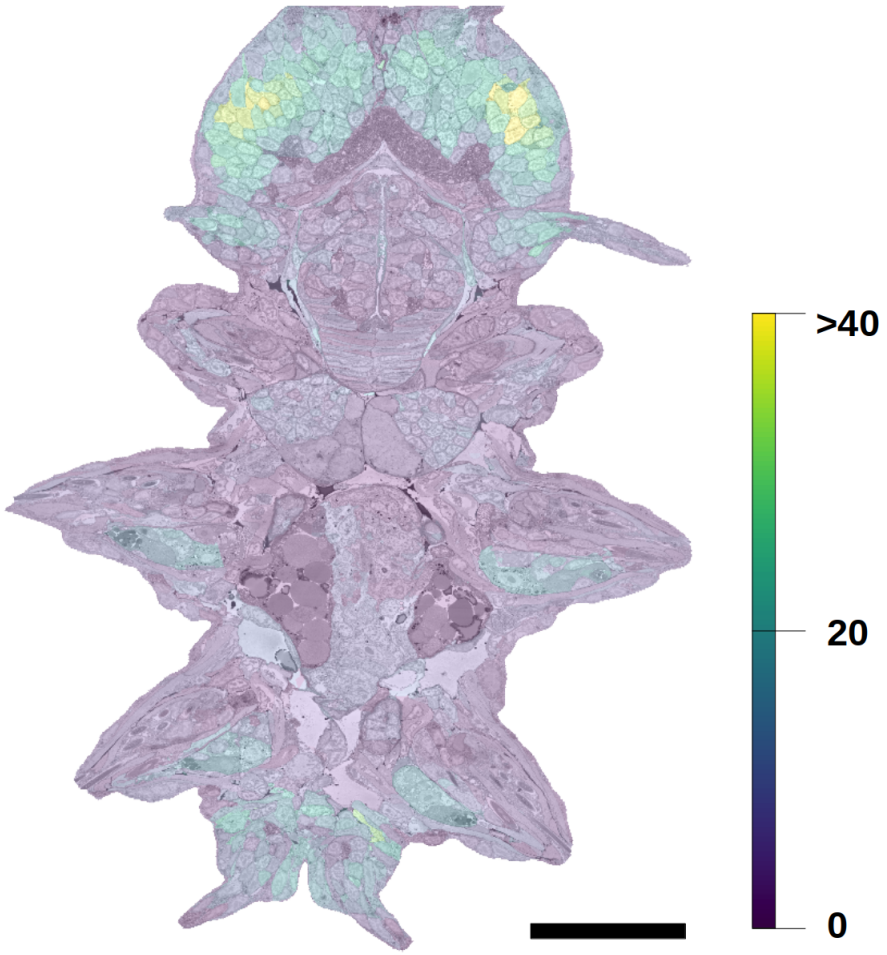


Figure S1: Gene expression density in the animal after the Virtual Cell assignment. The colourmap represents the number of genes assigned to a cell.

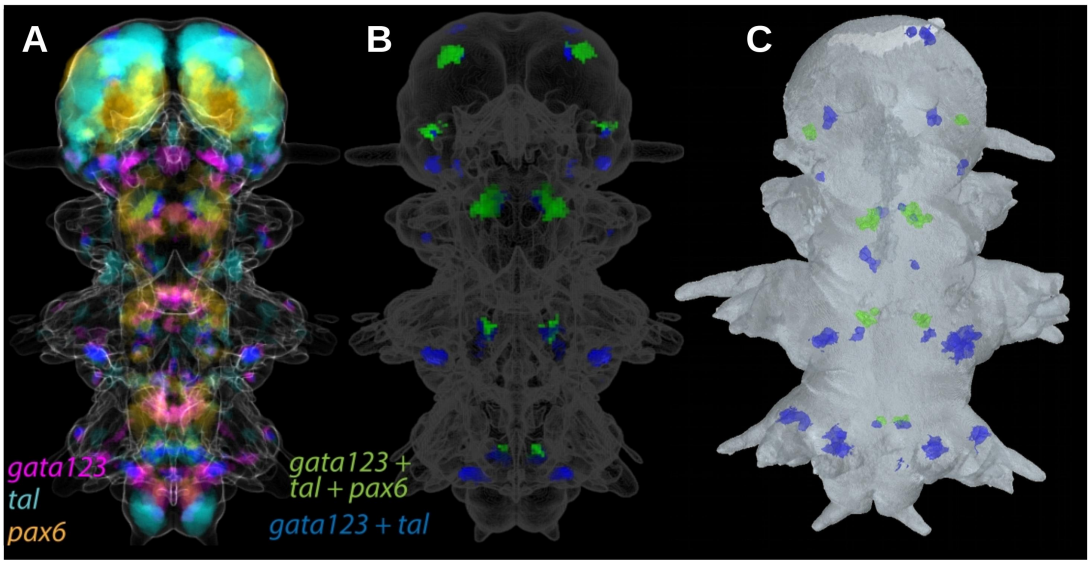


Figure S2: Gene expression maps and Virtual Cells (VCs). **A.** Gene expression maps for genes *gata123*, *tal* and *pax6* are shown in the animal volume. **B.** Visualised are the Virtual Cells that express all three genes (green) or just *gata123* and *tal* (blue). **C.** The cells segmented in the EM volume that have been assigned the VCs from B are shown in the EM volume.

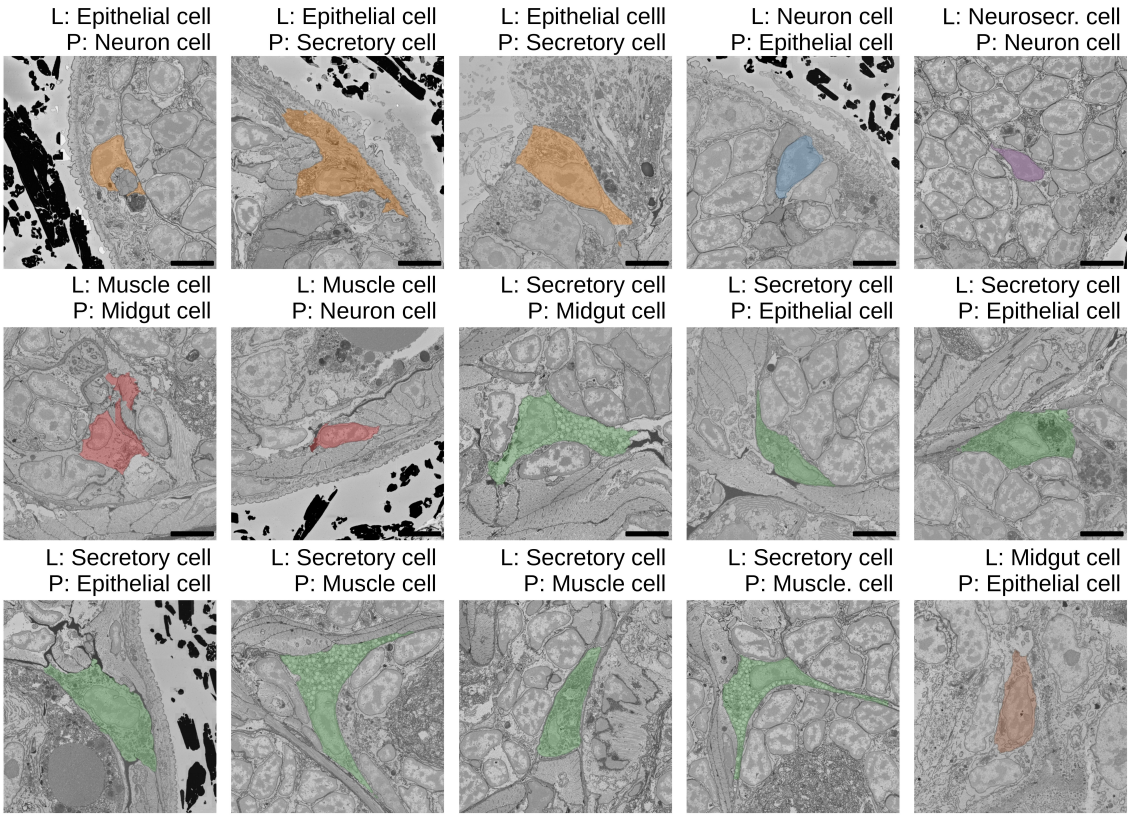


Figure S3: Errors of the logistic regression model. For each wrongly predicted cell the true label (L) and the predicted class (P) are shown.

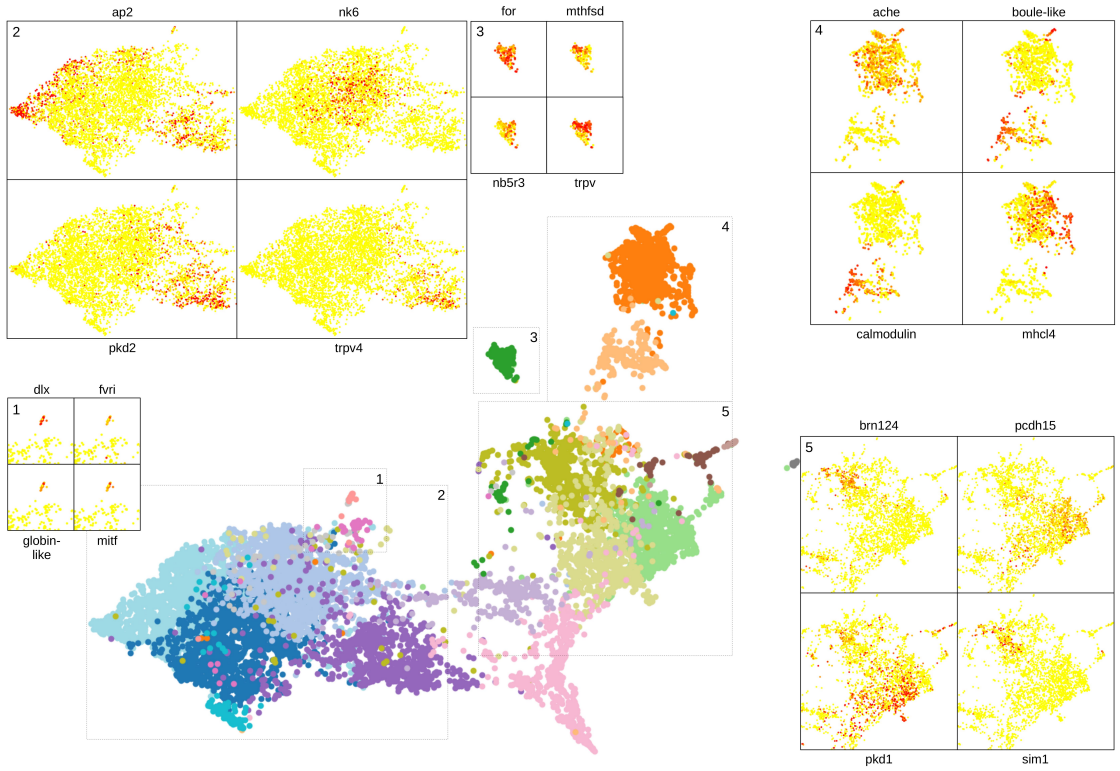


Figure S4: Specific gene expression. Selected genes specific for the clusters of rhabdomeric photoreceptors (1), neurons (2), midgut cells (3), muscles (4) and epithelial cells (5) are plotted on the UMAP representation from Figure 5.10.

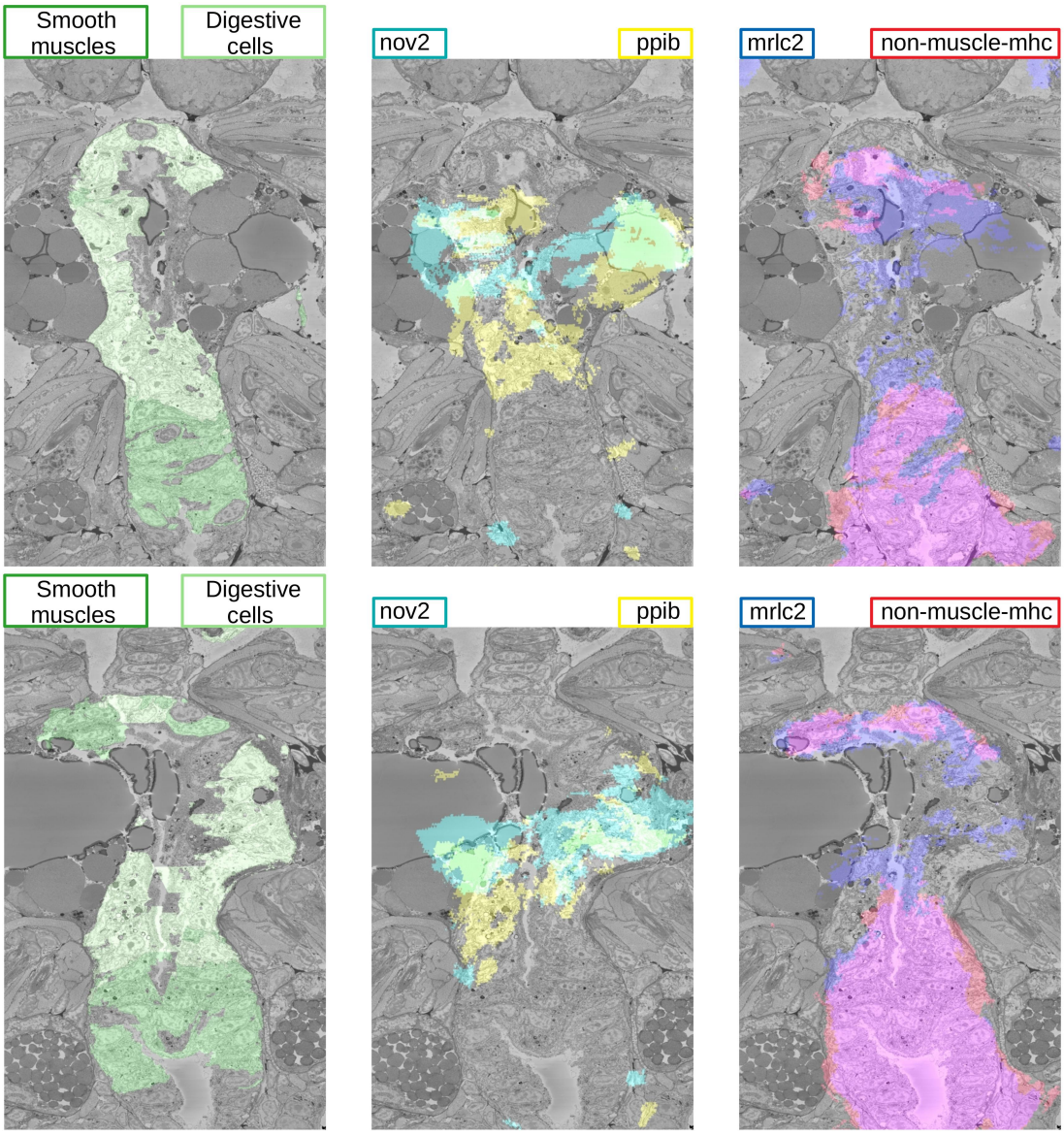


Figure S5: Two clusters of midgut cells and their specific genes. Two EM sections (upper and lower rows) of the animal midgut are shown, visualising the location of the two clusters (left panel), the expression of the genes *nov2* and *ppib*, specific to the cluster of epithelial digestive cells (central panel) and the expression of the genes *mrlc2* and *non-muscle-mhc*, specific to the cluster of smooth muscles (right panel).

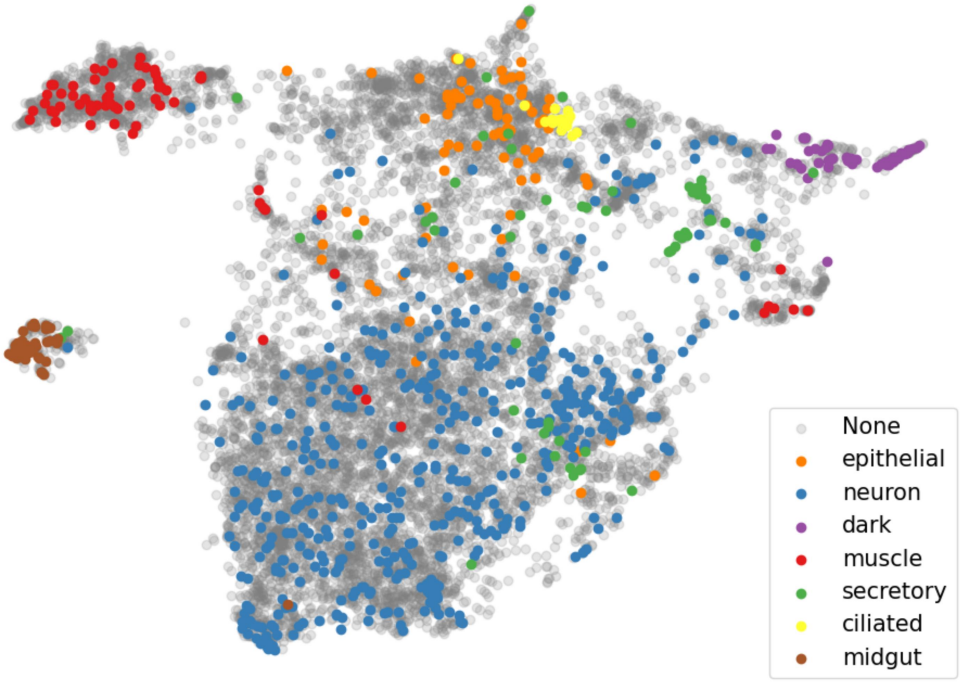


Figure S6: UMAP representation of all cells in the animal based on their nuclear MorphoFeatures. The manually annotated cell classes are shown in colors.

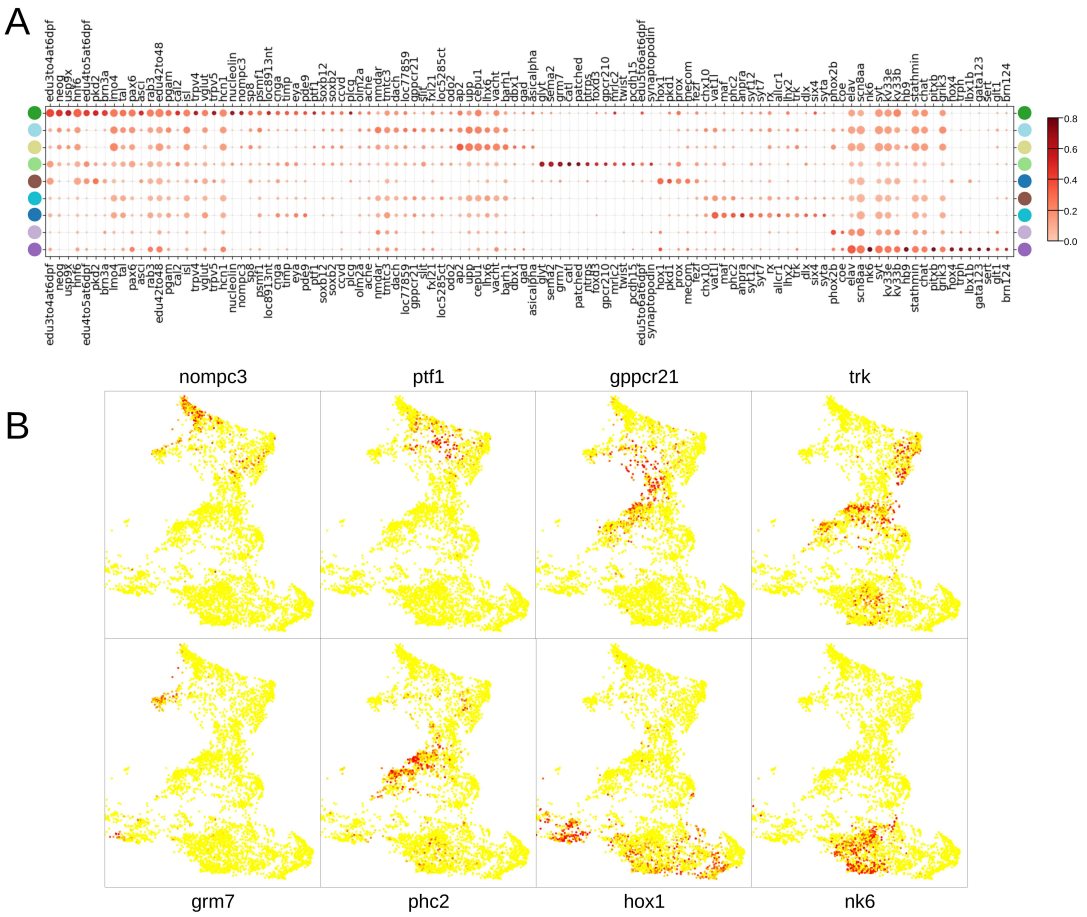


Figure S7: Gene analysis of ganglia clusters. **A.** Gene expression dot plot. The dot size reflects the fraction of the cluster expressing a gene; the dot colour reflects the gene specificity to a cluster (see Methods). **B.** Some of the genes from **A** plotted on the ganglia part of the UMAP representation from Figure 5.18.

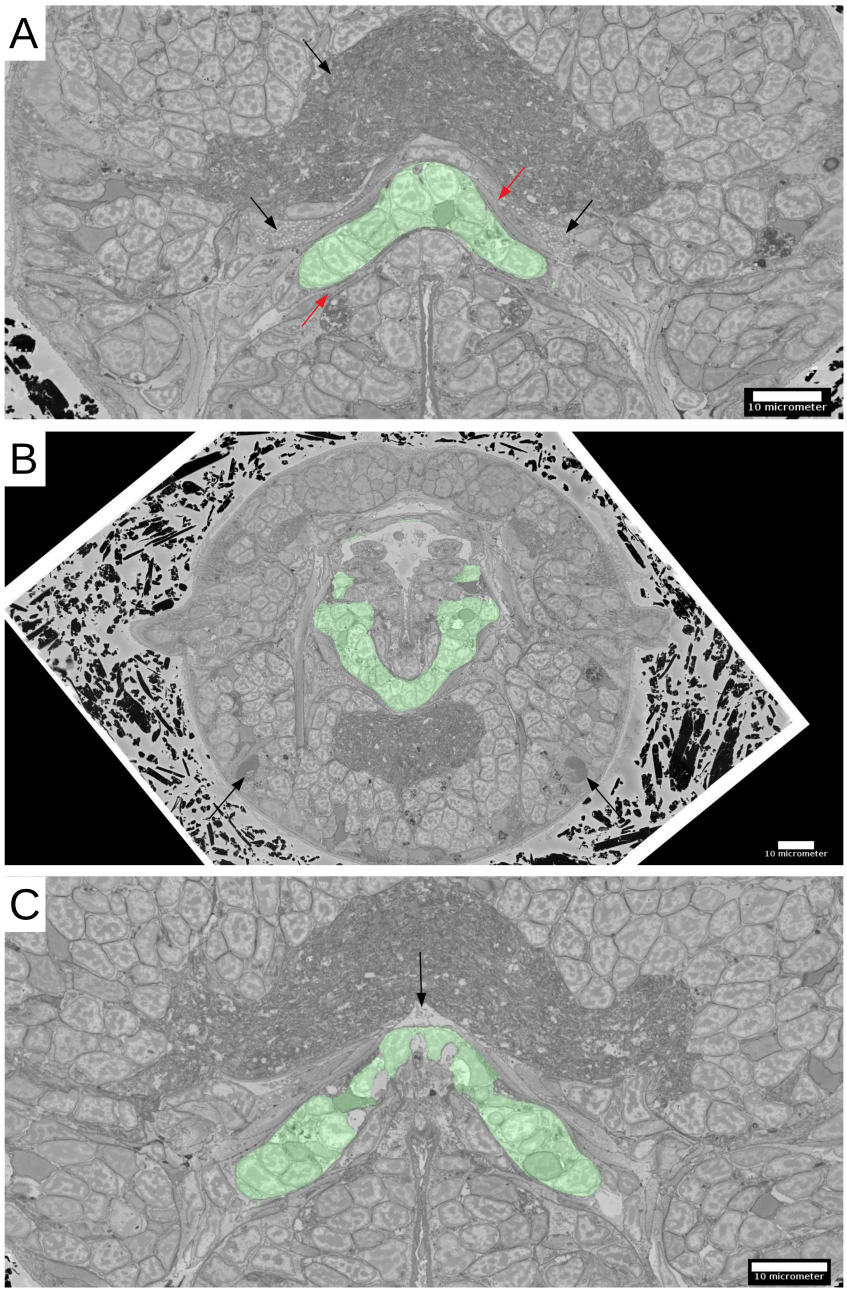


Figure S8: Infracerebral gland location and anatomy. **A.** Dorso-ventral view. The black arrows point to the neuropil and surrounding secretory cells, the red arrows - to surrounding muscles. **B.** Anterior-posterior view. The black arrows point to the posterior pair of adult eyes. **C.** Dorso-ventral view, ventrally to (A). The black arrow points to a cavity that might be a developing blood vessel.

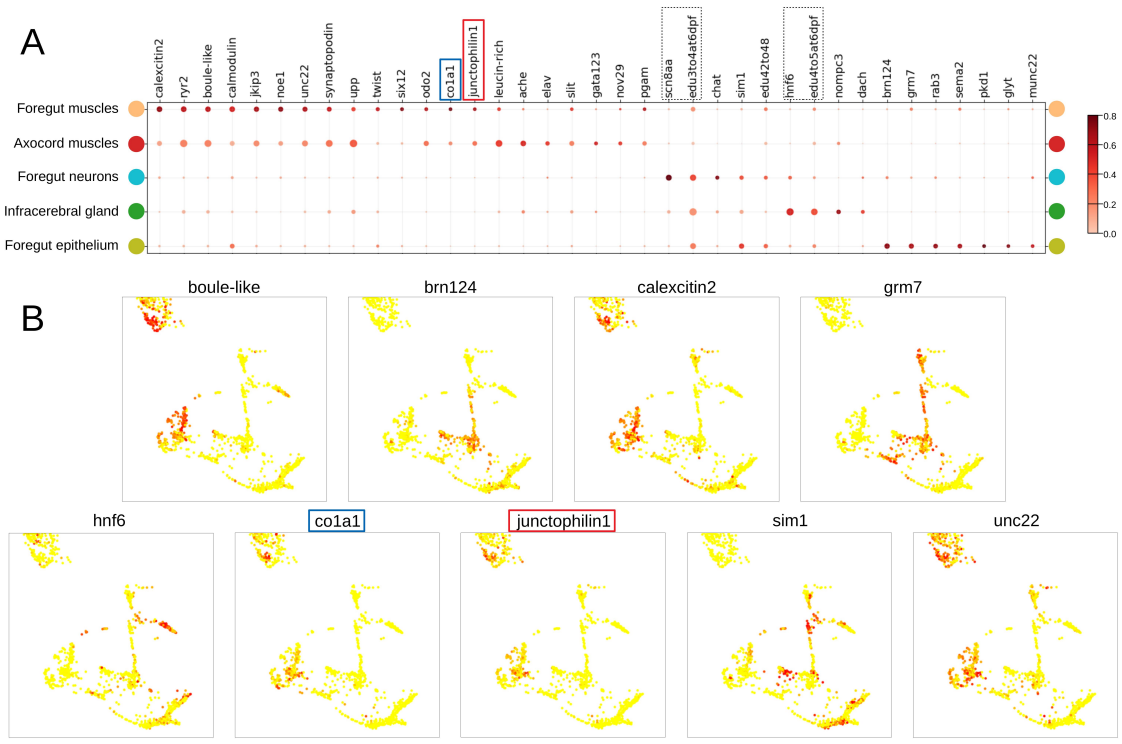


Figure S9: Gene analysis of foregut clusters. **A.** Gene expression dot plot. The dot size reflects the fraction of the cluster expressing a gene; the dot colour reflects the gene specificity to a cluster (see Methods). **B.** Some of the genes from **A** plotted on the foregut part of the UMAP representation from Figure 5.18.