Dissertation

submitted to the
Combined Faculty of Natural Sciences and Mathematics
of the Ruperto Carola University Heidelberg, Germany

for the degree of
Doctor of Natural Sciences

Presented by

Jonas Hartmann, M.Sc.
born in Wallisellen ZH, Switzerland

Oral examination: 14. March 2019

Complexity in Developmental Systems:

Toward an Integrated Understanding of Organ Formation

Referees: Dr. Stefano De Renzis

Prof. Dr. Ursula Klingmüller

Abstract

During animal development, embryonic cells assemble into intricately structured organs by working together in organized groups capable of implementing tightly coordinated collective behaviors, including patterning, morphogenesis and migration. Although many of the molecular components and basic mechanisms underlying such collective phenomena are known, the complexity emerging from their interplay still represents a major challenge for developmental biology.

Here, we first clarify the nature of this challenge and outline three key strategies for addressing it: precision perturbation, synthetic developmental biology, and data-driven inference. We then present the results of our effort to develop a set of tools rooted in two of these strategies and to apply them to uncover new mechanisms and principles underlying the coordination of collective cell behaviors during organogenesis, using the zebrafish posterior lateral line primordium as a model system.

To enable precision perturbation of migration and morphogenesis, we sought to adapt optogenetic tools to control chemokine and actin signaling. This endeavor proved far from trivial and we were ultimately unable to derive functional optogenetic constructs. However, our work toward this goal led to a useful new way of perturbing cortical contractility, which in turn revealed a potential role for cell surface tension in lateral line organogenesis.

Independently, we hypothesized that the lateral line primordium might employ plithotaxis to coordinate organ formation with collective migration. We tested this hypothesis using a novel optical tool that allows targeted arrest of cell migration, finding that contrary to previous assumptions plithotaxis does not substantially contribute to primordium guidance.

Finally, we developed a computational framework for automated single-cell segmentation, latent feature extraction and quantitative analysis of cellular architecture. We identified the key factors defining shape heterogeneity across primordium cells and went on to use this shape space as a reference for mapping the results of multiple experiments into a quantitative atlas of primordium cell architecture. We also propose a number of data-driven approaches to help bridge the gap from big data to mechanistic models.

Overall, this study presents several conceptual and methodological advances toward an integrated understanding of complex multi-cellular systems.

Zusammenfassung

Die Entwicklung tierischer Embryonen ist ein Prozess, bei dem Zellen eng zusammenarbeiten um vielfältig strukturierte Organe zu bilden. Dabei kommen gut koordinierte kollektive Abläufe zum Einsatz, insbesondere Musterbildung, Morphogenese und Zellmigration. Obwohl die molekularen Komponenten und Mechanismen, die diesen Phänomenen zugrunde liegen, bereits weitestgehend bekannt sind, ist die durch deren Wechselwirkung emergierende Komplexität nach wie vor eine grosse Herausforderung für die Entwicklungsbiologie.

Wir beleuchten zunächst diese Herausforderung genauer und schlagen drei Strategien zu ihrer Bewältigung vor: Präzisionsperturbation, synthetische Entwicklungsbiologie und datengestützte Inferenz. Dann präsentieren wir unsere Versuche, basierend auf zwei dieser Strategien neue Methoden zu entwickeln und die Prinzipien der Koordination kollektiver Entwicklungsprozesse zu studieren, wozu wir das Seitenlinienprimordium des Zebrabärblings als Modellsystem nutzten.

Wir adaptierten mehrere optogenetische Methoden mit dem Ziel, Zellmigration und Morphogenese durch Präzisionsperturbationen untersuchen zu können. Es erwies sich jedoch als nicht möglich, funktionale optogenetische Konstrukte zu entwickeln. Indirekt hat sich aus dieser Arbeit aber eine neue Methode zur Erhöhung kortikaler Kontraktilität ergeben, wodurch wir Hinweise darauf erhalten haben, dass die Zelloberflächenspannung eine wichtige Rolle in der Entwicklung des Seitenlinien-organs spielen könnte.

Davon unabhängig haben wir die Hypothese verfolgt, dass die Organbildung und die kollektive Migration des Seitenlinienprimordiums durch Plithotaxis gekoppelt sein könnten. Eine direkte Über-prüfung dieser Hypothese unter Einsatz einer neu entwickelten optischen Methode zum gezielten Anhalten migrierender Zellen sprach jedoch entgegen etablierter Annahmen nicht für die Existenz eines solchen Mechanismus.

Schliesslich haben wir ein computergestütztes System zur automatischen Segmentierung einzelner Primordiumzellen sowie zur Messung und Analyse charakteristischer Eigenschaften der Zell-architektur entwickelt. Insbesondere haben wir die verschiedenen Zellformen des Primordiums quantitativ beschrieben und als Referenz dazu genutzt, mehrere unabhängige Experimente zu einem Atlas der Zellarchitektur zu kombinieren. Basierend darauf schlagen wir Methoden vor, um die Lücke zwischen "Big Data" und mechanistischen Modellen zu schliessen.

Insgesamt präsentiert diese Studie mehrere konzeptuelle und methodologische Fortschritte in Richtung eines integrierten Verständnisses komplexer mehrzelliger Systeme.

Table of Contents

1 Introduction

1.1 Cell Collectives as Integrated Systems

The essential feature of multi-cellular life is that cells cooperate to collectively perform functions they could not perform as individuals, enabling the evolution of a diverse range of organisms that could otherwise not exist.

This is exemplified most strikingly in the diversity of multi-cellular structures and functions arising during animal development. Derived from a single zygote, genetically identical embryonic cells cooperate to form a multitude of differently organized assemblies that mature into fundamentally different functional organs.

Common collective behaviors required to achieve this are the establishment of spatial patterns of cell identity (patterning), the alteration of tissue geometry in three-dimensional space (morphogenesis), and the relocation of groups of cells within the embryo (migration) (see section 1.4).

To accomplish these tasks, cells must be capable of measuring their surroundings, communicating and interpreting information, changing their shape, and exerting mechanical forces – all in a well-coordinated fashion within and across tissues. Much is known about the basic molecular modules that underpin these cellular capabilities (see section 1.4) and it is through the use and reuse of such modules in a myriad different configurations that cell collectives produce the diverse phenomena observed in a developing embryo.

However, this integration of modules is not merely an additive assembly of building blocks. Instead, modules are linked into dynamic, highly coupled and multi-layered systems teeming with feedback loops and non-linearities that drive unexpected and context-sensitive system behaviors (see section 1.2). This emergent complexity explains why it remains extremely challenging to understand and predict collective cell behavior despite the large amount of knowledge available about the building blocks themselves. In other words, the very source of the versatility that accelerates the evolution of multi-cellular life also makes multi-cellular life hard to study.

It is therefore necessary to investigate multi-cellular systems from an integrated perspective (see section 1.3), taking into account the emergent properties and methodological challenges that arise from the interplay of multiple biological processes within a system. Only from such a perspective can we hope to fully elucidate the principles of collective cell behavior.

1.2 Complexity in Biological Systems

To confront the challenge presented by the highly interconnected and convoluted nature of cells and cell collectives, it is useful to consider the high-level properties of such systems, the most notable of which is *complexity*.

Despite being frequently used by biologists, the terms *complexity* and *complex system* lack universally acknowledged definitions and are often interpreted in wildly different ways or used in a colloquial and vague manner. This lack of a clear terminology and by extension the lack of a common logical framework or language for reasoning about complex systems adds another layer of difficulty to an already very hard problem.

Although addressing this issue in full is beyond the scope of this thesis, this section aims to carve out a useful definition of complexity and to introduce how it relates to biological systems, including its implications for study design. All research presented in this thesis is fundamentally inspired by these considerations.

1.2.1 Introduction & Definition

The term *complexity* has a different meaning in everyday language (where it is a synonym for *complicatedness*) and in each of multiple different fields of science, notably computer science [Dean, 2016], physics [Holovatch et al., 2017], and chemistry [Zayed et al., 2009]. However, the most useful notion of complexity for biological systems is what has also been referred to more specifically as *deterministic complexity* [Manson, 2001; Mazzocchi, 2008].

Deterministic complexity is best understood in contrast to two related pairs of terms: s*imple and complicated* and o*rdered and chaotic*. The first pair relates to system architecture, i.e. the number and diversity of parts and interactions making up the system. The second pair, ordered and chaotic, describes system behavior and in particular the predictability of system behavior. Both simple and complicated systems can exhibit ordered or chaotic behaviors.

Chaotic systems are unpredictable in the sense that any error in the estimation of their initial conditions or parameters will lead to a completely different prediction from the outcome that is observed. Although chaotic systems may be deterministic in principle and thus completely predictable given perfect information about the system, in practice they are unpredictable because even the measurement error on the initial conditions is enough for prediction and reality to diverge rapidly [Persson & Wagner, 1995; Manson, 2001].

As an example, consider the flow of a viscous liquid around a cylindrical obstacle (Fig. 1.1a-c) [Van Dyke, 1982]. This is a comparatively simple system since it only consists of the molecular components of the liquid and of the obstacle, which we can assume to be governed by relatively straightforward

and uniform kinetic and electrostatic rules. Given a low-viscosity fluid, a stable laminar flow is established (Fig. 1.1a). In this case, the trajectory of a drop of a dye added upstream of the obstacle can be predicted easily and accurately, even with a simplified model that does not account for every individual molecule involved. However, if the viscosity of the fluid is sufficiently high, the flow behind the obstacle becomes turbulent and for most initial positions of the drop of dye the final outcome is impossible to predict without a perfect model and perfect knowledge of the initial state (Fig. 1.1c). Note that such true chaos is likely rare in cellular and multi-cellular systems because it would prevent a controlled link between heritable information and fitness-relevant biological outcomes.

Complexity is a type of system behavior found at the interface of ordered and chaotic domains, combining aspects of both (definition 1). Remarkably, this combination does not simply result in ordered behavior with the occasional random fluctuation – instead, new and surprising behaviors emerge. This becomes immediately evident in the fluid flow example if the viscosity is tuned just right (Fig. 1.1b).

Definition 1: Let ***complexity*** be a system property manifest in system behaviors that combine aspects of ordered and chaotic behaviors. A system that has complexity is a ***complex system*** and exhibits ***complex behaviors***.

Complexity in this sense is distinct from both the colloquial use of the word, which indicates simply a high complicatedness (i.e. a large number of distinct parts and interactions), and from another common way in which the term complexity is used in biology, which is to indicate that a system produces macroscopic behaviors that are not trivially related to the system's components (often expressed in the phrase "the whole is greater than the sum of its parts"). However, this notion is already suitably encompassed in the term *emergence* (definition 2) and its conflation with complexity is therefore not useful. Incidentally, complexity (as defined here) is itself an emergent property.

Definition 2: Let ***emergence*** be the phenomenon of macroscopic system properties and behaviors arising from the interactions of the system's parts (and thus not being inherent in those parts). Macroscopic properties and behaviors arising in this way are ***emergent properties*** and ***emergent behaviors***.

Given the issues surrounding the multiple uses of the term complexity, it could be argued that the term should be left to its colloquial use and new terms should be introduced to capture the different concepts currently termed "complexity". Although this may be an important semantic question, it is a question best addressed elsewhere, which is why in this thesis the term complexity is used in accordance with definition 1.

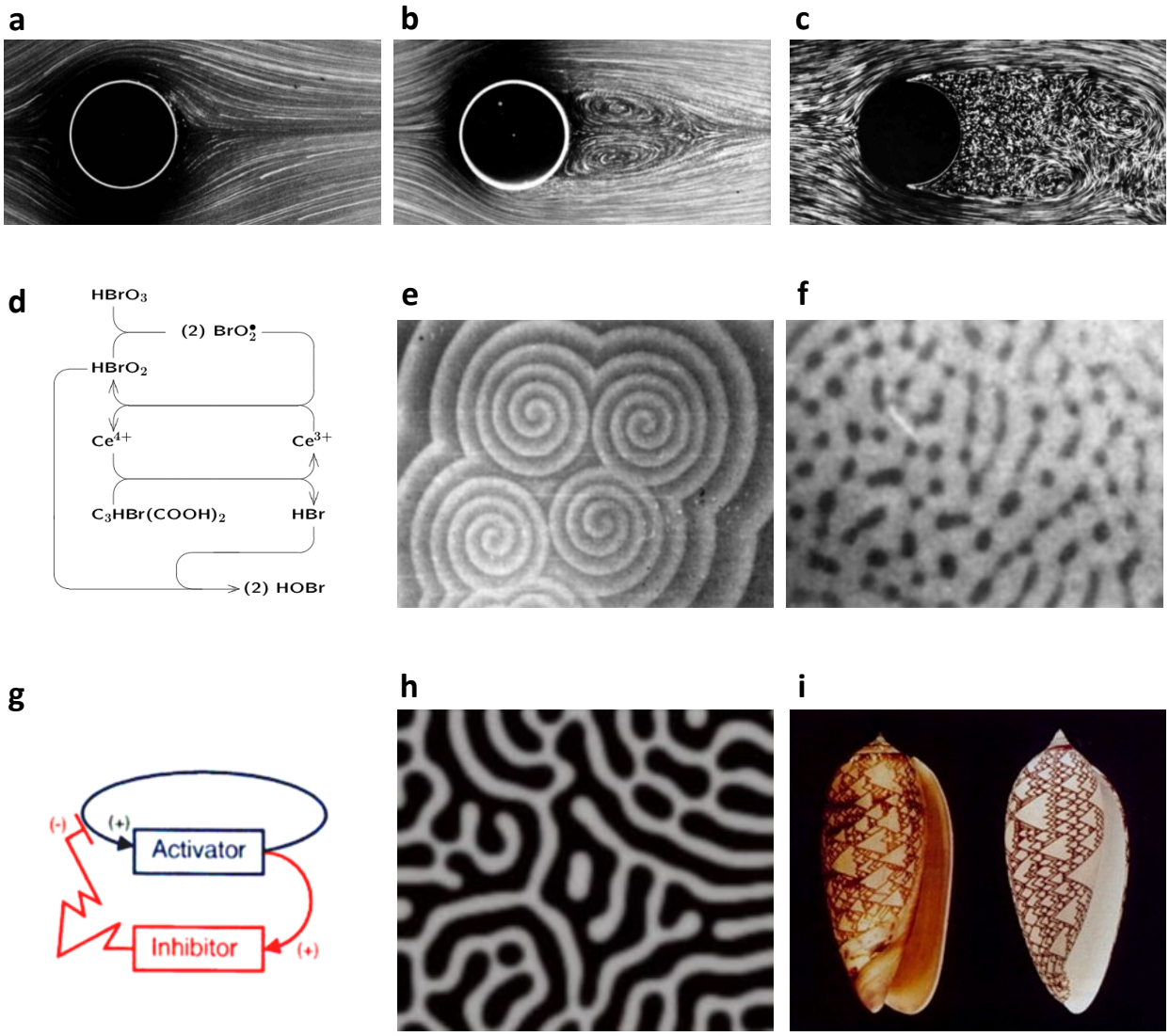

Figure 1.1: Complex systems and pattern generation from physics to chemistry to biology.

(a-c) Photographs showing viscous fluid flow around a cylindrical obstacle, resulting in ordered flow at low viscosity (a) and chaotic flow at high viscosity (c). If the viscosity is tuned just right, complex flow (b) emerges. Images reproduced with minor modifications (cropping, contrast enhancement) from [Van Dyke, 1982]. (d-e) Belousov-Zhabotinsky reactions implement a simple chemical oscillator based on the oxidation of cerous ions (Ce^{3+}) that feeds back to inhibit itself through the delayed production of bromide ions [Zhabotinski, 1991]. (d) shows the core scheme of the reaction, reproduced from [Zhabotinski, 2007]. (e) and (f) show patterns that emerge when the reaction takes place on a flat surface, reproduced with minor modifications (cropping, contrast enhancement) from [Zhabotinski & Zaikin, 1973]. (g-i) Turing/Gierer-Meinhardt models show that the simple activator-inhibitor system in (g) is sufficient to generate a wide range of patterns, such as the simulated stripes in (h) or the real (left) and simulated (right) sea shell patterns in (i). Reproduced with minor modifications (cropping, contrast enhancement) from (g) [Meinhardt, 2006], (h) [Kondo & Miura, 2010], and (i) [Meinhardt, 1998].

1.2.2 Biological Relevance

There are two main arguments for why complexity is relevant to biology. First, complexity tends to emerge in systems that feature non-linear interactions and both negative and positive feedback [Manson, 2001; Ross & Arkin, 2009], all of which are abundant across biological systems. Second, the behaviors spontaneously generated by complex systems are reminiscent of phenomena observed in biology, including symmetry breaking and pattern formation.

A rudimentary example of this can be found in chemistry in the form of Belousov-Zhabotinsky reactions [Belousov, 1959 (ru); Zhabotinsky, 1964 (ru); Zhabotinsky, 1991 (en)]. Through a relatively simple set of non-linear feedback loops (Fig. 1.1d), these purely chemical systems generate oscillations in time and – when spread out on a flat surface – intricate patterns in space (Fig. 1.1e-f) [Zhabotinski & Zaikin, 1971].

Similar feedback-based systems have been proposed as mechanisms of biological pattern formation. For instance, Turing already demonstrated in 1952 that patterns can be generated through the reaction of two substances that diffuse at different rates [Turing, 1952], a mechanism that was later generalized by Gierer and Meinhardt into a model that combines short-range self-activation and long-range self-inhibition (Fig. 1.1g) [Gierer & Meinhardt, 1972]. Turing/Gierer-Meinhardt models have since been demonstrated to be capable of generating a wide range of biologically relevant patterns in simulations [Meinhardt & Gierer, 2000; Kondo & Miura, 2010], including dots and stripes (Fig. 1.1h), and have been used to explain a number of real biological patterning events, including the formation of the vertebrate limb [Newman et al., 2018], the distribution of bird feather buds [Jung et al., 1998], and sea shell pigmentation (Fig. 1.1i) [Meinhardt, 1998]. More recently, they have also been further generalized to show that the inclusion of additional signaling factors (to a total of three or four interacting species) allows Turing patterns to emerge more robustly [Marcon et al., 2016] and that network topology can be used to predict and engineer key properties of such systems [Diego et al., 2018].

Reaction-diffusion-based patterning is only one example of how complex behaviors can underpin biological phenomena. Given the prevalence of feedback loops and non-linearities in other biological subsystems, including intracellular signaling, gene-regulatory networks and the cytoskeleton, as well as in combinations of all of the above (see section 1.4), there is little doubt that complexity is a pervasive and important property of biological systems.

In summary, the topology of biological systems implies that many of them are capable of exhibiting complex behavior and the capacity of complex systems for spontaneous pattern generation suggests that biology may widely use them as a means of self-organization.

1.2.3 The Methodological Consequence

Because variations in individual components can have a disproportionate effect on overall system behavior, complex systems are notoriously sensitive to initial conditions (although not as strikingly so as chaotic systems) and may exhibit a behavior known as a *cascade failure* [Manson, 2001; Buldyrev et al., 2010]. In a cascade failure, a perturbation of a single component propagates through the system and changes its state drastically, often completely breaking the dynamics observed in the unperturbed state.

Much of classical developmental biology is built on perturbation experiments such as gene knock-outs, which are often performed at large scale in the form of genetic screens on model organisms, an approach popularized by pioneering successes such as the screens of *C. elegans* movement [Brenner, 1974] and of *Drosophila* embryonic patterning [Nüsslein-Volhard & Wieschaus, 1980]. Classically, the differences between the wild-type and the resulting phenotype are taken as a starting point for inferring the function of the perturbed gene. For instance, a gene whose knock-out leads to the arrest of an otherwise migrating tissue might be considered to function in motility or guidance.

However, if biological systems are in fact complex systems, it is expected that genetic perturbations would frequently lead to cascade failures. In such cases, the difference between a perturbation phenotype and the wild-type will usually be uninformative and sometimes even misleading for attempts to infer the function of the perturbed gene [Welf & Danuser, 2014]. In the simple example mentioned above, the perturbed gene could just as easily be involved in proliferation or cell differentiation, which when perturbed could have knock-on effects that disrupt migration. This issue can be understood as *the methodological consequence (of complexity)* and will be referred to as such throughout this thesis.

Given the methodological consequence, it seems surprising that the genetic perturbation approach has historically proven incredibly powerful; indeed, it constitutes one of the main sources of most of the field's established knowledge. In part, this is because genetic screens are very effective as a means of finding the set of genes involved in a given system, a type of conclusion that unlike functional inference is not subject to the methodological consequence. In addition, the apparent success of genetics may also partially be explained by a positive selection bias, both in terms of which genes are studied (predominantly those that happen to yield interesting phenotypes in a screen) and which results are published (almost exclusively those where some sort of insight could be gained from the analysis).

In order to progress beyond such cases, it will be necessary to complement classical loss-of-function genetic studies with novel approaches more suitable for the dissection of complex biological systems. Fortunately, several such new methodologies – including those presented in this thesis – have recently been developed or are currently in development (see section 1.3). These new approaches

harbor the potential for biologists to investigate complex biological systems in an integrative rather than a reductive way, which will hopefully lead to more general and transferable biological models and to an increase in their explanatory and predictive power.

1.3 Novel Approaches for the Study of Complex Biological Systems

At present there are no simple solutions that fully overcome the methodological consequence of complexity. However, the rapid technological progress of the past decade together with new experimental and algorithmic tools has led to the advent of a set of methodologies that (in tandem with classical genetic and molecular work) can be expected to propel the biology of complex systems forward.

These novel approaches can be summarized in three categories: *precision perturbation*, *synthetic developmental biology*, and *data-driven biology*. All three are briefly introduced in this section and two of them – precision perturbation and data-driven biology – are at the core of much of the work presented in this thesis. Common to all of them is that they focus on native system states rather than perturbed states and thereby avoid the problem of cascade failure.

1.3.1 Precision Perturbation and Optogenetics

Although complex systems may exhibit cascade failure when perturbed, they are often robust within a limited domain (unlike chaotic systems), which has been called the *region of linear biology* [Welf & Danuser, 2014]. In this domain, a perturbation of a certain magnitude will have consequences of a similar magnitude, which can be expected to be caused through the same chain of reactions that mediates the natural behavior of the system. Therefore, tools for precise and subtle perturbation experiments (alongside tools for measurements and analyses of matching sensitivity) promise a way to overcome the methodological consequence by allowing complex biological systems to be studied within their linear region.

Perturbations can be made more precise by reducing their magnitude, confining them spatially or confining them temporally. Spatial confinement is especially interesting when the effects of a local perturbation can be observed as they propagate through the surrounding unperturbed system and temporal confinement is particularly useful when normal conditions can be maintained right until the experiment begins, which excludes artifacts resulting from chains of knock-on effects that begin earlier on in development and are unobserved.

Historically, many different strategies have been employed to achieve higher precision compared to complete genetic knock-outs. These include knock-downs where some residual gene product may persist and hypomorphic alleles and titration of drug treatments to reduce the magnitude of the perturbation. Furthermore, tissue-specific knock-downs, knock-outs or overexpression can confer spatial confinement, and temperature-sensitive alleles or heat-/drug-inducible promoters allow temporal confinement. All of these methods have been successful to some extent and often yielded new insights into biological systems.

The next step is to take precision perturbation to the single-cell level and below on the spatial scale and to the level of minutes or seconds on the temporal scale. Laser ablation was the first routine method to achieve this by means of light-based sample manipulation [Amy & Storb, 1965; Gayathri Vegesna et al., 2017]. It is, however, relatively crude and limited in terms of the specific nature of its interaction with the sample. Caged compounds and photolabile drugs that can be locally activated or degraded using light provide more specific biochemical activities but are only sparsely available and must first be delivered to the tissue of interest [Ellis-Davies, 2007]. All of these drawbacks now stand to be overcome by a recent addition to the precision perturbation toolbox: optogenetics.

Optogenetics combines the precision of light with the specificity of biochemical interactions [Repina et al., 2017; Guglielmi et al., 2016], which sets them up to be the ultimate precision perturbation tool. The recent burst of progress in the field began with the use of an algal channelrhodopsin that functions as a light-activated ion channel to optically control neuronal activity [Boyden et al., 2005]. Subsequently, an ever-growing set of naturally occurring light-responsive proteins or protein domains was adapted for optogenetic purposes and optimized to be sensitive to different wavelengths and to have different response kinetics [Tischer & Weiner, 2014; Karunarathne et al., 2015; Zhang & Cui, 2015]. Different approaches have been pursued for coupling light-dependent conformational change to desirable biochemical outputs, including chimeric fusions of different light-sensitive domains and signaling domains of receptors [Kim et al., 2005], releasable steric blocking of active sites [Wu et al., 2009], and recruitment of effectors from the cytoplasm to specific sites of activity by light-induced heterodimerization [Kennedy et al., 2010].

These tools have already been applied to control many cellular processes [Tischer & Weiner, 2014], including gene expression and epigenetics [Konermann et al., 2013], a wide range of signaling cascades [Karunarathne et al., 2015; Zhang & Cui, 2015], organelle transport and positioning [van Bergeijk et al., 2015], and cell contractility and mechanotransduction [Valon et al., 2017]. Now, the use of optogenetic tools is being expanded into multi-cellular tissues and developmental systems [Johnson & Toettcher, 2018], as highlighted in a recent study that brings a complete morphogenetic process under optogenetic control, namely epithelial invagination by means of apical constriction in the *Drosophila* blastoderm [Izquierdo et al., 2018]. A similar goal was also pursued as part of this thesis (see section 1.6.1).

The field of optogenetics is growing and maturing rapidly. Combined with careful experimental design and analysis, it will allow biological systems to be studied within their region of linear behavior and will thus enable significant advancements toward overcoming the methodological consequence of complexity.

1.3.2 Synthetic Developmental Biology

There are many more ways to break a complex system than there are to build one. More specifically, it is unlikely that there are many different ways of constructing a synthetic complex system that behaves in the same way as a natural complex system, especially if it is syntesized from largely the same components. It follows that any success at such an endeavor gives weight to the hypothesis that the natural system functions based on the same principles that guided the construction of the synthetic one. Thus, synthesis constitutes a means for testing hypotheses that are hard to test using perturbation due to the misleading effects of cascade failure.

In addition, synthetic approaches may allow a disentanglement of the complicated and complex aspects of a system. Multi-cellular systems tend to be both complex and complicated, with the latter making the study of the former much harder. Successfully reproducing key aspects of a system's behavior with only a subset of the naturally present parts and interactions indicates that this subset lies at the core of the system's behavior, whilst additional components may only have auxiliary functions (such as coupling the core system to upstream inputs and downstream outputs). In other words, synthetic approaches may demonstrate the *sufficiency* rather than just the *necessity* of a subsystem. Once such a sufficient subsystem is established, it is easier to study than its more complicated natural counterpart, even if its behavior remains complex.

Synthetic approaches such as reconstitution (that is the assembly of functional systems from existing biological building blocks) have long been used with great success in molecular biology [Kron & Spudich, 1986] and are increasingly employed in cell biology [Liu & Fletcher, 2009] but technical limitations have made it difficult to extend them to multi-cellular systems. The sheer number of components involved across multiple scales presents a massive challenge for any attempt to distill and synthesize a working synthetic model. Nevertheless, progress has been made toward this goal along two promising avenues – mathematical modeling and tissue engineering – and further steps are on the horizon.

Mathematical Modeling & Simulation

First-principle mathematical modeling and computational simulation is a powerful tool for *in silico* synthetic biology. Freed from the constraints of real-life biological engineering, synthetic systems can be created and iteratively improved upon in a fast and efficient fashion. Parameters can then be screened or fitted based on data from the natural system in order to investigate or demonstrate how well the model approximates reality and in order to make new predictions [Brodland, 2015].

Whilst computational models usually require drastic simplification, which is a potential downside, the mathematical language by which models are formulated also forces assumptions to be made explicit, which can identify researchers' biases and blind spots, i.e. aspects of the biological system that have not yet been investigated sufficiently [Brodland, 2015]. Furthermore, if a mathematical model is

capable of reproducing non-trivial behaviors observed in a natural system, it follows that the model is in fact sufficiently expressive to describe the core of the system. In this case, the simplified nature of the model turns from a weakness into a strength.

A question that is often left open is how to determine whether a model is actually reproducing *non-trivial* system behaviors. If a model with sufficiently many parameters is fit to data, its capability to reproduce the data in question is always trivial (this is also known as *overfitting*). By now, however, there exist a host of basic and advanced techniques to counter this issue, including investigation of the model's parameter space [Gramacy et al., 2004], information-theoretical criteria for model selection [Konishi & Kitagawa, 1996], and efficient computational parameter estimation and model selection using novel methods such as Approximate Bayesian Computation [Toni et al., 2009; Toni & Stumpf, 2010]. Furthermore, the best way of testing a computational model is the same as for any scientific model: by making novel predictions and checking experimentally whether they hold true, for instance by means of precision perturbation of the system under study (see section 1.3.1).

One particularly promising aspect of mathematical models is the ease with which their properties can be explored once they are established. Initial conditions, boundary conditions, parameters and model topology can be varied and the outcomes simulated. This makes it possible to chart the many different behaviors that even a relatively simple complex system can produce, providing an entry point for the discovery of general core principles that may be reused to produce very different outcomes in different contexts [Brodland, 2015].

A powerful example of the successful application of such techniques are the Turing/Gierer-Meinhardt systems discussed in section 1.2.2, which were both discovered and later massively extended by means of mathematical modeling [Kondo & Miura, 2010], including comprehensive model space exploration [Marcon et al., 2016].

Tissue Engineering

Experimentally synthesizing biological systems is harder than mathematical modeling but benefits from its inclusion of actual biological context, which imposes real-world constraints on the model and therefore naturally counters issues such as oversimplification and overfitting [Davies, 2017].

Presently, synthesis of tissue-scale biological systems from scratch by means of chemical and physical processes is still impossible to realize. Instead, early successes have come through approaches that use pre-existing biological entities (such as cultured cells) as a chassis within which genetically engineered systems can be implemented. Using genetic logic like transactivation and regression, systems such as biological oscillators and gradient-based pattern interpreters have been synthesized and each such success has contributed to an enhanced understanding of the principles underlying these systems [Purnick & Weiss, 2009; Davies, 2017].

Although some of this work has been aimed at synthetic morphogenesis [Teague et al, 2016], most of it so far was based on 2D cell cultures or on cells in solution [Davies, 2017]. Given recent successes in the field of organoid culture [McCauley & Wells, 2017; Artegiani & Clevers, 2018], this is liable to change in the near future. Currently, the organoid field is mostly focused on finding conditions that sufficiently mimic the *in vivo* environment to allow development to proceed quasi-normally, which does not strictly constitute synthesis. Nevertheless, because of their experimental accessibility and controlled environment, organoids will no doubt present an excellent chassis for advanced synthetic work.

One particularly interesting avenue is to engineer and control specifically the environment of cells and organoids. By varying initial and boundary conditions in a precise and controlled fashion and then observing the behavior of cells and cell collectives in these different contexts, the adaptive capacity of a biological system can be interrogated, which – akin to mathematical models – allows the delineation of core principles from which a wide range of context-specific behaviors can emerge. Interesting examples of such work include the generation of patterned adhesive surfaces to constrain and shape cells [Singhvi et al., 1994], the use of hydrogels to engineer microenvironments [Gong & Mills, 2018], and the controlled delivery of signaling molecules or drugs through microfluidics [Occhetta et al., 2015].

The synthetic approach is not limited to *in vitro* studies; as more precise and less invasive tools for the manipulation of biological systems become available (see section 1.3.1), it becomes feasible to employ the same approach in the context of living organisms. This presents another step up in terms of technical challenge but also closes the gap between synthetic and natural systems, allowing hypotheses about the latter to be tested by bringing them under the control of the former.

1.3.3 Data-Driven Biology and Inference Without Perturbation

Perturbation experiments are widely considered necessary in order to demonstrate causal relationships instead of 'mere' correlations. However, given the difficulty of interpreting perturbation experiments performed on complex systems, approaches to test hypotheses or infer causal models without resorting to perturbation are likely to be extremely valuable [Vilela & Danuser, 2011; Welf & Danuser, 2014].

The ever increasing sensitivity and scale at which biological data can be acquired, coupled with the rapid growth of computational power and data analysis tools over the past decade, are opening up several new avenues to pursue in this direction, all of which rely on *natural variation* within either a population of samples or a stream of time points, as opposed to inducing variation exogenously by means of perturbation.

Exploratory Machine Learning

One such avenue is the detection of non-random structures within large amounts of data, i.e. *big data*. This is the domain of *machine learning*, which has seen rapid progress over the past seven or so years and is a perfect match for the rapidly growing *omics* approach in biology [Libbrecht & Noble; Camacho et al., 2018; Ching et al., 2018].

When large datasets with hundreds of features and thousands of samples are collected – for instance by means of single-cell transcriptomics – the distribution of samples within the feature space is expected to be highly structured rather than random. Machine learning approaches, in particular *unsupervised learning*, present a powerful means of detecting and exploring this structure. Thus, information can be retrieved from a single large dataset describing the wild-type population rather than from a comparison of specific features between two conditions such as a wild-type and a perturbation [Haghverdi et al., 2015; Buettner et al., 2015; Angerer et al., 2017; Deng et al., 2018].

This approach has the potential to become a means for hypothesis generation that does not rely on large-scale perturbation screening but instead detects components or dynamics of interest based on their distribution in a wild-type population.

Generative Models

Another avenue is *data-driven generative modeling*. In addition to the synthetic first-principle approach to mathematical modeling discussed in section 1.3.2, machine learning allows models to be built in a mechanism-agnostic, data-driven fashion.

Currently, this is accomplished by taking general-purpose models, such as neural networks, and training them to reproduce the structure observed within a large dataset, thus learning a generative model of the data [Salakhutdinov, 2015]. Whilst useful for technical applications such as image segmentation [Badrinarayanan et al., 2016] and dimensionality reduction [Way & Greene, 2018], this methodology is limited due to a substantial drawback common to present-day neural network approaches, namely that the resulting models are usually not interpretable [Lipton, 2017; Samek et al., 2017; Doshi-Velez & Kim, 2017]. In essence, the final model is simply a complicated mathematical function that generates data looking similar to the training data. This function does not reflect anything about the real-world system under study in an interpretable way. Fortunately, this problem is well-known in the machine learning field and much work is being done to address it [Lundberg & Lee, 2017; Alverez-Melis & Jaakkola, 2018].

A particularly promising prospect might lie at the interface of first-principle modeling and data-driven modeling: given a set of pre-established constraints, components and interactions, a data-driven inference engine builds and evaluates first-principle models that – once a good fit is found – are readily interpretable for researchers. Hybrid first-principle/neural-network models are also a

promising option to explore, as demonstrated recently by the development of an impressively successful hybrid algorithm for automated chemical retrosynthesis [Segler et al., 2018].

Causal Inference from Basal Fluctuations

A third and highly promising avenue is the possibility of *inferring causal relationships* from time course measurements of *basal fluctuations* [Vilela & Danuser, 2011; Welf & Danuser, 2014]. Components of biological systems are always subject to stochastic variation over time, introduced at the biochemical level by thermal fluctuations (often exacerbated by low abundances) [Averbukh et al., 2018] and at larger scales by supervening stochastic effects such as transcriptional bursting [Raj & Van Oudenaarden, 2008]. If these fluctuations can be measured precisely enough across multiple components of a system, they can be exploited to track causal relationships within the system by determining how strongly fluctuations in one component are transferred to another. In other words, natural fluctuations can be exploited as micro-perturbations of the system. This avoids the methodological consequence because such fluctuations are by definition within the linear region of system behavior – otherwise they would frequently cause cascade failures within wild-type systems.

The actual mathematical implementation of such an inference approach is non-trivial, but fortunately there already exists a good starting point in the field of mathematical economics, where the notion of *Granger causality* was developed for exactly this purpose [Granger, 1969; Vilela & Danuser, 2011]. Granger causality holds that a time-dependent variable X is causally upstream of another time-dependent variable Y if the past dynamics of X are informative for predicting the future of Y better than what would be possible from the past dynamics of Y alone. This is based on the two intuitions that a cause must occur before its effect and that a cause must have unique information about its effect [Granger, 1969; Eichler, 2012]. Early implementations of Granger causality were limited to linear models with a single dependent variable, non-parametric and multivariate generalizations have been developed since [Dhamala et al., 2008; Barret et al., 2010].

Although Granger causality has become a popular tool in neurobiology for the analysis of information flow in the brain [Friston et al., 2013], it has so far not been widely adopted for biological systems, likely due to the challenging mathematics and perhaps due to a lack of sufficiently high-quality time series data. However, recent use cases have begun to appear, for instance in causal network inference from gene expression data [Finkle et al., 2018] and from high-throughput imaging data [Lock et al., 2014]. Furthermore, it is known that a generalized implementation of Granger causality capable of coping with the feedback-dense topology of complex biological systems is currently being developed in the lab of Gaudenz Danuser [Welf & Danuser, 2014; and unpublished data].

Data Science for Biology

All of the approaches described above fall within the scope of the young field of *data science*, which utilizes computational tools to study systems in ways that go beyond what is possible with human reasoning alone. Data science is at the heart of a powerful trend toward the use of data-driven methods and machine learning across virtually all fields of science [see e.g. Chen et al., 2018; Radovic et al., 2018; Ray et al., 2018; Segler et al, 2018] and – even more notably – across much of the private sector [see e.g. Parloff, 2016; Lewis-Kraus, 2016; Evans, 2018]. The rapid spread of useful inventions across all of these fields also promises rapid advancements for data-driven biology.

However, there are also pitfalls and limitations associated with this trend, which should be given due consideration.

For instance, data science in some ways reflects a departure from the hypothetico-deductive model of classical science[1]. It still uses empirical data and still produces predictive models, but does so in a data-driven rather than a hypothesis-driven way. This is not a problem *per se* – it can and should be seen as an enrichment of classical science – but it may become a problem if classical science is replaced entirely by data-driven methods.

Consider for example that machine learning currently produces models that are predictive but not explanatory, due to the lack of interpretability described above [Lipton, 2017]. Unlike classical scientific models, such black box models cannot be checked for consistency with independent facts and accepted theories. They also cannot be generalized, extended or transferred to other problems by means of human reasoning. Last but not least, whilst purely predictive models may be useful in many scenarios of applied science, they do not satisfy human curiosity and do not produce human understanding, which is why on their own they cannot substitute the contribution of classical science to the formation of an enlightened society. It is therefore critical to ground discoveries from data-driven biology within accompanying mechanistic work.

A more practical limitation that is highly relevant for biology is the need for big data in order for most modern machine learning approaches (in particular *deep learning*) to perform well [Marcus, 2018]. Whilst the omics technologies can readily produce such big data, they usually do not capture biological context (such as the original distribution of cells within a tissue) and they usually require the fixation or destruction of the sample itself, preventing the observation of live dynamics. These limitations make it hard to study complex biological systems, even if cascade failures are eliminated as a source of error. On the other hand, methods that readily produce context-rich and *in vivo* data, such as fluorescence microscopy, are hard to scale up for measuring hundreds of components in thousands of samples, and the resulting data is difficult to cast into computational representations that are readily usable for data-driven investigation.

[1] To what extent the hypothetico-deductive model describes classical scientific practice accurately is admittedly debatable in the first place, but that debate is beyond the scope of this thesis.

In short, there is currently a dichotomy between *big data* and *rich data* in biology, which represents an interesting challenge for the field and is bound to spur new experimental and computational developments.

On the omics side, progress toward more rich datasets is being made by combining different omics approaches (*multi-omics*) [Huang et al., 2017; Argelaguet et al., 2018] and by developing novel high-throughput methods that work *in situ*, one example being imaging mass cytometry [Giesen et al., 2014]. On the microscopy side, the throughput of acquisition can be scaled up by automated microscopy, including cases where the data itself informs the next acquisition (called *feedback microscopy* or *smart microscopy*) [Tischer et al., 2014], e.g. to track moving cells [Rabut & Ellenberg, 2004] or to identify and follow interesting spontaneous events such as cell divisions [Conrad et al., 2011]. The limited number of channels in fluorescence microscopy and thus the limited number of components that are simultaneously observable has also been tackled, for instance by cycling through multiple stainings on the same sample with intermittent washing or bleaching steps [Lin et al., 2015; Gut et al., 2018].

On the computational side, there are some ongoing efforts to improve the analysis of small-scale but context-rich data, although the field is currently mainly focused on big data. The important goal of overcoming the limited multiplexing capabilities of fluorescence microscopy can also be addressed computationally, namely by means *atlas mapping*. Much like different geopolitical and geological features can be overlaid on a geographical reference map, different measurements acquired during different experiments can be mapped onto each other based on a common reference measurement.

Recently, this approach has been used to merge time courses of mitosis of 28 independently imaged endogenously tagged proteins, creating a dynamic protein atlas of cell division [Cai et al., 2018]. The use of generative machine learning models to create a similar atlas of adherent human induced pluripotent stem cells has also been reported recently [Johnson et al., 2018]. Atlas approaches are not limited to the cellular scale, as illustrated by a whole-embryo gene expression atlas of *Platynereis dumerilii* generated by whole-mount *in situ* hybridization and subsequent registration [Vergara et al., 2017]. Even dynamic subcellular processes can be studied by atlas mapping, demonstrated impressively in a study that employs highly time-resolved temporal registration of lamellipodial proteins based on local movement of the lamellipodial front to extract a temporal hierarchy of the actin regulatory cascade [Lee et al., 2015]. Finally, atlas methods can be used to combine big data and rich data approaches, for instance by mapping single-cell expression data back into the embryo through reference measurements such as *in situ* hybridization data [Satija et al., 2015; Achim et al., 2015].

Taken together, the computational methods outlined in this section present a massive opportunity for the future of biological research in general and for the study of complex biological systems in

particular – provided that potential pitfalls are avoided and that current challenges can be overcome, which is among the main goals of the work described in this thesis.

1.4 The Interplay of Patterning, Morphogenesis and Migration

As introduced in section 1.1, cells and cell collectives perform a range of impressive feats of self-organization in the context of the developing embryo. In order to build a functioning adult organism from a single-celled zygote, cells must grow and divide (proliferation), adopt specific spatial configurations (patterning), change their molecular composition (differentiation) and subcellular architecture (e.g. acquisition of polarity), change their shape and generate three-dimensional tissue structures (morphogenesis), change their position within organs or within the embryo entire (migration), and sometimes self-destruct in a controlled manner (apoptosis).

Although there are still open questions about the molecular basis of each of these phenomena, much is already known about the components and mechanisms involved. However, developmental processes do not occur in isolation; for fully functional organs and organisms to form, they must be combined in myriad ways and coordinated in space and time over multiple orders of magnitude, giving rise to sophisticated and robust developmental programs. Relatively little is known, comparatively speaking, about this integration and the resulting emergent behaviors.

This section provides a brief overview of the established knowledge in the field, in particular the multi-cellular phenomena of *patterning*, *morphogenesis* and *collective migration*, the integration of which is the main subject under study in this thesis.

1.4.1 Patterning

Patterning is the generation of non-random distributions of cellular features within tissues or organisms. Any cellular feature can be patterned, including gene expression, biochemical activities, polarity and subcellular organization, cell shape, and mechanical properties. Usually, multiple such features are patterned alongside each other and often the pattern of one induces or reinforces the pattern of another. Patterns can be generated in space or time or both, and again one may lead to the other, a famous example of this being somite segmentation in vertebrates [Oates et al., 2012]. Patterning can be transient or made permanent by coupling it to differentiation through expression of transcription factors that prevent reversibility or through epigenetic modifications.

Most molecular components found to be relevant for patterning are embedded in signaling or signal integration modules. This includes diffusible signals (morphogens) and their corresponding receptors, most of which are G-protein coupled receptors (GPCRs) or receptor tyrosine kinases (RTKs) [Alberts et al., 2002]. Also included are modules for direct contact signaling such as the Notch-Delta module [Artavanis-Tsakonas et al., 1999] and – discovered more recently – modules for mechanosensing such as the cadherin-catenin complex [Buckley et al., 2014]. In addition to sensing and receiving signals, cells must process and integrate the information they receive, which happens both through intracellular signaling cascades involving for instance MAPK, PI3K or Rho GTPases [McKay and

Morrison, 2007] and through changes in gene expression that are parsed through gene regulatory networks [Levine & Davidson, 2005]. Finally, there are a plethora of additional factors allowing modulation and fine-tuning of patterns, including regulators of receptor sensitivity (e.g. β-arrestins) [Reiter & Lefkowitz, 2006] and of morphogen diffusion (e.g. heparan sulfate glycans; HSPGs) [Yan & Lin, 2009].

A series of commonly occurring higher-order mechanisms or processes, here referred to as developmental *motifs*, have been discovered that make use of the aforementioned molecular modules to implement patterning events.

One of the most fundamental motifs is that of morphogen gradients (Fig. 1.2a) [Rogers & Schier, 2011]. Local production and subsequent diffusion of a morphogen produces a concentration gradient, which can be maintained and shaped through sink activity and modulation of diffusion [Wartlick et al., 2009]. Cells can measure the local levels and thereby gain information about their location relative to the source, which informs their cellular response [Ashe & Briscoe, 2006]. By integrating multiple gradients from different sources, tissues can be patterned into exceedingly precise and complicated domain structures [Briscoe & Small, 2015], as seen for instance in the well-known case of the continuously refined body axis patterning at the blastoderm stage of *Drosophila* [Jaeger et al., 2012].

Another common motif is lateral inhibition (Fig. 1.2b), which allows the generation of salt-and-pepper patterns as well as local pattern refinement [Sjöqvist & Andersson, 2017]. In lateral inhibition, cells moving along a particular developmental trajectory signal to their neighbors to prevent them from following the same trajectory, which includes preventing them from producing the lateral inhibition factor themselves. In a field of roughly uniform cells, those expressing a lateral inhibition factor at a slightly higher level than others will inhibit its expression in their neighbors, which – usually in combination with an additional positive feedback – leads to the emergence of a salt-and-pepper pattern [Turing, 1952], as seen for instance in *Drosophila* neurogenesis [Cabrera, 1990]. Alternatively, cells can express a lateral inhibition factor in response to a signaling cue such as a morphogen or chemokine gradient. In this case, the cell receiving the highest amount of inductive signal will suppress the others [Sjöqvist and Andersson, 2017], a mechanism used to refine shallow morphogen and chemokine patterns, as seen in *C. elegans* vulval development [Sternberg & Horvitz, 1989] and in angiogenic sprouting [Phng & Gerhardt, 2009].

At the next higher level, motifs such as gradient sensing and lateral inhibition are combined to yield complex pattern generators that can create a near-infinite variety of different patterns. One such generator is the Turing/Gierer-Meinhardt model discussed in section 1.2.2, which combines short-range self-activation with long-range self-inhibition and is thought to underlie patterning events ranging from zebrafish pigment stripe formation [Yamaguchi et al., 2007] to Hydra body plan organization [Meinhardt, 1993]. Depending on the parameters, boundary conditions and initial

conditions, this combination of feedback loops can lead to very different outcomes [Kondo & Miura, 2010].

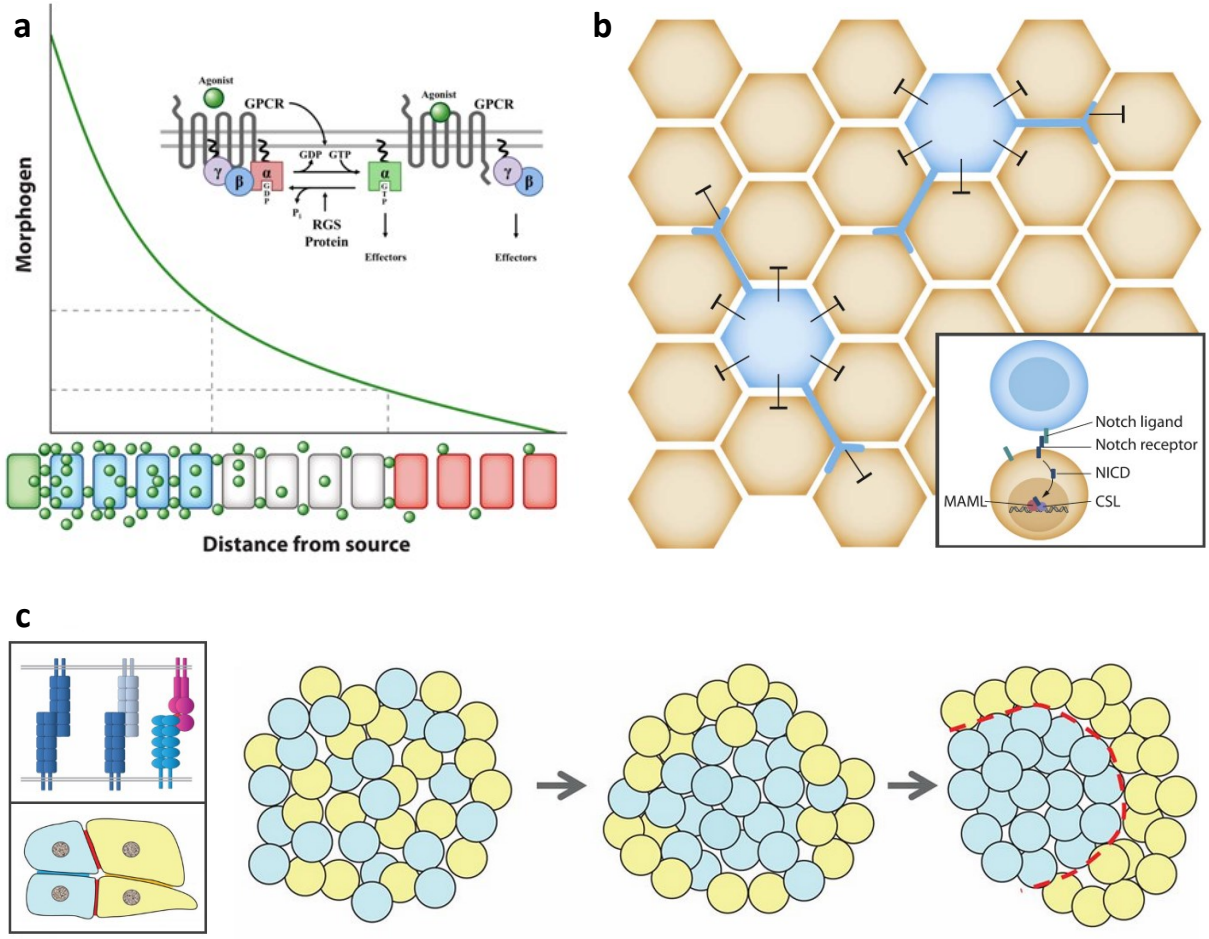

Figure 1.2: Common modules and motifs that drive patterning events.

(a) Diffusible signaling and sensing modules such as morphogens (green spheres) and GPCRs (see inset) allow cells to communicate over long distances. Morphogens diffusing away from a source (green cell on the left) can form a gradient that encodes spatial information. This in turn informs the behavior of receiving cells which may for instance differentiate differently (blue, white and red cells) depending on whether they sense a morphogen concentration above certain thresholds (dashed gray lines). Inset adapted from [Stewart et al., 2012], main figure reproduced from [Rogers & Schier, 2011]. (b) Cell-cell contact signaling modules such as Notch signaling (inset) are employed to mediate lateral inhibition, which can generate and refine patterns locally. For instance, cells expressing a ligand on their surface (blue cells) can inhibit surrounding cells (brown cells), generating a salt-and-pepper pattern. The distance of inhibition can be extended beyond immediate neighbors through filopodial protrusions. Figure adapted from [Sjöqvist & Andersson, 2017]. (c) Homophilic and heterophilic cell-cell adhesion molecules (upper left inset) mediate differential adhesion between cells (lower left inset) through homotypic (blue and yellow junctions) and heterotypic (red junctions) adhesion. A mixed aggregate of cells differentially expressing such adhesion molecules can cluster and ultimately separate through a passive physical unmixing process (left to right series). Figure adapted from [Fagotto, 2014].

Not all patterning is based on cell-cell communication. One important exception is the differential adhesion hypothesis (DAH), which was first brought forward to explain homotypic cell sorting in embryonic amphibian cells [Townes & Holtfreter, 1955] by proposing that mixed populations of cells favoring homotypic over heterotypic adhesion would naturally segregate as a consequence of physical energy minimization (Fig. 1.2c) [Steinberg, 1970]. DAH-based phenomena have since been found to be involved in a range of developmental processes [Steinberg, 2007]. In more recent work, the DAH has been generalized by reformulating it in terms of interface tension rather than adhesion, producing the differential interface tension hypothesis (DITH) [Brodland, 2002], which also explains patterning phenomena emerging from differential cell surface tension, for example the sorting of high-tension cells to the inside of the embryo during mouse blastocyst formation [Maître et al., 2016].

1.4.2 Morphogenesis

Morphogenesis is the adoption of specific geometric configurations by cells, tissues and organisms. Most morphogenetic processes involve shape changes of individual cells, which through collective action in a tissue compound into larger-scale deformations such as curving, folding, elongation or compaction. For such collective transformations to proceed normally, tight spatial and temporal coordination of the individual cells' behaviors is required.

Because morphogenesis requires physical forces to be generated and transmitted, the molecular module predominantly responsible for morphogenetic processes is the actomyosin cytoskeleton [Munjal & Lecuit, 2014]. Cortical actomyosin mediates the mechanical stability and dynamics of cell shapes and – through cross-cellular coupling via adhesion molecules – also maintains supracellular mechanical integrity within tissues [Lecuit & Lenne, 2007]. The forces generated by actin polymerization and/or by motor proteins such as non-muscle myosin II (MyoII) allow actin to form extensile or contractile networks [Belmonte et al., 2017], respectively mediating the outgrowth of cell membrane domains such as lamellipodia and filopodia [Ridley, 2011], and the contraction of entire cells or of specific subcellular sections such as the apical domain [Salbreux et al., 2012; Martin & Goldstein, 2014]. The actomyosin module is controlled by an extensive set of regulators [Zaidel-Bar et al., 2015], most notably the small GTPases RhoA, Rac and Cdc42, which themselves are part of a complex regulatory network [Van Aelst & D'Souza-Schoray, 1997; De Curtis & Meldolesi, 2012]. In addition to the actomyosin module, multiple other modules with the capacity of modifying physical properties or behaviors of the cell can be involved in morphogenetic events, including microtubules [Cearns et al., 2016], adhesion molecules [Niessen et al., 2011], ECM components [Rozario & DeSimone, 2010], and intermediate filaments [Sanghvi-Shah & Weber, 2017].

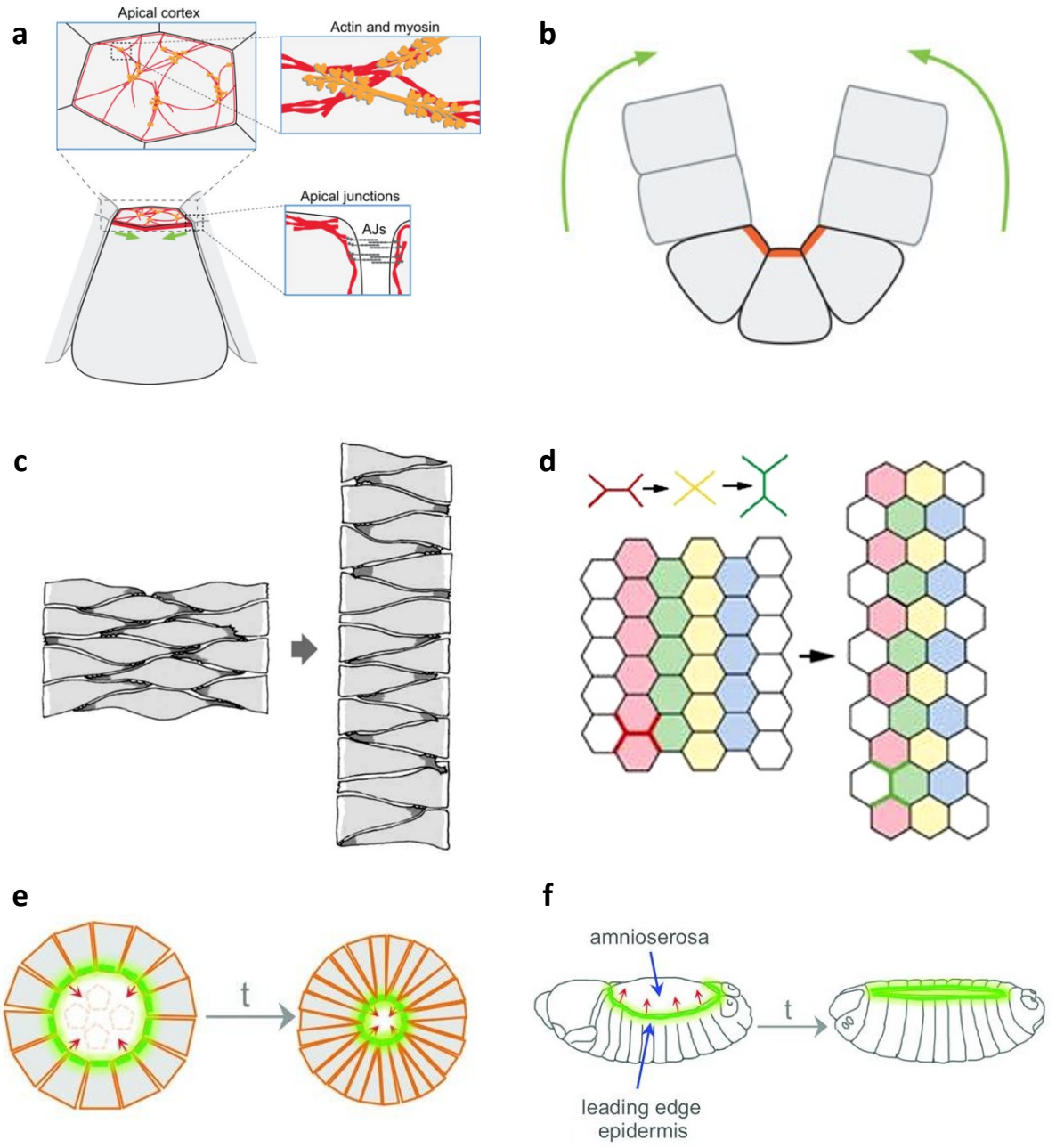

Figure 1.3: Motifs of morphogenesis.

(a-b) Constriction of actomyosin networks is a common force generator of morphogenetic cell deformations. In an epithelial sheet, local constriction on the apical side together with force transmission through adherens junctions (a) can lead to tissue-scale bending and invagination (b). Figures adapted from [Martin & Goldstein, 2014]. (c-d) Convergent extension of tissues can occur either through laterally polarized protrusion and contraction (c) or through junctional shortening and neighbor exchange (d). Both produce a convergence of the tissue along one axis and simultaneously and extension along the other. Figures adapted from (c) [Keller et al., 2000] and (d) [Bertet et al., 2004]. (e-f) Two examples of supracellular actin cables that allow tissues to act as a mechanical unit: the response of embryonic epidermal cells to local wounding (e) and *Drosophila* dorsal closure (f). Figures adapted from [Röper, 2013].

A number of morphogenetic motifs that build upon cells' ability to deform themselves and to exert forces on their surroundings are frequently observed to play a role in different morphogenetic events.

The most-studied such motif is apical constriction, wherein a contractile actomyosin network contracts the apical surface of cells in an epithelial sheet (Fig. 1.3a) [Martin & Goldstein, 2014]. Depending on the geometrical configuration and the mechanical properties of surrounding tissues, this can (1) drive invagination and tissue folding (Fig. 1.3b) as seen in *Drosophila* gastrulation [Sweeton et al., 1991], (2) lead to the formation of radially organized rosettes such as during pancreatic branching morphogenesis [Villasenor et al., 2010], or (3) exert a pulling force on the surrounding tissue as for example in *Drosophila* dorsal closure [Solon et al., 2009].

Another common motif is convergent extension, which is the convergence of a tissue along one axis coupled to the concomitant extension along another [Tada & Heisenberg, 2012]. Convergent extension occurs for instance during *Xenopus* gastrulation [Shih & Keller, 1992] and in *Drosophila* germ band extension [Irvine & Wieschaus, 1994]. Whilst the former is driven by polarized protrusive activity that generates directional cell traction (Fig. 1.3c) [Keller et al., 2000], the latter occurs through a myosin-powered junctional remodeling process that shortens cell-cell junctions along the converging axis, which in turn drives neighbor exchanges and cell intercalations that increase the total length of junctions along the extending axis (Fig. 1.3d) [Bertet et al., 2004].

Due to junction-mediated force transmission cells in a tissue form a collective mechanical system. This is particularly evident in the formation of supracellular actin cables and rings, a third common motif of morphogenesis (Fig. 1.3e-f) [Röper, 2013]. The strong mechanical coupling provided by such cables can allow tissues to act as a single mechanical entity, as for instance in the aforementioned dorsal closure process in *Drosophila*, where force generation by apical constriction of the central amnioserosa cells is converted into an irreversible ingression of the peripheral epidermis by means of a continuous actin cable that shrinks as closure proceeds [Solon et al., 2009].

1.4.3 Collective Migration

Both single cells and entire tissues have the ability to migrate large distances within the developing embryo or in the adult organism. Most commonly, they do so either by crawling along an ECM surface such as a basal lamina (e.g. border cells in *Drosophila* [Montell, 2003]) or by dragging themselves through a three-dimensional meshwork of cells and ECM within another tissue (e.g. in angiogenic sprouting [Betz et al., 2016]). In order for migrating cells to reach their intended destination a guidance cue is usually required, which may take the form of an extracellular gradient of a soluble chemoattractant (chemotaxis) [Dormann & Weijer, 2003] but may also take a range of other forms, including ECM-bound chemoattractants (haptotaxis) [Weber et al., 2013] or gradients in substrate stiffness (durotaxis) [Lo et al., 2000].

The molecular underpinnings of cell migration have been studied extensively. Generally, cells must adopt a polarity along the axis of migration, a process that is directed by the guidance cue and its receptors and orchestrated by the cytoskeleton (Fig. 1.4a), the small GTPase system, and a variety of other factors depending on the particular context [Affolter & Weijer, 2005]. Cells must then protrude in the direction of migration, for instance by forming lamellipodia or other protrusions through actin polymerization (Fig. 1.4a) [Le Clainche & Carlier, 2008] or alternatively by selectively releasing the membrane from the cortex and pushing it outward through internal pressure (blebbing) [Charras & Paluch, 2008]. In order to generate traction force, attachments to the surrounding ECM must then be established (e.g. focal adhesions) [Le Clainche & Carlier, 2008], which is generally followed by myosin-mediated rear-end contraction to actually move the cell body in the direction of the protrusion [Chrzanowska-Wodnicka & Burridge, 1996]. Finally, rear-end ECM attachments must be dissolved in order for migration to progress [Rid et al., 2005].

These cellular and molecular mechanisms underlie both the migration of single cells and the migration of cell collectives. However, collectively migrating cells must additionally maintain cell-cell contacts and mechanical integrity through junctional machineries (Fig. 1.4b-c) [Friedl & Gilmour, 2009; Ilina & Friedl, 2009]. This provides a basis for emergent supracellular behaviors, including the displacement of coupled cells without their active contribution [Ilina & Friedl, 2009], the coordination of cell polarity within the moving group [Theveneau et al., 2010], and coordinated protrusion and retraction behaviors [Vitorino & Meyer, 2008]. Importantly, the mechanics of moving cell groups are complicated and likely complex [Trepat et al., 2009] and may directly feed back on cell behavior by means of plithotaxis, wherein the principal axis of mechanical stress serves as a guidance cue for cell migration [Tambe et al., 2011; Trepat & Fredberg, 2011], possibly through a mechanosensitive cadherin-keratin complex [Weber et al., 2012].

There are a number of additional capabilities conferred to migrating tissues by virtue of their collectivity. For instance, a cell collective may be able to sense guidance cues with greater sensitivity, precision and robustness, since the total area of signal integration is larger [Malet-Engra et al., 2015; Varennes et al., 2016]. Similarly, a collective can migrate far more robustly than an single cell, as directionality can be maintained even if some cells in the tissue occasionally lose polarity, especially in cases where the collective motion of the tissue directly contributes to the polarization of its constituent cells [Haas & Gilmour, 2006].

Importantly, collectively migrating cells generally establish a tissue-scale directional asymmetry in addition to the migrational polarity of individual cells [Vitorino & Meyer, 2008; Rørth, 2012]. A common pattern is for cells at the leading edge (so-called leader cells) to adopt a mesenchyme-like organization, characterized by polarization in the direction of migration, high protrusive activity at the leading edge and more fluid-like cell surface dynamics, whereas cells in the rear (follower cells) become more epithelial-like, exhibiting apico-basal polarity, tight-junction formation and more

elastic cell mechanics (Fig. 1.4c) [Pastor-Pareja, 2004; Lecaudey et al., 2008; Fischer et al., 2009; Friedl & Gilmour, 2009]. These phenotypic leader-follower differences are also reflected at the level of gene expression, for instance in different expression levels of chemokine receptors [Aman & Piotrowski, 2008], metalloproteinases [Nabeshima et al., 2000], and adhesion proteins [Revenu et al., 2014]. Interestingly, follower cells may simultaneously perform additional tasks, such as morphogenesis and differentiation, which must be coordinated with migration to achieve the desired final embryo architecture (see section 1.4.4).

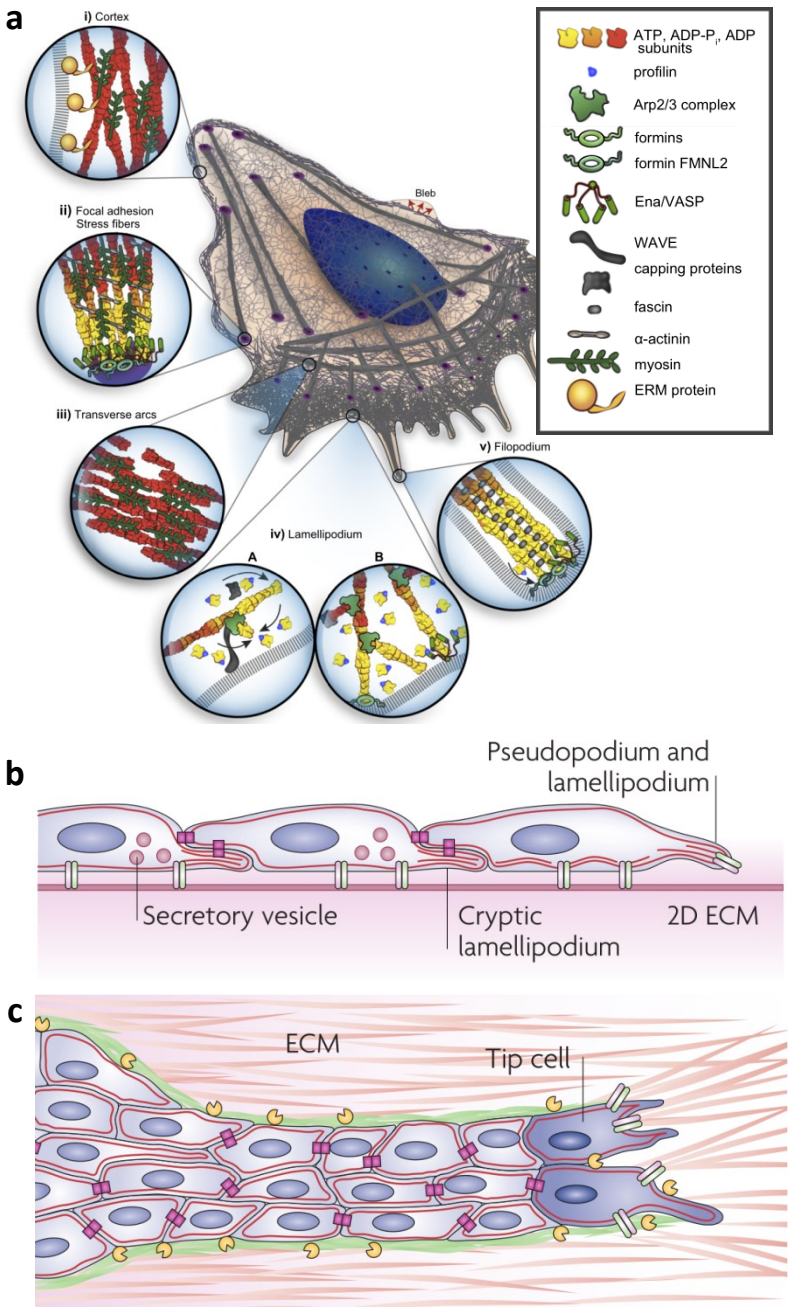

Figure 1.4: Individual and collective cell migration.

(a) A single migratory cell showing the polarized organization of the actin cytoskeleton and associated proteins, which mediate leading edge growth by branched actin polymerization and rear-end retraction by myosin-based contractility. Adapted from [Blanchoin et al., 2014]. (b-c) Collective cell migration on a 2D ECM surface (b) and within a 3D ECM matrix (c). Collectively migrating cells show a similar pattern of actin organization along the axis of migration but simultaneously maintain junctions with their partners. In addition, migrating collectives generally exhibit tissue-scale asymmetry, with highly motile and protrusive tip/leader cells (darker blue cells in c) and more epithelial followers. Communication between collectively migrating cells may occur through biochemical cues but also through tissue mechanics or remodeling of the ECM by secretion of new ECM components (green) or metalloproteinases (yellow). Both (c) and (d) are adapted from [Friedl & Gilmour, 2009].

1.4.4 Interplay

In recent years, a number of examples of feedback between morphogenesis, patterning and/or migration have been found, indicating that these processes do not occur in a simple sequential step-by-step fashion during embryogenesis [Gilmour et al., 2017].

For instance, morphogenesis can feed back onto patterning through changes in tissue shape which affect the distribution and dynamics of diffusible signals. This was found to be the case during the embryonic formation of gut villi, where tissue bending leads to a geometry that favors the local accumulation of the morphogen *sonic hedgehog* at the tip of nascent villi. This in turn leads to a local upregulation of *BMP*, which inhibits intestinal stem cell fate and thus restricts the stem cell niche to regions outside the villi [Shyer et al., 2015].

A related mechanism was discovered to act in the zebrafish lateral line, where rosette-shaped clusters of cells organize around a central microlumen, a small interstitial space sealed off from the rest of the tissue by tight junctions. Within this microlumen, apically secreted *FGF* is concentrated, which feeds back on the behavior of the rosette cells [Durdu et al., 2014]. In this case, the microlumen acts as a spatial insulator of signaling, preventing nearby cells that are not directly participating in rosette formation from seeing high levels of *FGF*. In addition, the microlumen may fulfill a coordinative function: since it is only capable of trapping *FGF* if it is fully formed, it may act as a control point for proper morphogenetic rosette assembly, coordinating the completion of the morphogenetic event with the downstream events triggered by *FGF* (see also section 1.5.2).

Another link between patterning and morphogenesis is established by mechanical forces and their transduction into biochemical signals. As groups of cells undergo shape changes and movements, they exert mechanical forces both within that group and beyond it. These forces are sensed through mechanically-driven conformational changes of junctional [Charras & Yap, 2018] or actin-binding proteins [Harris et al., 2018] and subsequently integrated into cellular decision making through a variety of mechanisms [Mammoto & Ingber, 2010], including direct modification of gene expression by nuclear translocation of *β-Catenin* [Farge, 2003], *NF-κB* [Chen et al., 2003] or *YAP/TAZ* [Dupont et al., 2011]. A wide range of developmental processes such as *Drosophila* anterior-gut induction [Farge, 2003], murine blood vessel remodeling [Lucitti et al., 2007], human mesenchymal stem cell lineage committment [McBeath, 2004], and many others [Mammoto & Ingber, 2010] have been shown to be at least partially under mechanical control.

The addition of collective migration into the mix further complicates things. Collectively migrating tissues are generally patterned themselves, with leader cells in front and follower cells in the rear, as well as potentially intermediates in between [Vitorino & Meyer, 2008; Rørth, 2012; Friedl & Gilmour, 2009]. This pattern can be established and maintained by a number of different mechanisms, for instance Notch-based lateral inhibition and subsequent differentiation [Ghabrial & Krasnow, 2006;

Hellström et al., 2007], leader cell-mediated local modification of the extracellular matrix [Nabeshima et al., 2000], or morphogen signaling between leaders and followers [Aman & Piotrowski, 2008]. Given that migrating tissues are subject to complicated patterns of mechanical stress [Trepat et al., 2009], tissue mechanics are bound to also play a major role not only in their guidance [Tambe et al., 2011] but also in their internal organization. This role, however, remains to be elucidated.

As discussed in section 1.4.3, leader cells are generally mesenchymal-like but follower cells may adopt an epithelial-like organization, including apico-basal polarity and tight junctions [Pastor-Pareja, 2004; Lecaudey et al., 2008]. This makes it possible for followers to undergo additional epithelial morphogenesis concomitant with migration, be it through apical constriction, convergent extension, and/or through the formation of supracellular actin structures. Prominent examples include the formation of a tubular inner lumen during tracheal branching morphogenesis [Caussinus et al., 2008] and the assembly of radially organized rosettes during lateral line migration in fish [Villablanca et al., 2006; Lecaudey et al., 2008]. In such systems, collective migration and follower cell morphogenesis must be properly coordinated, as the relative timing of the two determines the final outcome. Thus, migration can have considerable influence on the overall organization and behavior of a tissue and *vice versa*.

Taken together, these examples illustrate the complicated and undoubtedly complex nature of integrated developmental processes. This runs contrary to an earlier view of development which generally assumed a loosely linear progression from patterning to differentiation to morphogenesis, inspired chiefly by the temporal separation of main body axis patterning and the first morphogenetic events in *Drosophila* [Nüsslein-Volhard & Wieschaus, 1980]. If developmental processes are instead feeding back on each other in a multitude of ways, the resulting emergent complexity dramatically increases the range of possible dynamics and outcomes they can produce (see section 1.2). In this case, the established knowledge discussed in this section likely represents only the very tip of the iceberg, leaving many specific mechanisms and general principles yet to be discovered.

1.5 The Zebrafish Posterior Lateral Line Primordium as a Model

To study the complex interplay of developmental processes in an *in vivo* context we made use of the zebrafish posterior Lateral Line Primordium (pLLP) as a model system.

The pLLP is a group of about 100 cells that collectively migrate directly under the skin along the flank of the developing zebrafish embryo from the back of the head to the tip of the tail, depositing a number of rosette-shaped clusters along the way that go on to form the lateral line sensory system in adult fish (Fig. 1.5) [Ghysen & Dambly-Chaudière, 2007]. It is an ideal model system for this study both because it represents a case where patterning, morphogenesis and collective migration are tightly integrated and because of its experimental accessibility.

This section introduces the pLLP, describes the known mechanisms of self-organization at work within it and briefly explains its strengths as an experimental model system.

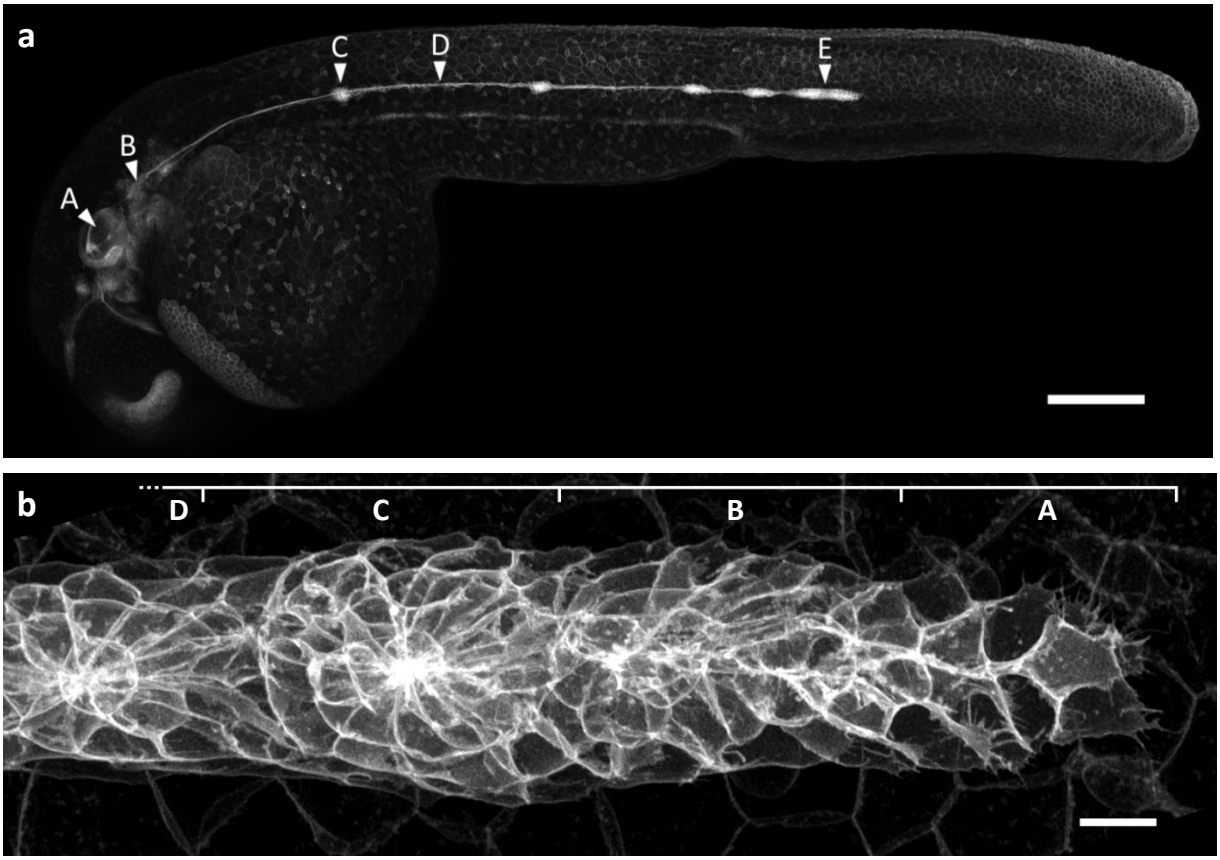

Figure 1.5: The zebrafish posterior lateral line primordium.

(a) Stitched maximum intensity-projected overview of a zebrafish embryo at approx. 35hpf. The lateral line and a few other tissues (including the periderm) are labeled by *claudinB::lyn:EGFP* [Haas & Gilmour, 2006]. Arrows indicate (A) the developing otic vesicle, (B) the lateral line ganglion, (C) a developing (pro)neuromast, (D) interneuromast chain cells, and (E) the pLLP itself. Scale bar: 200µm. (b) Maximum intensity-projected top view of the pLLP with *lyn:EGFP* membrane labeling, showing highly protrusive leader cells (section A), an intermediate transition zone (section B), epithelialized follower cells forming a rosette (section C), and an assembled proneuromast that will soon be deposited (section D). Scale bar: 10µm. Both images were acquired on the Zeiss LSM880 AiryScan (see Materials & Methods, section 2.2.1).

1.5.1 Developmental Context and pLLP Physiology

The lateral line is a mechano-sensory system dedicated to the perception of water movement in fish and amphibians, endowing them with a sense of "touch-at-a-distance" important for a diverse range of behaviors including schooling, hunting, and courtship [Dijkgraaf, 1963]. It consists of a set of discrete sensory organs called *neuromasts* that are dotted across the surface of the adult body in a species-specific pattern, each locally sensing water motion through mechano-sensory hair cells and forwarding the information to the brain via lateral line neurons [Dijkgraaf, 1963; Ghysen & Dambly-Chaudière, 2004].

Lateral line development overall is a complicated multi-stage process starting with the formation of cranial placodes and continuing throughout adult life as additional neuromasts are added to account for the organism's growth [Schlosser, 2006; Ghysen & Dambly-Chaudière, 2007]. Here, we use as our model specifically the posterior Lateral Line Primordium (pLLP) in zebrafish embryos, which combines collective migration and simultaneous morphogenesis to distribute a first set of neuromasts along the flank of the fish.

The pLLP derives from the lateral line placode, a group of cells of the cranial ectoderm located on either side of the head just behind the otic placode (Fig. 1.5a) [Schlosser, 2006]. Following the placodal cells' delamination at 18–20hpf, they split into a smaller anterior group (about 20 cells) which undergo neurogenesis to form the neurons of the lateral line ganglion and a larger posterior group (about 100 cells) which form the lateral line primordium itself [Schlosser, 2006; Ghysen & Dambly-Chaudière, 2007].

The primordium begins to collectively migrate along the horizontal myoseptum at about 20hpf, crawling on top of the basement membrane directly underneath the embryonic surface ectoderm [Ghysen & Dambly-Chaudière, 2007]. It assumes the typical tissue-scale polarity with mesenchymal-like leaders at the front and epithelial-like followers at the rear (Fig. 1.5b) [Lecaudey et al., 2008]. Follower cells undergo apical constriction to form discrete rosette-shaped clusters of about 20 cells (so-called *proneuromasts*) [Lecaudey et al., 2008], which synchronously slow their migration and decouple from the rest of the migrating primordium [Gompel et al., 2001; Haas & Gilmour, 2006]. In this way, the primordium intermittently deposits five proneuromasts along the flank of the fish before reaching the tail at about 40hpf, at which point it fragments to form 2–3 terminal proneuromasts in relatively quick succession [Gompel et al., 2001].

Following deposition, proneuromast cells further differentiate into mechanosensory hair cells, support cells and mantle cells, organized concentrically from inside to outside [Hernández et al., 2007]. The hair cells project kinocilia outside the periderm (protected by a mucous cupula secreted by the mantle cells), which allows them to sense motion in the surrounding water [Dijkgraaf, 1963].

In between proneuromasts, the pLLP continuously deposits a chain of more mesenchymal-like progenitors called *interneuromast cells* which later go on to form additional neuromasts (Fig 1.5a) [Gompel et al., 2001; Grant et al., 2005; Ghysen & Dambly-Chaudière, 2007]. Furthermore, the primordium is closely pursued by outgrowing axons projected from the lateral line ganglion, which innervate proneuromasts upon deposition [Metcalfe, 1985; Gilmour et al., 2004]. Together with glial cells that migrate along these neurites and provide myelination [Gilmour et al., 2002], they form the lateral line nerve, which is responsible for relaying the neuromasts' mechanical measurements to the brain.

1.5.2 Self-Organization of the pLLP

The mechanisms underlying the organization and behavior of the pLLP are known to some extent and have been found to be interesting cases of developmental self-organization.

Patterning of the pLLP into a leader and follower zone is thought to be implemented by mutual repression of *Wnt/β-catenin* and *FGF* signaling, where high *Wnt* activity in leaders suppresses *FGF* signaling via *Sef* and high *FGF* signaling in followers suppresses *Wnt* via *Dkk1* (Fig. 1.6a) [Aman & Piotrowski, 2008; Ma & Raible, 2009]. Furthermore, *FGF* signaling in followers is amplified through positive feedback, partly by *Fgf*-driven expression of *Fgfr1* [Aman & Piotrowski, 2008] and partly by rosette formation (see below) [Durdu et al., 2014]. Such simultaneous mutual repression and self-amplification is a common patterning motif that implements a bistable switch [Ferrell, 2014] and is a special case of both lateral inhibition and the Turing/Gierer-Meinhardt system (see sections 1.2.2 and 1.4.1). There may also be a positive feedback loop stabilizing the leader state in the pLLP but its mechanism is as of yet unknown. With their specific domains of activity stably established, *Wnt* and *FGF* signaling can control leader-specific and follower-specific gene expression, respectively (Fig. 1.6a) [Aman & Piotrowski, 2008; Lecaudey et al., 2008].

Migration of the pLLP is guided by the chemokine *Cxcl12a/Sdf1a*, which is produced along the horizontal myoseptum and sensed by the cells of the primordium through the canonical chemokine receptor *Cxcr4b* [David et al., 2002]. However, unlike one might expect, *Cxcl12a* is not externally pre-patterned as a gradient for the pLLP to follow. Instead, it is uniformly distributed within the interstitial space of the horizontal myoseptum, which is impressively demonstrated by the fact that the primordium can perform a "U-turn" and migrate back toward the head in cases where *Cxcl12a* has been selectively removed in a segment of the embryo's trunk [Haas & Gilmour, 2006]. Directional information is therefore not derived from the environment but is created by the tissue-scale polarity of the primordium itself: *FGF* signaling in the follower cells induces the expression of the non-canonical chemokine receptor *Cxcr7b*, which is thought to be a non-signaling scavenger receptor that locally reduces the extracellular concentration of *Cxcl12a* (Fig. 1.6b-c) [Burns et al., 2006; Dambly-Chaudière et al., 2007]. Indeed, when *Cxcr4b* activity is visualized using a tandem fluorescence timer

approach, it appears graded from front to back – and this gradient disappears upon *Cxcr7b* depletion (Fig. 1.6b) [Donà et al., 2013]. Thus, selective expression of a scavenger receptor in the follower cells carves a self-generated gradient from the uniformly present chemokine which in turn informs the directional polarity of the primordium.

Interestingly, expression of *Cxcr4b* is not required across the entire pLLP, as just a few *Cxcr4b*-expressing cells can rescue migration of a mutant primordium by taking the position of tip cells [Haas & Gilmour, 2006]. They are not, however, mechanically pulling the entire primordium along but instead somehow confer directionality to the *Cxcr4b*-mutant cells, indicating that migrational polarity can be induced by pathways other than *Cxcr4b* signaling [Haas & Gilmour, 2006]. *Fgf* has been suggested to fulfill this function by acting as a chemokine produced and secreted by leader cells under *Wnt* control [Dalle Nogare et al., 2014] (whilst being prevented from signaling in the leaders themselves by *Wnt*-based feed-forward inhibition via *Sef*, as discussed above) but evidence for this model remains weak and somewhat contradictory to the established role of *Fgf* as an inducer of follower cell epithelialization [Lecaudey et al., 2008; Durdu et al., 2014]. As an alternative, it has been suggested that pulling forces generated by motile leader cells might polarize follower cells along the same axis through plithotaxis [Lecaudey & Gilmour, 2006], although conclusive experiments in support of this model are still missing.

Morphogenesis of the pLLP is driven by apical constriction of follower cells leading to rosette formation [Lecaudey et al., 2008]. This behavior is under the control of *FGF-Ras-MAPK* signaling and mediated by apically localized *RhoA-Rock-MyoII* activity, as one might expect in a classical sequence of patterning and morphogenesis events [Lecaudey et al., 2008; Harding & Nechiporuk, 2012]. There is, however, also a direct feedback from rosette architecture to *FGF* signaling (Fig. 1.6d-e) [Durdu et al., 2014]. As a rosette assembles, the apical surfaces of participating cells enclose a small interstitial space known as a *microlumen*, which is sealed shut by tight junctions. This microlumen then begins to fill up with *Fgf* ligand, which continues to be secreted apically by rosette cells (Fig. 1.6d). Consequently, *FGF* signaling is both amplified and spatially restricted such that the cells participating in a mature rosette are selectively exposed to very high signaling levels, which induces them to terminate migration and detaches the proneuromast from the migrating primordium (Fig. 1.6e) [Durdu et al., 2014]. In other words, rosette maturation and deposition is coordinated through a feedback between rosette architecture and *FGF* signaling.

As all of these self-organizing sub-systems are interlinked both biochemically (in particular through *FGF* signaling) and biomechanically, they together form an integrated higher-order system that guides the behavior of the pLLP – a perfect model for studying the emergent effects resulting from such interplay.

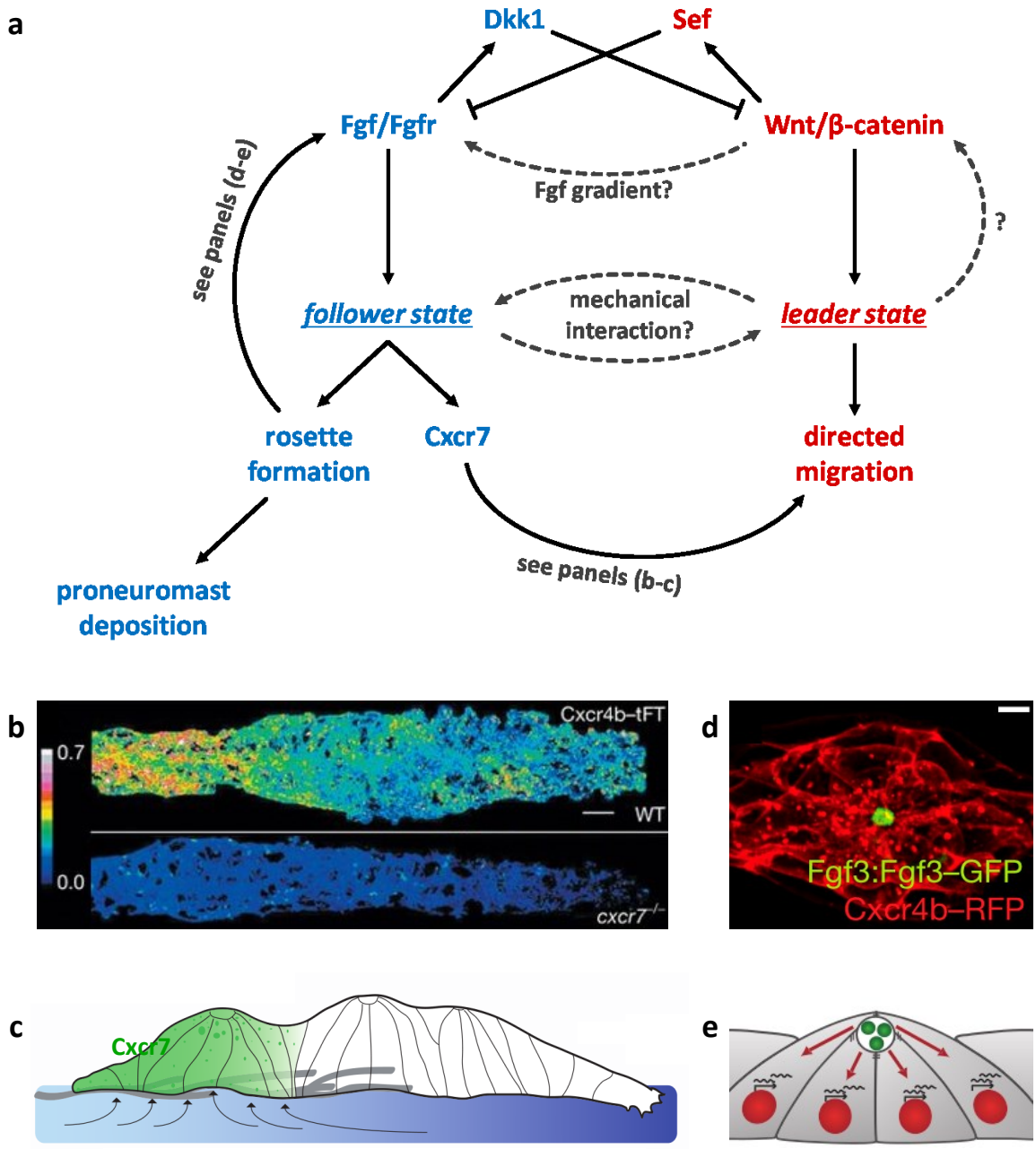

Figure 1.6: Mechanisms of self-organization in the pLLP.

(a) Simplified interaction graph of the most relevant known and hypothesized mechanisms of pLLP self-organization involving the mesenchymal-like leader state and the epithelial-like follower state, including mutual inhibition of the *Wnt* and *Fgf* domains, positive feedback within each domain (in particular via luminal signaling in the follower cells, see d-e), the chemokine gradient generated by the scavenger receptor *Cxcr7* (see b-c), a putative *Fgf* gradient, and putative mechanical interactions such as plithotaxis. (b) Ratiometric images of a *Cxcr4b* tandem fluorescence timer: low ratios indicate high *Cxcr4b* signaling, which can be seen in the front of the wild-type primordium or in the entire tissue in the absence of scavenger receptor *Cxcr7*. Scale bar: 10µm. (c) Illustration of the self-generated gradient model, where follower cells carve a gradient from an otherwise uniform chemokine distribution. (d) Proneuromast showing accumulation of *Fgf3* in the microlumen. Scale bar: 5µm. (e) Illustration of luminal signaling, where local trapping of a ligand leads to restricted and amplified signaling in cells participating in the lumen. (b) and (c) are adapted from [Donà et al., 2013] and (d) and (e) from [Durdu et al., 2014].

1.5.3 Advantages as a Model System

Aside from its biological significance as an example of integrated patterning, morphogenesis and migration, the lateral line model system features a number of technical advantages.

Since pLLP migration and development occurs directly underneath a thin transparent layer of embryonic skin, live imaging is straightforward. Existing tissue-specific *Gal4* [Distel et al., 2009] and chemically inducible *LexPR* [Emelyanov & Parinov, 2008; Durdu et al., 2014] driver systems coupled with the ease of zebrafish transgenics allow for relatively rapid and versatile development, testing and use of new live reporters and genetically encoded perturbation constructs. This combination of genetic and optical accessibility makes the pLLP ideal for the adaptation of optogenetic constructs into an *in vivo* context.

These advantages are compounded with the general advantages of zebrafish as a model organism [Lele & Krone, 1996; Nüsslein-Volhard & Dahm, 2002], including its ease of maintenance and husbandry, its large number of progeny resulting from crosses, its external development which allows embryos to be collected without sacrificing adults, and its well-established genetics and transgenics, including recently developed efficient CRISPR-based genome editing methods [Auer & Del Bene, 2014; Burger et al., 2016].

Taken together, the multi-facetted nature of the lateral line primordium combined with its experimental accessibility make it an ideal model for the establishment of novel approaches to study complex biological systems and for the investigation of how development integrates signaling, shape and movement to give rise to diverse and intricate organs and organisms.

1.6 Aims of this Study

Two overall aims were pursued in this work. Firstly, the discovery of mechanisms and principles of coordination and integration of different aspects of multi-cellular development, in particular patterning, morphogenesis and collective migration. Secondly, the establishment of methods capable of better addressing the challenge presented by the complexity of biological systems, in particular optogenetic precision perturbation and data-driven computational analysis, and the subsequent employment of these methods toward accomplishing the first aim.

These aims were approached from a variety of different angles which can be grouped into three distinct but synergistic projects:

1) Adapting optogenetic tools for the precise manipulation of chemokine and actin signaling in the lateral line primordium.
2) Investigating whether rosette morphogenesis feeds back to regulate pLLP migration by orienting and increasing mechanical tension along the tissue.
3) Developing a computational framework for image-based 3D single-cell segmentation and quantitative analysis of cell shape and architecture across the pLLP.

2 Materials and Methods

2.1 Molecular Biology and Zebrafish Work

2.1.1 Molecular Biology

The plasmids used in this study were generated by MultiSite Gateway cloning (Invitrogen, USA). Sequences of interest were cloned from zebrafish embryonic (30hpf) cDNA extracted using the RNeasy Micro Kit (Qiagen, Germany) and reverse transcribed using SuperScript III reverse transcriptase (Invitrogen, USA) or were amplified directly from existing templates. KAPA HiFi hot-start polymerase (Kapa Biosystems, USA) was employed for PCR amplification and products were checked by Sanger sequencing performed by an off-site service (GATC Biotech, Germany). PCR products were inserted into MultiSite Gateway entry clones by BP reactions, which were checked by sequencing and subsequently combined with promoters (see section 2.1.4) and if required with other coding sequences (e.g. for fluorescent proteins) into Tol2-flanked expression vectors by LR reactions according to the Tol2kit protocol [Kwan et al., 2007]. The destination vectors used contained one of three co-injection markers: *clmc2::EGFP* (green heart), *cryebb::ECFP* (blue eyes), or *cryebb::mKate2* (red eyes)[1]. Final expression vectors were double-checked by restriction analysis using FastDigest (Thermo Fisher Scientific, USA) or NEB (New England Biolabs, USA) restriction endonucleases. Bacterial transformation was performed using Stellar (Clontech Laboratories, USA) or TOP10 (Invitrogen, USA) chemically competent *E. coli* cells.

In-vitro transcription (IVT) for the production of *Tol2* transposase mRNA used to generate stable transgenic lines from expression vectors (see section 2.1.4) as well as for the production of mRNA used to test expression constructs prior to the generation of stable lines was performed using the mMESSAGE mMACHINE kit (Ambion, USA).

2.1.2 Optogenetic Constructs and chemoARHGEF

PA-Rac1 [Wu et al., 2009] was kindly provided by Prof. Dr. Klaus Hahn and Gateway-cloned into expression constructs *LexOP::mCherry:PA-Rac1*[2] and *UAS::mNeonGreen:PA-Rac1* (mNeonGreen was kindly provided by Prof. Dr. Alex Hajnal). PA-Cxcr4b was originally described as a chimera of bovine rhodopsin and human CXCR4 [Xu et al., 2014]; we modified the intracellular domains to match zebrafish *Cxcr4b* and had the resulting sequence synthesized by GeneArt gene synthesis (Thermo Fisher Scientific, UAS). This sequence was then cloned into a *UAS::PA-Cxcr4b:EGFP* expression

[1] These destination vectors had been previously established by a former member of the lab, Andreas Kunze.
[2] This construct was cloned by a former member of the lab, Erika Donà.

vector[1]. bOpsin [Karunarathne et al., 2013a] was kindly provided by Prof. Dr. N Gautam and cloned into *UAS::bOpsin:tagRFP-T* (tagRFP-T was kindly provided by Dr. Jan Ellenberg).

The components of the CRY2-CIBN system [Kennedy et al., 2010] were also kindly provided by Prof. Dr. N Gautam and were used to engineer the membrane recruitment system. A set of 8 constructs was created based on CIBN: *UAS::CIBN:CaaX*, *UAS::CIBN:mNeonGreen:CaaX*, *UAS::Lyn:CIBN*, *UAS::Lyn:mNeonGreen:CIBN, bAct::CIBN:CaaX*, *bAct::CIBN:mNeonGreen:CaaX*, *bAct::Lyn:CIBN* and *bAct::Lyn:mNeonGreen:CIBN*, all directed to the plasma membrane by fusion to either the prenylation motif CaaX [Casey & Seabra, 1996] or the palmitoylation sequence of the tyrosine-protein kinase *Lyn* (here abbreviated as simply Lyn) [Teruel et al., 1999]. *bAct* denotes the near-ubiquitously active β-actin promoter. CRY2 was cloned into both middle and 3' entry vectors to allow the combinatorial generation of a range of CRY2-effector fusions, although initial tests were performed without effectors, using either *UAS::CRY2:tagRFP-T* or *UAS::tagRFP-T:CRY2*, or matching versions under the *bAct* promoter.

The four effector constructs were assembled using LR reactions as described in section 2.1.1. The *GRK3ct* used in *UAS::CRY2:tagRFP-T:GRK3ct* was another kind gift by Prof. Dr. N Gautam. For the *RGS4Δ* of *UAS::CRY2:tagRFP-T:RGS4Δ* we tested both the rat version (rnRGS4Δ, also kindly provided by Prof. Dr. N Gautam) and a cloned version of the zebrafish RGS4 (zfRGS4Δ) identified via ZFIN (zfin.org). For *UAS::CRY2:tagRFP-T:Tiam1a-Gd* and *UAS::CRY2:tagRFP-T:ARHGEF25b-Gd*, the closest zebrafish homologues to the corresponding effectors reported in literature [Levskaya et al., 2009; Van Unen et al., 2015] were found using NCBI's BLAST (blast.ncbi.nlm.nih.gov) [Altschul et al., 1997] and the catalytically active GEF domains were predicted using SMART (smart.embl-heidelberg.de) [Letunic & Bork, 2017]. They were then cloned from zebrafish embryonic cDNA as described in section 2.1.1.

To construct constitutively membrane-localized versions of the *Tiam1a* and *ARHGEF25b* effectors for chemically inducible expression under the LexOP promoter, the same active domains cloned for the CRY2 versions were coupled to Lyn to enforce membrane localization, yielding *LexOP::Lyn:tagRFP-T:Tiam1a-Gd* and *LexOP::Lyn:tagRFP-T:ARHGEF25b-Gd*.

Any relevant sequences that are not already published elsewhere can be found in appendix A1.

2.1.3 Fish and Embryo Handling

Growth, maintenance and breeding of zebrafish (*Danio rerio*) was handled according to standard procedures [Westerfield, 2000]. In accordance with EMBL internal policy 65 (IP65) and European Union Directive 2010/63/EU, all experiments were performed on embryos younger than 3dpf. Unless stated otherwise, embryos were kept in E3 buffer at 27°C or 30°C.

[1] This work was performed by a former member of the lab, Erika Donà.

2.1.4 Transgenic and Mutant Lines

For transient expression and transgenic line establishment, embryos were microinjected at the one-cell stage with an injection mix containing 0.05% phenol red and either 200ng/µl of mRNA or 25ng/µl of plasmid DNA and 100ng/µl of *Tol2* transposase in nuclease-free water (Thermo Fisher Scientific, USA). Embryos were then either imaged or raised to adulthood and screened by checking F1 progeny for expression of fluorescent co-injection markers included in the backbone of expression vectors (see section 2.1.1). Experiments on stable transgenic lines were performed either on F1 embryos positive for the selection markers or on subsequent generations derived from selected F1 embryos.

Unless otherwise specified, one of two transactivation systems was used to drive expression of transgenic constructs. The first consists of a *6xUAS* promoter driven by a *GAL4/UAS:mCherry* enhancer trap line (ETL GA346) [Distel et al., 2009] where the mCherry coexpression marker has been removed by CRISPR/Cas9-mediated deletion [unpublished data[1]]. The second is comprised of a *LexOP* promoter driven by the transactivator *LexPR*, which is chemically inducible by the drug mifepristone (RU486) [Emelyanov & Parinov, 2008] and is expressed in the lateral line under control of the *Cxcr4b* upstream regulatory region in a BAC transgenic line [Durdu et al., 2014]. The two expression systems are referred to as *UAS* and *LexOP* throughout this thesis.

Several of the transgenic and mutant lines used in this study have been generated by other members of lab or by other labs and have been or will be described elsewhere. These include *claudinB:: Lyn:EGFP* (green membrane marker) [Haas & Gilmour, 2006], *cxcr4b::NLS:tdTomato* (red nucleus marker) [Donà et al., 2013], ETL GA346 (Gal4-UAS driver line, described above), *cxcr4b::LexPR* (LexPR-LexOP driver line, described above), *Cxcr4b::Cxcr4b:tagRFP* (used as a red membrane marker) [Donà et al., 2013], *Cxcr7::Lyn:mKate2:sfGFP* (a red-green membrane marker labeling the *Cxcr7* expression domain) [unpublished data[2]], *LexOP::NLS:mIRFP* (a far-red nucleus marker) [unpublished data[3]], *UAS::B4galT1(1-55Q):tagRFP-T* (trans-Golgi) [unpublished data[4]], *UAS::CDMPR:tagRFP-T* (trans-Golgi network and late endosomes) [unpublished data4], *UAS::mKate2:GM130* (cis-Golgi) [unpublished data4], *UAS::mKate2:Rab5a* (early endosomes) [unpublished data4], *UAS::mKate2:Rab11a* (recycling endosomes) [unpublished data4], cxcr4b$^{t26035/t26035}$ (*Cxcr4b-/-* null mutant) [Knaut et al., 2002], cxcr7$^{sa16/sa16}$ (*Cxcr7-/-* null mutant) [Kettleborough et al., 2013], and cxcl12a$^{t30516/\ t30516}$ (*Cxcl12a-/-* null mutant) [Valentin et al., 2007].

All other transgenic lines were made as detailed above and in section 2.1.1. They include the UAS versions of the optogenetic constructs described in section 2.1.2, namely *UAS::mNeonGreen:PA-Rac1*, *LexOP::mCherry:PA-Rac1*, *UAS::PA-Cxcr4b:EGFP*, *UAS::bOpsin:tagRFP-T*, *UAS::Lyn:CIBN*, *UAS:: CRY2:tagRFP-T:GRK3ct*, *UAS::CRY2:tagRFP-T:RGS4Δ*, *UAS::CRY2:tagRFP-T:Tiam1a-Gd*, *UAS::CRY2:*

[1] This work was performed by a former member of the lab, Andreas Kunze, and will be described elsewhere.
[2] This line was generated by a former member of the lab, Alejandra Guzman Herrera, and will be described elsewhere.
[3] This line was generated by another member of the lab, Elisa Gallo, and will be described elsewhere.
[4] These lines were generated by another member of the lab, Mie Wong, and will be described elsewhere.

tagRFP-T:ARHGEF25b-Gd, *LexOP::Lyn:tagRFP-T:Tiam1a-Gd*, and *LexOP::Lyn:tagRFP-T:ARHGEF25b-Gd*. They also include four actin reporters: *UAS::mNeonGreen:UtrCH*, *UAS::tagRFP-T:UtrCH*, *UAS::LifeAct:mNeonGreen*, and *UAS::LifeAct:tagRFP-T*. Here, *UtrCH* is the actin-binding CH-domain of *Utrophin* and *LifeAct* is an actin-binding peptide derived from the yeast protein *ABP140*. Both have previously been described as probes suitable for visualizing F-actin *in vivo* [Burkel et al., 2007; Riedl et al., 2008]. *UtrCH* was kindly provided by Dr. Péter Lénárt and *LifeAct* by Prof. Dr. Roland Wedlich-Söldner. Finally, they also include a *UAS::mNeonGreen:Gγ9* reporter line. The G-protein Gγ9 has been shown to translocate from the cell membrane to intracellular membranes upon GPCR activation and could thus be used as a reporter of GPCR signaling activity [Saini et al., 2007; Karunarathne et al., 2013a]. The Gγ9 used here was a kind gift by Prof. Dr. N Gautam.

2.1.5 Chemical Treatments

If required, embryo pigmentation was prevented by treating embryos with 0.002% N-phenylthiourea (PTU) (Sigma-Aldrich, USA) from 24hpf. Unless otherwise stated, expression of LexOP constructs was chemically induced by treatment with 20µM of Mifepristone (RU486) (Sigma-Aldrich, USA) from 25hpf until the time of mounting (usually approx. 33hpf).

Para-nitroblebbistatin and azido-blebbistatin treatments [Képiró et al., 2012] were at first carried out as has been described for the lateral line primordium [Képiró et al., 2015], using 0.1% DMSO in E3 as a negative control, 10µM of Para-nitroblebbistatin (Optopharma, Hungary) in E3 with 0.1% DMSO as a positive control, and 1µM of azido-blebbistatin (Optopharma, Hungary) in E3 in 0.1% DMSO. Embryos aged approximately 30hpf[1] were incubated in these solutions in the dark for 10min, then mounted as described in section 2.2.1, and after solidification of the agarose incubated with the inhibitors for another 10min prior to 2-Photon uncaging (see section 2.2.5). However, based on experience with other compounds the incubation time before and after mounting was later increased to 30min each and different concentrations of the inhibitors were tested.

For the EDTA treatment to reduce cell-cell adhesion and thus induce cell rounding, embryos were mounted normally (see section 2.2.1) and prepared for imaging at the microscope. Then, the E3 in the dish was exchanged for 0.5M EDTA in PBS. Imaging was performed between 30min and 1h after the addition of the drug.

[1] Képiró and colleagues noted "1 day post fertilization" in their protocols. However, the position of the pLLP in their figures best corresponds to approx. 30hpf at our incubation temperatures.

2.2 Microscopy

2.2.1 Live Imaging: Sample Preparation and Microscopes

Embryos were prepared for live imaging between 26 and 36hpf, depending on experiment specifics. If required, embryos with the correct genotype and/or transgene expression were selected under a widefield epifluorescence microscope. Embryos were then manually dechorionated using forceps and anaesthetized with 0.01% tricaine (Sigma-Aldrich, USA) prior to being briefly immersed in 1% peqGOLD Low Melt Agarose (Peqlab, Germany) in E3 (containing 0.01% tricaine and kept liquid at 42°C) and immediately transferred to MatTek Glass Bottom Microwell Dishes (35mm petri dish, 10mm microwell, 0.16-0.19mm coverglass) (Mattek Corporation, USA). Using a weighted needle tool, the embryos were gently arranged such that one of their lateral sides flatly rests directly atop the glass slide. Following agarose cooling, E3 with 0.01% tricaine was added to the dish. Up to 20 embryos were mounted in a single dish for immediate imaging but at most 10 embryos were mounted for overnight time course experiments. Mounting of early embryos at 8 to 12hpf to test expression and activity of mRNA constructs was performed in the same way but without any tricaine.

The following microscopes were used: PE Ultraview ERS spinning disk confocal (PerkinElmer, USA) (henceforth ERS), PE Ultraview VoX spinning disk confocal (PerkinElmer, USA) (henceforth VoX), Zeiss LSM780 (NLO) scanning confocal with 2-Photon capability (Carl Zeiss, Germany) (henceforth LSM780), Zeiss LSM880 with AiryScan technology (Carl Zeiss, Germany) (henceforth LSM880), Olympus FV1200 with pulsed UV ablation lasers (Olympus, Japan) (henceforth FV1200). Unless stated otherwise, images were acquired using either 20X 0.8NA air objectives or 40X 1.2NA water objectives with Immersol W immersion fluid (Carl Zeiss, Germany). Where relevant, a more detailed account of microscopy parameters is given with the description of particular experiments.

2.2.2 Optogenetics: Activation and Imaging

For global activation, a channel was configured to use the activating wavelength as if it were used to image a corresponding fluorophore. Where possible, another channel was configured to use a non-activating wavelength for simultaneous imaging of a reporter. To measure the initial state, a stack or short time course was first acquired using only the non-activating channel. Following this, a time course using both channels was run to observe the consequences of illumination with an activating wavelength.

Overnight time courses with low time resolution were also run using both wavelengths to test for tissue-scale phenotypes. In particular, primordium migration was scored at approximately 50hpf by measuring the distance from the end of the embryo's yolk extension to the tip of the primordium and normalizing it to the length of the tail (specifically the distance from the end of the yolk extension to the end of the tail) by division, yielding the "pLLP / tail ratio".

For local activation, non-activating channels were configured identically but the FRAP tool (on the VoX) or the region tool (on the LSM780) were set up to use an activating wavelength at a comparable power to illuminate only a small specific region of the sample. Time courses were started using the non-activating channel and after a short time the activating wavelength was triggered to briefly illuminate the region of interest, followed by continued imaging with the non-activating channel.

Activating wavelengths tested were 440nm and 458nm for PA-Rac1, 488nm for PA-Cxcr4b, 405nm, 440nm and 458nm for bOpsin, and 488nm for the CRY2-CIBN system, although other wavelengths were also tested briefly for each construct. Imaging at 561nm and 633nm was considered non-activating for all constructs and was used to image independent reporters. Some reporters were only available with EGFP or mNeonGreen labeling, in which case imaging them was considered to be either equivalent with global activation (for PA-Cxcr4b and CRY2-CIBN) or to possibly be slightly activating (for PA-Rac1 and bOpsin).

For all constructs, we kept embryos in the dark prior to mounting and imaging to avoid possible side-effects of premature activation. Importantly, when working with the CRY2-CIBN system we noticed early on that exposure to ambient light or bright-field illumination during mounting is sufficient to induce complete and long-lasting translocation. We therefore worked with red ambient lighting and introduced a yellow long-pass filter (commonly known as a yellow Post-It note) into our bright-field illumination light path, which proved effective in preventing premature activation.

2.2.3 chemoARHGEF Imaging

Chemical induction of chemoARHGEF or *NLS:mIRFP* expression via the LexPR-LexOP system was performed as described in section 2.1.5 for all endpoint experiments. High-resolution 3D stacks were acquired at the LSM880 as described in section 2.2.7.

For live imaging of induction, embryos were also treated with 20µM RU486 at 24hpf and then mounted at 28hpf as described in section 2.2.1, with both the agarose and E3 medium additionally containing 30µM RU486. Time-lapse imaging was performed using adaptive feedback microscopy to track migrating primordia as described in section 2.2.8.

2.2.4 Acute Laser Ablation of the *Cxcr7* Expression Domain

Laser ablation of the *Cxcr7*-expressing domain of the pLLP was performed at the FV1200 for multiple embryos (aged approx. 30hpf) in a single dish in quick succession, followed by a transfer of the dish to the ERS or VoX for fast time course imaging of the pLLP's response.

Cxcr7 expression was visualized using the *Cxcr7::Lyn:mKate2:sfGFP* BAC transgenic line, which expresses a dual-color (red/green) membrane label under the control of the *Cxcr7* upstream regulatory region. Since this BAC has been shown to recapitulate the expression pattern of *Cxcr7* and

to rescue *Cxcr7-/-* mutants when used to drive expression of *Cxcr7* itself [unpublished data[1]], we here consider it a *bona fide* reporter of native *Cxcr7* expression.

To image the response to ablation, fish carrying this reporter were crossed either to the green membrane label line *claudinB::Lyn:EGFP* or to *Cxcr4b::Cxcr4b:tagRFP* (here used simply as a red membrane marker), yielding embryos with all lateral line membranes labeled in one color and only membranes of *Cxcr7*-positive cells labeled in the other color.

Ablation was performed by firing the FV1200 355nm pulsed UV laser (power: 40%, dwell time: 2.0µs/pxl, objective: UAPON40XW340) onto a series of 15 to 25 single spots distributed across *Cxcr7*-positive cells at the back of the prim, which we found to be the best approach for thoroughly destroying the *Cxcr7*-positive domain without causing dramatic damage to the surrounding tissue. In each dish, 7/10 mounted embryos were treated by ablation, leaving the other 3 as untreated controls.

Time courses acquired on the ERS or VOX were converted to 8bit, stitched and cropped as described in section 2.3.2 and finally projected into kymographs using the *Reslice* tool in Fiji [Schindelin et al., 2012].

2.2.5 Uncaging of Azido-Blebbistatin

Uncaging of azido-blebbistatin was performed on the LSM780. We used *claudinB::Lyn:EGFP* embryos treated with azido-blebbistatin, para-nitroblebbistatin or DMSO as described in section 2.1.5. For each mounted embryo, we first acquired a stack of *Lyn:EGFP* using standard imaging conditions, then performed uncaging, and finally acquired another stack.

For uncaging, we sought to match the conditions described in the literature [Képiró et al., 2015] but technical limitations prevented us from exactly reproducing them, which is why we conducted a series of tests with different settings. Using a plan-apochromat 20X 0.8NA M27 air objective and 800nm 2-Photon laser light, we tested different combinations of laser power (1% to 5% in steps of 1%), pixel sizes (0.83µm, 0.55µm, 0.42µm) and pixel dwell times (25µs and 50µs). We illuminated a rectangular ROI tightly fit over the entire pLLP and scanned through the tissue using 100 slices per stack, adjusting the total number of stacks to approximately match the reported total of 8min of continuous illumination.

Following uncaging, embryos were kept mounted in the agarose dish in E3 buffer with 0.01% tricaine and were either imaged at the VoX or ERS spinning disk microscopes to observe cell and tissue-scale phenotypes over time or were left until the next morning (approx. 50hpf) to score the effects of the drug treatments and uncaging on long-term migration using a widefield epifluorescence microscope. Since only the primordium close to the coverglass was irradiated during uncaging, the primordium on

[1] This work was performed by another member of the lab, Mie Wong, and will be described elsewhere.

the other side of each embryo provided a direct internal control for the effect of 2-Photon irradiation itself.

Despite testing a number of different conditions, we were unable to reproduce the results reported in [Képiró et al., 2015]. We found that 2-photon irradiation either had no effect when intensities were set too low or dwell times too short, or – unexpectedly – that it led to an arrest of migration independently of whether the embryos were treated with azido-blebbistatin or not (data not shown). In no case did we find that the combination of azido-blebbistatin treatment and 2-photon irradiation had a specific non-additive effect on primordium migration. We therefore ultimately decided to make use of the unexpected finding that an appropriate amount of 2-photon irradiation caused an arrest of migration on its own (see sections 2.2.6 and 3.2.2).

2.2.6 2P-Arrest of Follower Migration

Arrest of follower migration by 2-Photon irradiation was performed with embryos transgenic for *Cxcr4b::NLS:tdTomato* and either *claudinB::Lyn:EGFP* or *UAS::mNG:UtrCH*. Treatment was conducted at the LSM780, where we first acquired a stack with standard 3D confocal settings, then irradiated the pLLP with 2-Photon light, and subsequently acquired another standard stack. After treating 3 of 4 embryos in a dish in quick succession, the dish was moved to the ERS or VoX and fast time course acquisition was started on all 4 embryos, the unperturbed case serving as a matched control.

For 2-Photon irradiation, we used a plan-apochromat 20X 0.8NA M27 air objective, a pixel size of 0.42µm, a pixel dwell time of 50µs, and 800nm 2-Photon light at 3% power, which we measured to equate 33mW before the objective. Notably, decreasing the power to 1% resulted in no phenotype and increasing it to as little as 5% resulted in visible damage to the primordium. We selected a rectangular ROI covering either the entire primordium (for complete arrests during initial tests) or only the followers (from the back of the primordium to the center of the frontal-most rosette) and we scanned across a z-stack ranging from the apical to the basal side with a total of 100 slices, resulting in a total of approximately 5min of irradiation per sample.

Subsequent multi-position time course imaging of the leader cells' response was performed at the ERS or VoX using a 40X 1.2NA water objective, acquiring stacks of 16 slices with a step size of 2.5µm at a speed of 5min or 3min per time point. Data analysis is described in section 2.3.4.

2.2.7 3D Live Imaging at the Zeiss LSM880 AiryScan

As a basis for single-cell segmentation, high-resolution 3D stacks (voxel size: 0.099µm in xy, 0.225µm in z) of the membrane marker *claudinB::Lyn:EGFP* were acquired on the LSM880 using a 40X 1.2NA water objective and the AiryScan FAST mode [Huff, 2016] with a piezo stage and bi-directional scanning, which allowed an entire volume to be imaged in approximately 20 seconds (or 40 seconds for two colors). Subsequently, the built-in 3D AiryScan deconvolution was run with 'auto' settings.

Importantly, achieving optimal image quality required adjustment of the stage such that the incident light beam coming from the objective was exactly normal to the plane of the cover glass. To achieve this, we used 633nm reflected light and xz line scanning to visualize the cover glass interface and manually optimized the pitch of the stage, then repeated the process with a yz line scan. This was done every time a new dish was placed on the stage, although adjustments were not always necessary.

In addition to the *claudinB::Lyn:EGFP* green membrane label, secondary markers were imaged using red or far-red fluorophores. These include *Cxcr4b::NLS:tdTomato* (nuclei), *UAS::tagRFP-T:UtrCH* (F-actin), *UAS::B4galT1(1-55Q):tagRFP-T*[1] (trans-Golgi), *UAS::CDMPR:tagRFP-T* (trans-Golgi network and late endosomes), *UAS::mKate2:GM130* (cis-Golgi), lysotracker deep-red (Thermo Fisher Scientific, USA) (acidic compartments), *UAS::mKate2:Rab5a* (early endosomes), and *UAS::mKate2:Rab11a* (recycling endosomes). Furthermore, *claudinB::Lyn:EGFP* was imaged in a number of homozygous mutant backgrounds, namely *Cxcr4b-/-* (chemokine receptor mutant), *Cxcr7-/-* (scavenger receptor mutant), and *Cxcl12a-/-* (chemokine mutant).

The comparison images in figure 3.7 were acquired (a) on the LSM880 using a classical confocal scanning mode with the objective and settings matching those used for AiryScan acquisition and (b) on the VoX using a matching objective and identical z-step size and other settings optimized for high-quality images (high laser power, 200ms exposure, 77 sensitivity). Both volumes were deconvolved using Huygens Remote Manager (v3.4.1) (Scientific Volume Imaging B.V., The Netherlands) with the corresponding standard settings.

2.2.8 Adaptive Feedback Microscopy for On-Line Tracking of the pLLP

Automated feedback microscopy at the LSM880 was implemented using our python development stack (see section 2.3.1) in conjunction with the *Microscopy Pipeline Constructor* (MyPiC) macro [Politi et al., 2018], which provides a means of interacting with the *ZEN Black* software (Carl Zeiss, Germany) (Fig. 3.21b).

The pipeline, named *prim tracker 880* (pt880), is based on a very high-speed, very low-quality 3D pre-scan of the primordium, usually acquired using conventional confocal settings (not AiryScan) with very large pixel size (0.82µm), low pixel dwell-time (2.18µs), and large z step size (3µm) (Fig. 3.21c). When ZEN/MyPiC automatically saves this stack to a pre-specified directory, the running pt880 process detects the newly generated image file, loads it and masks the primordium (Fig. 3.21d) through a simple image analysis workflow consisting of automated thresholding and subsequent object filtering to retain only the largest object. Currently, this workflow is optimized to work with the bright *claudinB::Lyn:EGFP* marker and cannot handle more dim or mosaic markers. However, the modular design of pt880 should allow its reconfiguration to different markers or even different

[1] This and the following 5 datasets were acquired by another member of the lab, Mie Wong.

tissues with relative ease, provided a corresponding image analysis workflow can be developed. Based on the mask, an appropriate 3D adjustment of the stage position is calculated and fed back to MyPiC via the Windows registry, which then triggers stage movement and the acquisition of a high-quality AiryScan FAST mode stack (as described in section 2.2.7) (Fig. 3.21e). Using MyPiC, these steps can be configured to be repeated across multiple positions (i.e. multiple embryos in a dish) and then looped over time to generate a time course (Fig. 3.21a).

There are a number of fail-safes integrated into pt880, including the option of limiting the total possible stage movement in z (to avoid the risk of damaging the objective in case tracking goes wrong for some reason), the detection of cases where a primordium has likely moved out of the frame since the last time point (leading to a greater stage movement to catch up), and the detection of cases where segmentation likely failed for some reason (leading to a default amount of stage movement).

The pt880 software can be customized for different use-cases relatively easily. For instance, it has been used with only minor modifications as a software autofocus for lateral line chain cells[1] and as an on-line tracker for zebrafish embryonic microglia in the brain[2] (data not shown). However, pt880 is currently still in *alpha* (preliminary implementation, not feature complete). It will be extended and refined in due course and the code will be released as open source software subsequently. Until then, the code is available on request.

[1] This work was done together with another member of the lab, Elisa Gallo.
[2] This work was done with a member of the lab of Francesca Peri, Katrin Möller.

2.3 Computational Work

2.3.1 Software Development Stack

Software development was performed using the Anaconda distribution (Anaconda Inc., USA) of python 2.7.13 (64-bit) (Python Software Foundation, USA) [Van Rossum, 1995] on the Windows 7 Professional SP1 (64-bit) operating system (Microsoft Corporation, USA).

A number of scientific libraries and modules were employed: numpy 1.11.3 [Travis & Oliphant, 2006] and pandas 0.19.2 [McKinney, 2010] for numerical computation, scikit-image 0.13.0 [Van der Walt et al., 2014] and scipy.ndimage 2.0 [Jones et al., 2001] for image processing, scikit-learn 0.19.1 [Pedregosa et al., 2011] for machine learning, matplotlib 1.5.1 [Hunter, 2007] and seaborn 0.7.1 [Waskom et al., 2016] for plotting, networkx 1.11 [Hagberg et al., 2008] for graph-based work, tifffile 0.11.1 [Gohlke, 2016] for loading of TIFF images, and various scipy 1.0.0 [Jones et al., 2001] modules for different purposes. Parallelization was implemented using dask 0.15.4 [Dask Development Team, 2016].

Prototyping, workflow management and exploratory data analysis were performed in Jupyter Notebooks (jupyter 1.0.0, notebook 5.3.1) [Kluyver et al., 2016], including interactive visualization using ipywidgets 7.4.1. Software engineering was conducted using the Spyder IDE (spyder 3.2.4) [Raybaut et al., 2018] or Notepad++ 6.7.5 [Ho, 2016]. Version control was handled with Git 2.12.2.windows.2 [Torvalds et al., 2018] linked with an EMBL-hosted instance of the GitLab repository manager (GitLab, USA).

All software is available on request. The software for single-cell segmentation, cluster-based embedding, atlas prediction and related analysis will be made freely available as open source software following publication of the project's findings. The software for automated feedback microscopy of the pLLP will be made freely available as open source software following some further extension, testing and refinement.

2.3.2 Image Preprocessing

Some basic image processing was performed using the Fiji distribution [Schindelin et al., 2012] of ImageJ 1.52g [Schneider et al., 2012] (henceforth referred to as Fiji). In particular, all images in formats other than the TIFF format (.tif) were converted to TIFF using Fiji and all images with bit-depths higher than 8bit were converted to 8bit by rescaling, using a simple automated Fiji macro. Rescaling was performed either between the minimum and maximum value of the image/stack (if intensity differences between samples were not relevant) or between a fixed, manually determined minimum and maximum value applied across all images of a particular experiment (to preserve relative intensity differences between samples). Multi-position tilings along an embryo were stitched using the *Grid/Collection Stitching* plugin [Preibisch et al., 2009]. In some cases, images were cropped

to reduce the amount of empty space surrounding the object of interest. For the purpose of figure generation (but never prior to any quantitative data analysis), some images were aligned/reoriented using Fiji's *Straighten* tool.

2.3.3 Single-Cell Analysis of chemoARHGEF Phenotypes

Following acquisition (see section 2.2.3), stacks were preprocessed as described in section 2.3.2 and single cells were segmented as described in section 2.3.5.

Mean intensities of the red (chemoARHGEF) and/or the far-red (*NLS:mIRFP*) channel were extracted for each cell and normalized for each condition independently by linear rescaling of the minimum-maximum range to between 0 and 1. Normalized mean intensity was then used to split the cell population into highly expressing and non/weakly expressing cells based on a percentile threshold: cells with an intensity above the 80th percentile were considered highly expressing.

Cell sphericity was measured as the mean deviation of cell surface ISLA landmarks (see section 2.3.6) from a sphere with a radius equal to the mean distance of surface landmarks from the centroid. This measure was then inverted and normalized between 0 and 1. Thus, a value of 1 would indicate that a cell is perfectly spherical. Distance from the midline was calculated as the absolute second principal component of a PCA of the 3D coordinates of all centroids of a primordium, which corresponds to the perpendicular distance of a given centroid to the primordium's front-to-back midline.

A resampling analysis was performed to test whether the location of chemoARHGEF-positive cells alone can explain their sphericity or whether chemoARHGEF affects both location and sphericity. To this end, the 3D centroid positions of all chemoARHGEF (N=20) and wild-type (N=26) primordia were overlaid onto the same spatial frame of reference (the TFOR; see section 2.3.6). For every cell with high chemoARHGEF expression, all wild-type cells in its immediate spatial proximity (closer than 5µm; about one cell radius) were found using a KDTree ball point query approach (using the *scipy.cKDTree* class) and one of them was selected at random. The sphericities of all thus selected wild-type cells was then compared to the sphericities of the corresponding chemoARHGEF-positive cells. To increase the robustness of the result, this sampling was computed 1000 times, with the random selection among available wild-type neighbors for each chemoARHGEF-positive cell being different each time. The plot in figure 3.4. shows the result from a single representative sampling run and reports the average Cohen's d across all runs.

All other statistical analysis was performed as described in section 2.3.10.

2.3.4 Analysis of 2-Photon Arrest Dynamics

Time courses of pLLP migration following 2-photon arrest of the follower cells acquired at the VoX or ERS microscope were preprocessed as described in section 2.3.2 and subsequently maximum

intensity z-projected using a Fiji macro. All analysis described in this section was performed based on such projections.

Binary masks of the moving primordia were obtained from the *NLS:tdTomato* channel (nuclei) and from the *lyn:EGFP* channel (membranes) (if available) using a masking workflow consisting of the following steps: median-filtering with a 5pxl square structural element to reduce detector noise, Gaussian filtering with σ=10pxl (nuclei) or σ=5pxl (membranes), automated threshold detection (described below), binarization with the detected threshold, removal of smaller objects (all below the median object size), and finally edge smoothing using 10 iterations of binary dilation followed by 10 iterations of binary erosion, both with a disc-shaped structural element (r=5pxl). This workflow was applied to each time point separately.

```
# Code Snippet 2.1 (python 2.7)
# thresholds <- list of all thresholds
# n_objects  <- list of corresponding object counts

# Run through all thresholds
for i in range(len(thresholds)):

    # Check if the peak is before the current value
    if max(n_objects[:i]) > n_objects[i]:

        # Check if the current value is below a given fraction of the peak
        if n_objects[i] <= max(n_objects[:i]) / peak_fraction:

            # Accept current threshold
            target_threshold = thresholds[i]
            break

        # Alternatively: check if current value is followed by an increase
        elif n_objects[i+1] > n_objects[i]:

            # Accept current threshold
            target_threshold = thresholds[i]
            break

    # Fallback: If 0 is reached, the previous threshold is accepted
    if n_objects[i] == 0:

        # Accept previous threshold
        target_threshold = thresholds[i-1]
        break
```

The automated threshold detection algorithm is based on object counting and inspired by a semi-automated approach for smFISH spot detection [Raj et al., 2008]. It works by running through all possible thresholds (255 in an 8bit image), binarizing the image with each threshold and counting the number of separated foreground objects. The resulting empirical function, $n_{objects} = f(threshold)$, typically shows an initial increase in the number of objects (as the background is split) followed by a sharp decrease (as background objects are discarded) and then a long flat tail with only minor increases (as foreground objects are split) and a slow decrease (as foreground objects are discarded). The region just after the sharp decrease is ideal for thresholding, as it discards most background

objects whilst retaining most foreground objects. Here, we Gaussian-smoothed (σ=3) the function above and used the heuristic approach detailed in Code Snippet 2.1 to automatically determine a suitable threshold within the ideal region.

Note that *peak_fraction* is a user-defined parameter, which in this case was set to 2.0. We optimized all workflow parameters by visual inspection of the output and manual adjustment. We also visually inspected all outputs following the workflow's execution, ensuring that usable results were generated across time points and samples.

Using the binary masks, we extracted the pLLP's tip position and computed its differential over time to obtain tip speed. As not all movies had the same spatial and temporal resolution, the positions and speeds were converted from pixels to micrometers and temporal resolution was increased to 1min across all samples (from originally 5min or 3min) by local linear interpolation.

We also computed tip cell protrusion lengths, which we approximated as the difference between the tip of the nucleus mask and the tip of the membrane mask – in other words, the distance from the frontal-most point of the tip cell's nucleus to the frontal-most point of its membrane protrusions. However, because leader cell protrusions are difficult to segment accurately, we considered the automated extraction of this measure from the binary masks to be unreliable. We therefore used the Fiji's *Manual Tracking* plugin [Cordelières et al., 2017] to manually annotate nucleus and membrane tip positions, which we then parsed into a format ready for analysis using a python script.

For movies containing the actin label *UAS::mNeonGreen:UtrCH* we created a mask based on the red nuclei and then reduced this mask to the tip region by including only the frontal-most 40'000 pixels. On the *mNeonGreen:UtrCH* images we first performed local background removal by subtracting the local mean (computed within a 30x30pxl square structural element) and setting negative pixels in the resulting images to zero. This removes the background signal of unbound probes, making the remaining signal a proxy of F-actin abundance. We then measured the mean of the background-subtracted intensity over time in the masked tip region. Furthermore, we employed a function from the OpenCV python bindings (Itseez, USA) to compute the Farneback optical flow [Farnebäck, 2003] of the actin signal (without background subtraction) across the entire image and subsequently extracted mean and standard deviations of both optical flow magnitudes and optical flow angles within the tip region.

All of the above measures showed considerable noise from one time point to the next. Therefore, we also computed a smoothed version of each measurement using a 1D Gaussian smooth (σ=9min) to emphasize continuous trends.

Data visualization and analysis was performed with the tools and methods described in sections 2.3.1 and 2.3.10. Importantly, all analyses involving time in some form used *time post irradiation*, which differs for each embryo in a dish due to the 5min of irradiation time (+1min of handling) required per

embryo. Thus, imaging of the embryo irradiated last begins at t=0min, imaging for embryo irradiated second to last begins at t=6min, and imaging for the embryo irradiated first begins at t=12min.

Code written for this project is available on request.

2.3.5 Single-Cell Segmentation Pipeline

Stacks of primordia with membranes labeled by *claudinB::Lyn:EGFP* were acquired at the LSM880 as described in section 2.2.7 and preprocessed as described in section 2.3.2.

Samples that also contained the *cxcr4b::NLS:tdTomato* nuclear label showed some unavoidable bleed-through into the green membrane channel, increasing the background within cells. Making use of the simultaneously acquired red channel (which does not feature any bleed-through from *Lyn:EGFP*), linear unmixing was applied to remove the contribution of *NLS:tdTomato* (C, contaminant image) from the green channel (M, mixed image) and thus retrieve the cleaned *Lyn:EGFP* image (U, unmixed image). To do so, we assumed that the signal in M was composed according to eq. 2.1 and that we could therefore retrieve U by subtracting the contamination term from M (eq. 2.2).

$$M = U + a \cdot C \quad \text{(eq. 2.1)}$$

$$U = M - a \cdot C \quad \text{(eq. 2.2)}$$

The bleed-through factor a was determined by iteratively searching for a value of a where the correlation $CORR$ (eq. 2.3) between the cleaned image U and the contaminant image C is minimal.

$$CORR = pcc(C,\ |M - a_i \cdot C - mean(M - a_i \cdot C)|) \quad \text{(eq. 2.3)}$$

Note that the values of U used to compute this correlation were centered around their mean and subsequently converted to absolute values, which ensures that unreasonably high values of a are punished because overly unmixed regions start correlating with C again. The correlation value itself is the Pearson Correlation Coefficient (pcc) of the thus prepared images. This approach robustly removed *NLS:tdTomato* bleed-through and produced unmixed images that – unlike the raw images – could be segmented successfully.

Following acquisition, preprocessing and if necessary linear unmixing, the membrane images were fed into the segmentation pipeline. The pipeline consists of the following steps:

1. Median 3D smooth with cuboid *3x3x3* voxel structural element to reduce shot noise.
2. Gaussian 3D smooth with σ=*3pxl* to further reduce noise and smoothen structures.
3. Thresholding to retrieve mask of foreground objects (membranes).
   The appropriate threshold was automatically detected using an approach similar to the one described in section 2.3.4. However, here a base threshold was determined as the most frequent value in the image histogram. Starting from this base threshold, a limited range of positive offsets (usually 0 to 10 in steps of 1, for lower-quality images 0 to 40 in steps of 2)

was scanned iteratively, counting the number of connected components in the inverse of the mask generated by applying each given threshold. The threshold resulting in the largest number of such connected components was considered the best threshold and was applied to generate the final membrane mask.

4. Removal of disconnected components using morphological hole filling.
5. Labeling of connected components on the inverted membrane mask. This ideally yields one connected component per cell, that is the cytoplasm.
6. Removal of connected components smaller than 1'000 voxels (artifacts) and assignment of connected components larger than 1'000'000 voxels as background objects.
7. Watershed using labeled connected components as seeds and the smoothed input image (with an additional 3D Gaussian smooth with *σ=3pxl* on top of steps 1 and 2) as topography. The background objects surrounding the prim were also considered seeds.
8. Assignment of zero label to background objects and removal of any objects disconnected from the primordium by retaining only the single largest foreground object.

The parameters of this pipeline were manually optimized by inspection of outputs during test runs.

All primordia segmented by this approach were manually double-checked and rare cases exhibiting substantial segmentation errors were excluded from downstream analysis.

2.3.6 ISLA and CBE

Intensity-biased Stochastic Landmark Assignment (ISLA)[1] (Fig. 2.1a) was performed on cropped-out volumes of single segmented cells, with voxels outside the segmentation mask set to zero and voxel intensities normalized such that their sum equals 1 (by dividing each by the sum of all). Landmarks were assigned by considering the normalized voxel intensities as the probabilities of a multinomial distribution from which 2000 points were sampled (with replacement). Landmark coordinates were then transformed from pixels to microns to account for anisotropic imaging.

When generating point clouds to represent pure cell shape (Fig. 2.1a; right-hand side), the intensity image used for ISLA was not the actual membrane marker image. Instead, a binary image of the inner hull of the segmentation mask was used. When generating point clouds to represent intensity distributions (e.g. for the additional markers of cellular architecture detailed in section 2.2.7) (Fig. 2.1a; left-hand side), a simple background subtraction was performed prior to ISLA to prevent landmarks from being assigned spuriously due to background signal. The background level was determined as the mean intensity within the masked cell and was subtracted from each voxel's intensity value, with resulting negative values set to zero.

[1] ISLA was partially inspired by a similar method developed in the lab of Prof. Dr. Julie Theriot, who kindly described their approach to us. To the best of our knowledge, their approach as not been published yet and therefore cannot be cited here.

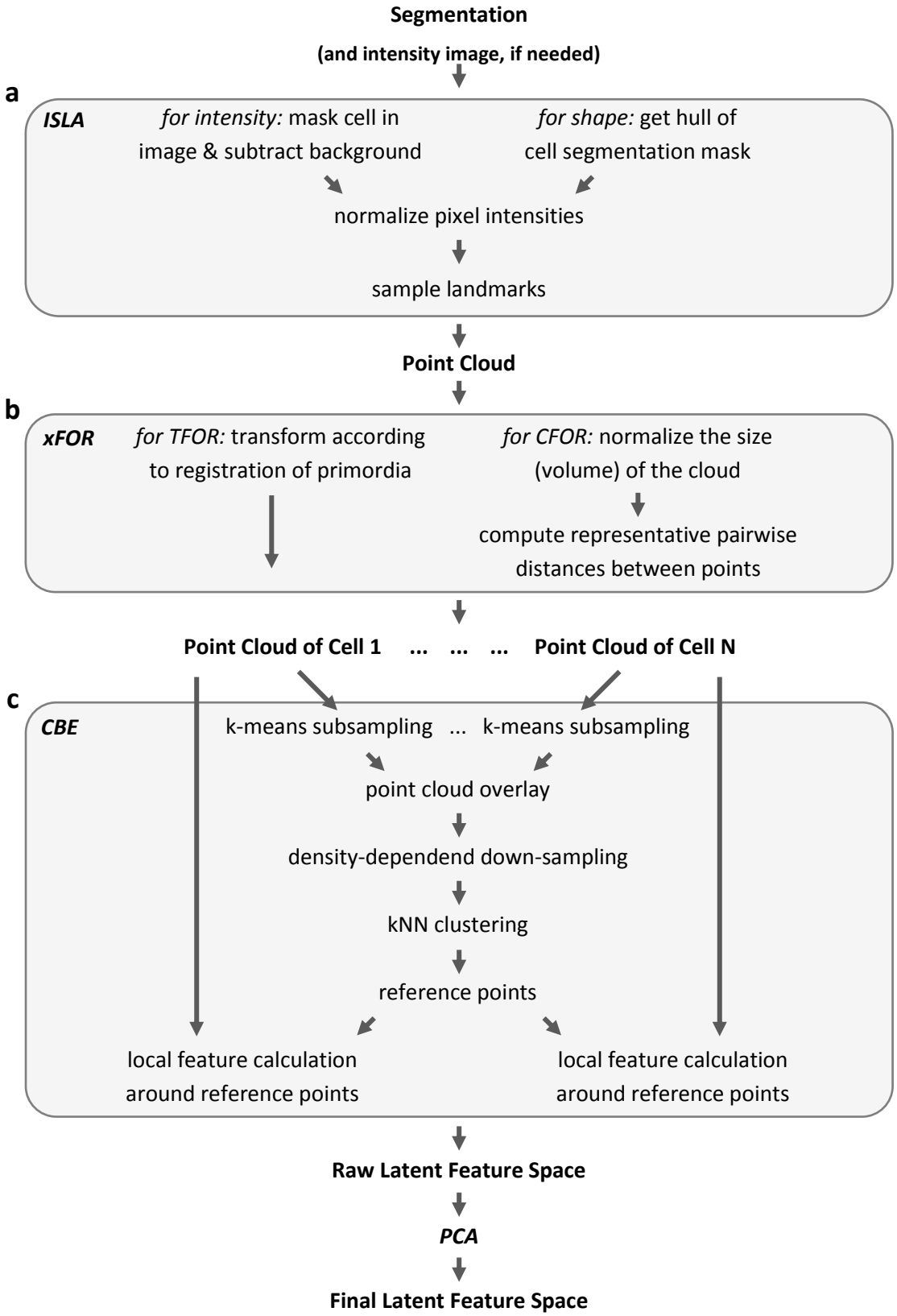

Figure 2.1: Overview of point cloud extraction and embedding pipeline.

Data flows from 3D single-cell segmentations in the form of labeled image volumes as well as corresponding intensity images (if required) to (a) ISLA point-cloud sampling from either a masked intensity distribution or a segmentation hull, to (b) conversion to either the Tissue Frame of Reference (TFOR) or the Cell Frame of Reference (CFOR), and finally to (c) latent feature extraction by Cluster-Based Embedding (CBE). The modular nature of the pipeline allows individual modules (gray boxes) to be replaced by other algorithms if needed. For instance, extraction of point cloud moments was also attempted as an alternative to CBE.

To enable extraction of latent features from a matched *Tissue Frame of Reference* (TFOR) (Fig. 2.1b, left-hand side), primordia were aligned using a simple PCA-based approach that does not require image registration. 3000 landmarks were sampled (using ISLA) from a given primordium's overall segmentation mask and the PCA of this matrix (3000 landmarks in 3 dimensions) was computed. Given that the pLLP's longest axis is always its front-rear axis and the shortest axis is always the apico-basal axis, such a PCA transformation snaps primordia that had been acquired at a slight slant into an aligned frame of reference. To complete the alignment, the primordial point clouds were then translated such that the frontal-most point becomes the spatial origin (0.0, 0.0, 0.0). The same transformation can then be applied to landmarks extracted from individual cells, which results in cellular point clouds that are oriented exactly as if the primordia had been registered prior to their extraction.

The only issue that required resolution in order for this approach to work as intended is the directional ambiguity of the principal component (PC) axes, the sign of which can differ between samples. This was safely resolved by inspecting the contribution of the image axes to each PC in order to match the highest-contributing image axis to its corresponding PC, flipping the PCs sign if necessary. This ensures that the sign of PCs always follows the sign of the image axis. Thus, a correct alignment will be produced so long as the primordia are always acquired at slants less than 90° from a perfectly aligned position, which our imaging workflow could guarantee.

To create a *Cell Frame of Reference* (CFOR) that is invariant to size and rotation (Fig. 2.1b, right-hand side), point cloud sizes were first normalized such that the sum of the magnitudes of all centroid-to-landmark vectors is 1, which corresponds to a normalization for cell volume. Second, cellular point clouds were cast into a pairwise distance (PD) representation. In the PD space, each point of the cloud is no longer characterized by three spatial coordinates but instead by the distances to every other point of the cloud. This representation is rotationally invariant but also extremely high-dimensional (an LxL matrix, where L is the number of landmarks). To reduce this dimensionality, only the 10th, 50th and 90th percentiles of all pairwise distances for each point were chosen to represent the point (resulting in an Lx3 matrix), which we reasoned would encode both local and global relative spatial location. However, although we have empirically found this approach to work reasonably well for the cells of the pLLP, it should be noted that we have not yet formally investigated the potential loss of information resulting from the re-representation of point clouds in such a reduced pairwise distance space.

To determine cluster centers for Cluster-Based Embedding (CBE) (Fig. 2.1c), point clouds from multiple samples were centered on their respective centroids and overlaid. K-means clustering was performed on this overlay cloud (using scikit-learn's *MiniBatchKMeans* implementation) with *k=20*. The resulting cluster centers were used as common reference points for the next step.

Several measures were taken to improve the robustness and performance of cluster detection. First, individual cellular point clouds were downsampled from 2000 points to 500 points prior to being overlaid (using k-means clustering with *k=500* clusters, the centers of which were used as the new landmarks). Second, not all available cells were used in the overlay. Instead, a number of primordia (at least 10, at most 25) were selected and only their cells were used in the overlay, whereas the resulting cluster centers were used as reference points across all available samples. Third, the entire overlaid point cloud was downsampled using a density-dependent downsampling approach inspired by [Qiu, 2011], yielding a final overlaid cloud of at most 200'000 points, which allowed reference cluster centers to be computed reasonably efficiently.

Density-dependent downsampling was performed using a simplified version of the algorithm described in [Qiu, 2011]. First, the local density (LD) of points is found, which here is defined as the number of points in the local neighborhood, i.e. within a sphere where the radius is the median pairwise distance between all points multiplied by a given factor (here 5). Next, a target density (TD) is determined, which in accordance with [Qiu, 2011] was set to be the third percentile of all local densities. Now, points are downsampled such that the probability of keeping each point is given by equation 2.4. If necessary, the resulting downsampled distribution is further reduced by random sampling in order to reach the maximum of 200'000 points.

$$p(keep\_cell\_i) = \begin{cases} 1, & if\ LD_i < TD \\ \frac{TD}{LD_i}, & otherwise \end{cases} \quad \text{(eq. 2.4)}$$

The reason density-dependent downsampling was chosen is to avoid cases where high-density agglomerations of landmarks in a particular region accumulate multiple clusters and thus deplete lower-density regions of local reference points; density-dependent downsampling preserves the overall shape of the overlaid point cloud whilst reducing local density peaks.

Following the determination of common reference points by k-means clustering, CBE proceeds by extracting features describing the local landmark distribution around the reference points for each separate cellular point cloud. A number of such features were implemented, including the number of landmarks in the local neighborhood of reference points, the number of landmarks assigned to the reference point by the k-means clustering itself, the local density of landmarks at the reference point determined by a Gaussian Kernel Density Estimate (KDE), and the mean Euclidean distance of the 25 nearest neighbors of each reference point. The results were similar with each of these approaches and ultimately the last option (mean Euclidean distance of nearest neighbors) was chosen for all analyses based on the expressiveness of the resulting shape space and the ease of its computation.

The feature extraction described above yields an *n-by-k* latent feature space, where *n* is the number of cells and *k* the number of shared reference clusters (here *k=20*). To align individual features with

biological relevance and to allow the removal of less relevant features, this space was transformed by Principal Component Analysis (PCA).

In addition to CBE, an alternative embedding based on the moments of the PFOR or CFOR point clouds was also generated as a comparably simplistic baseline. We computed the 1st raw moments (eq. 2.5), the 2nd centralized moments (eq. 2.6) and the 3rd to 5th normalized moments (eq. 2.7) (55 features in total) and once again used PCA to re-express these features in a more compact and expressive fashion.

$$rM_{1[ijk]} = mean\left(C_z \cdot i + C_y \cdot j + C_x \cdot k\right) \quad \text{(eq. 2.5)}$$

$$cM_{2[ijk]} = mean\left(\left(C_z - rM_{1[100]}\right)^i \cdot \left(C_y - rM_{1[010]}\right)^j \cdot \left(C_x - rM_{1[001]}\right)^k\right) \quad \text{(eq. 2.6)}$$

$$nM_{m[ijk]} = \frac{cM_{m[ijk]}}{std\left(C_z - rM_{1[100]}\right)^i \cdot std\left(C_y - rM_{1[010]}\right)^j \cdot std\left(C_x - rM_{1[001]}\right)^k} \quad \text{(eq. 2.7)}$$

In both equations, C_d is the array of all point cloud coordinates along the spatial dimension d, M_m is the set of raw (rM_m), centralized (cM_m) or normalized (nM_m) moments of the m-th order, and $[i, j, k]$ includes all combinations of length 3 drawn from the integer range $[0, \dots, m]$ that satisfy $i + j + k = m$. All array operations are element-wise and $mean(\dots)$ and $std(\dots)$ compute the mean and standard deviation of a given array.

2.3.7 Prediction and Visualization of Morphological Archetypes

Four morphological archetypes were manually annotated in 26 primordia (see Fig. 3.17a), yielding 93 leader cells, 241 outer rosette cells, 182 inner rosette cells and 108 between-rosette cells (624 cells in total). Only the most clear examples of the respective archetypes were labeled.

Cell archetype prediction was performed using a Support Vector Classifier (*sklearn.svm.SVC*) with a Radial Basis Function (RBF) kernel, using the PFOR embedding of cell shape as input features. We optimized the SVC hyperparameters using a grid search with 5-fold cross-validation over 5 orders of magnitude surrounding the scikit-learn default values, settling ultimately on *C=1.0* (penalty) and *gamma=0.05* (RBF kernel coefficient). The defaults were used for all other settings.

The confusion matrices in figure 3.17b were produced by randomly splitting the annotated cells into a training set (436 cells) and a test set (188 cells). Predictions for the entire dataset (16'974 cells) were generated following training with all 624 manually annotated cells.

The archetype space was constructed by inferring the classification probabilities for each class (using *sklearn.svm.SVC.predict_proba*) and performing a PCA on them. The 3D and 2D visualizations in figure 3.18 were then generated by plotting the first three or the first two principal components, respectively.

2.3.8 Multi-Channel Atlas Prediction

To construct multi-channel atlases, two-color stacks of the green membrane marker *claudinB:: Lyn:EGFP* and the red or far-red secondary markers listed in section 2.2.7 were collected.

For the selective registration approach, ISLA and PFOR-CBE were performed on the segmentation masks and for each target cell the 10 nearest neighbors were retrieved from each of the secondary channels' shape spaces using *scipy.cKDTree.query*. An open source python implementation of the Iterative Closest Point (ICP) algorithm was used [retrieved from *github.com/ClayFlannigan/icp* on 13.08.2018 and used under the Apache License version 2.0] to register the segmentation-derived ISLA point clouds of the selected cells. The 5 cells with the highest loss after segmentation were discarded and the ISLA point clouds derived from the secondary channel of the remaining 5 cells were transformed to match the registration.

To reconstruct images from registered point clouds, they were first scaled back from real scale (µm) into the original image scale (voxels) and transposed to align with the bounding box of the target cell. Next, Gaussian Kernel Density Estimation (KDE) was performed using *scipy.stats.gaussian_kde*, creating a reconstructed intensity image. Finally, the target cell's segmentation mask was used to paste this reconstruction back into the correct voxels in the target cell's primordium.

For the machine learning approach, embedded spaces were extracted for all cells and all channels. Because expression of the secondary markers was sometimes heterogeneous across the primordium, only cells with a secondary marker intensity above the 33rd percentile were used as training data. Latent features extracted from the segmentation mask were used as input features and latent features extracted from the secondary channels were the target values for which regressors were being trained. To select the best machine learning model, the following regressors were tested using 5-fold cross-validation: k-nearest neighbors regression (*sklearn.neighbors. KNeighborsRegressor*), random forest regression (*sklearn.ensemble.RandomForestRegressor*), elastic net regression (*sklearn.linear_model. ElasticNet*), Lasso regression (*sklearn.linear_model.Lasso*), a multi-layer perceptron (*sklearn.neural_network. MLPRegressor*), and a support vector regressor with an RBF-kernel (*sklearn.svm.SVR*). Hyperparameters were optimized on the *NLS:tdTomato* nuclear marker using a 5-fold cross-validated grid search of 5 orders of magnitude surrounding the scikit-learn defaults. Performance was evaluated and optimized across different secondary channels and latent feature embeddings, with the primary aim being high explained variance but also giving some consideration to computational efficiency (training and prediction time).

The selected final model is the SVR regressor with hyperparameters *C=10.0* (penalty), *gamma=0.005* (RBF kernel coefficient) and *epsilon=0.1* (penalty-free epsilon-tube). It was trained for each secondary channel on all available data for that channel and then applied to predict that channel's embedded space for all other cells.

2.3.9 Data Visualizations

Micrographs were visualized using Fiji. Scale bars were added with "*Analyze>Tools>Scale Bar*". All other plots were created with the python tools described in section 2.3.1.

The "exploded view" of the segmented pLLP (Fig. 3.8b) was generated by first determining the centroids of each segmented cell and then shifting them apart by scaling their x and y coordinates by a single user-specified factor. Next, an appropriately up-scaled empty image stack was created and populated one by one with the segmented cells, which were copied from the original image to the scaled position of their centroid. In this way, the cells are shifted apart uniformly but are not themselves scaled or otherwise transformed. The python code for generating this visualization is called *tissueRipper* and is available as open source at *github.com/WhoIsJack/tissueRipper* (note: *WhoIsJack* is the GitHub alias of the author of this thesis).

The correlation heatmaps (Fig. 3.13a-b) were created using seaborn's *clustermap* function with the pairwise Pearson correlation coefficients between all engineered and latent features. The corresponding bigraphs (Fig. 3.31c-d) were generated using a custom plotting function based on the networkx module. The edges were colored according to the signed value of the Pearson correlation coefficient and sized according to the absolute value. Edges with an absolute correlation coefficient smaller than 0.3 were omitted. The nodes of the engineered features were sorted to reduce edge crossings and group similar nodes, which was achieved by minimizing the following custom loss function:

$$loss = \sum_{i=0}^{f_E} \sum_{j=0}^{f_L} \left| \frac{O_i}{f_E} - \frac{j}{f_E} \right| \cdot \left| pcc\left(E_i, L_j\right) \right| \qquad \text{(eq. 2.8)}$$

where f_E and f_L are the number of engineered and latent features, respectively. O is the given sort order of the engineered features, i.e. any permutation of the integer interval $[0, f_E]$. Finally, $\left| pcc\left(E_i, L_j\right) \right|$ is the absolute Pearson correlation coefficient of the values of the i-th engineered feature and the j-th latent feature. In essence, this loss function is the sum of all Euclidean rank distances between engineered and latent features, weighted by their corresponding absolute Pearson correlation coefficients. Minimization was performed by random shuffling of the sort order and retaining only shuffles that reduced the loss until no change was observed for 2000 consecutive shuffles.

Due to the high dimensionality of the single-cell analysis data, many of the visualizations used during data analysis were interactive, allowing the analyst to select the particular features or dimensions to visualized "live" within Jupyter notebooks. Interactivity was mainly implemented using ipywidgets and in some cases using matplotlib's interactivity features. As the format of this thesis does not support interactive visualizations, the most relevant or representative plot specifications were chosen for the various figures.

2.3.10 Statistical Analysis

Unless otherwise stated, the following definitions and statistical methods were used.

N refers to the number of embryos/primordia and *n* to the number of cells.

Statistical significance for comparisons between two conditions was estimated without parametric assumptions using a two-tailed Mann-Whitney U test (*scipy.stats.mannwhitneyu* with keyword argument *alternative='two-sided'*). For linear regressions, the significance of the slope being non-zero was estimated using a two-tailed Wald test with t-distribution of the test statistic (in *scipy.stats.linregress*). In figures, significance is indicated as **ns** ($p \geq 0.01$), ***** ($p < 0.01$), ****** ($p < 0.001$) or ******* ($p < 0.0001$).

Significance tests with large sample sizes such as those encountered during single-cell analysis tend to indicate high significance regardless of whether the difference between populations is substantive or technical [Sullivan & Feinn, 2012], which is why we report effect sizes rather than statistical significance in such cases. This applies in particular to the quantitative analysis of chemoARHGEF (section 3.1.4). Effect size was estimated using Cohen's d [Cohen, 1988], which was computed using a custom python function. Resulting values can be described as *no effect* ($d \approx 0.0$), a *small effect* ($d \approx 0.2$), a *medium effect* ($d \approx 0.5$) or a *large effect* ($d \approx 0.8$) [Cohen, 1988].

3 Results

3.1 Adapting Optogenetic Tools for the Precise Perturbation of Chemokine and Actin Signaling in the pLLP

The potential power of optogenetics as a tool for minimal and precise perturbation of biological systems was introduced in section 1.3.1. Here, we sought to adapt existing optogenetic tools for use in the pLLP and to employ them to study the interplay of migration and morphogenesis.

3.1.1 Selection and Overview of Optogenetic Tools

We chose to adapt tools that had been shown to work in other systems and that were specifically designed to manipulate chemokine signaling (and thus migrational polarity) or actin organization and dynamics (and thus force generation and cellular architecture). We reasoned that control over these two key players would allow us to perform a broad range of experiments from domain-specific activation to single-cell perturbation and to observe the consequences both within directly perturbed cells and within unperturbed cells in the same tissue, which together would reveal even the most intricate interrelations of chemokine signaling and tissue mechanics in the primordium.

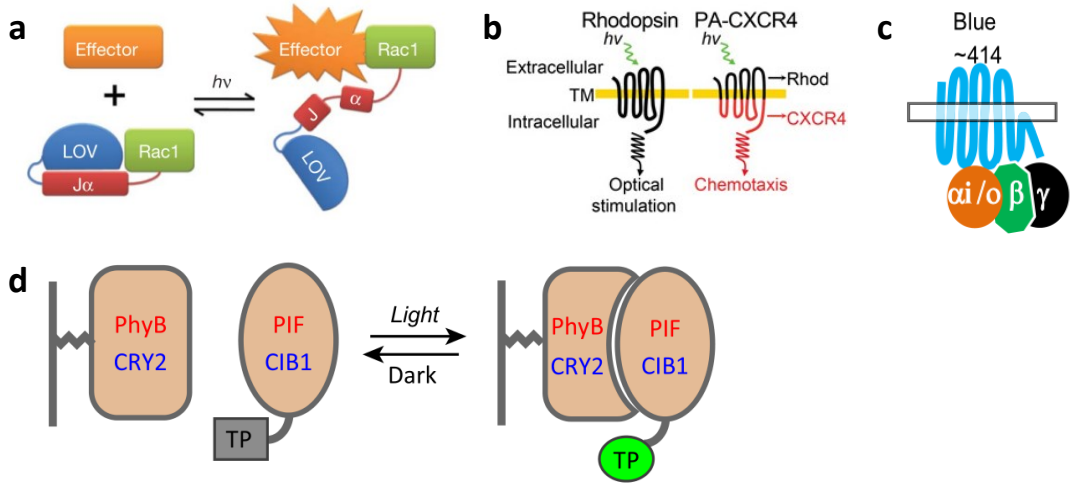

Figure 3.1: Optogenetic tools adapted to control chemokine and actin signaling.

(a) PA-Rac1: The LOV-Jα domain blocks the active site of constitutively active Rac1 in the dark, preventing its interaction with effectors (left). Light induction releases the block and effectors are activated (right). Reproduced from [Wu et al., 2009]. (b) Rhodopsin is a naturally occurring light-sensitive GPCR (left). PA-CXCR4 was created by exchanging rhodopsin's intracellular domains for those of CXCR4 (right). Reproduced from [Xu et al., 2009]. (c) bOpsin is a retinal blue light-sensitive GPCR that happens to signal through $G\alpha_i$, which is why it can be repurposed directly to mimic chemokine signaling in non-retinal cells. Reproduced from [Karunarathne et al., 2013a]. (d) Light-dependent heterodimers such as PhyB-PIF or CRY2-CIB1 can be used to optically control Target Proteins (TPs) through recruitment. Here, one of the dimerization partners is tethered to the cell membrane and the other is recruited upon light induction, bringing along a target protein that is inactive in the cytoplasm but binds and activates interaction partners once localized to the membrane. Adapted from [Zhang & Cui, 2015].

The relevant optogenetic tools can be loosely categorized into two types: single-component systems based on conformational change (Fig. 3.1a-c) and two-component systems based on recruitment (Fig. 3.1d) [Tischer & Weiner, 2014; Karunarathne et al., 2015; Repina et al., 2017].

In single-component systems, optical stimulation of a light-sensitive domain leads to a specific conformational change, activating the biochemical function of the protein in question. We adapted three such proteins to the lateral line: a photoactivatable version of the small GTPase *Rac1*; PA-Rac1 [Wu et al., 2009], a light-sensitive chimera of *rhodopsin* and *Cxcr4b*; PA-Cxcr4b [Xu et al., 2014], and a naturally light-sensitive receptor capable of activating the *Cxcr4* pathway; bOpsin [Karunarathne et al., 2013a].

PA-Rac1 is a constitutively active form of human Rac1 fused to the photosensitive LOV2-Jα domain of *Avena sativa* phototropin1 [Wu et al., 2009]. The LOV2-Jα domain sterically blocks the active site of Rac1 in the dark but releases it upon blue light illumination, allowing Rac1 to bind and activate its effectors (Fig. 3.1a) [Wu et al., 2009]. PA-Rac1 has been used to drive lamellipodia formation and motility in cultures of HeLa cells and mouse embryonic fibroblasts [Wu et al., 2009], in *Drosophila* border cells [Wang et al., 2010] and in zebrafish neutrophils and neural crest cells [Yoo et al., 2012; Scarpa et al., 2015].

PA-CXCR4 is a chimeric bovine rhodopsin, a naturally light-sensitive GPCR whose intracellular domains have been replaced by human CXCR4 to change rhodopsin's $G\alpha_t$ activity into CXCR4's $G\alpha_i$ activity (Fig. 3.1b) [Xu et al., 2014]. This form of PA-Cxcr4b has been used to polarize T-cells and to guide their migration [Xu et al., 2014]. To use it in the pLLP we modified the human CXCR4-domains to match those of zebrafish *Cxcr4b*, yielding PA-Cxcr4b[1].

Finally, bOpsin (short for blue opsin) is a human retinal opsin that has been found to be capable of signaling through $G\alpha_i$ (Fig. 3.1c) and has been repurposed to polarize HeLa cells and to induce neurite extension in rat hippocampal neurons [Karunarathne et al., 2013a] as well as to guide immune cell migration *in vitro* [Karunarathne et al., 2013b].

Recruitment-based tools make use of protein pairs that heterodimerize in response to optical stimulation. As a general strategy, one interaction partner is fused to a targeting domain which localizes it to a compartment of interest such as the cell membrane. The other partner is coupled to an effector protein that has no downstream targets in the cytoplasm and is thus inactive in the dark. Upon optical activation, the cytoplasmic component binds the pre-localized component and the effector – now recruited to its site of function – activates its downstream targets (Fig. 3.1d).

To bring recruitment-based tools to the lateral line, we adapted the CRY2-CIBN system (Fig. 3.1d) [Kennedy et al., 2010], which at the time was the most easy to use and most widely adopted

[1] This construct was designed by a former member of the lab, Erika Donà.

recruitment-based optogenetic tool [Tischer & Weiner, 2014; Zhang & Cui, 2015]. Indeed, the *Arabidopsis*-derived CRY2-CIBN heterodimerization system has been successfully used to optically trigger transcription [Hughes et al., 2012; Konermann et al., 2013], intracellular signaling [Zhang et al., 2014], and more recently morphogenesis by apical constriction [Izquierdo et al., 2018].

Here, we focused on inducing translocation to the cell membrane by recruitment of optically activated CRY2-effector fusion proteins to membrane-anchored CIBN. As effectors, we used two Rho GTPase GEFs to activate RhoA and Rac1, respectively, as was previously done in murine fibroblasts with the PhyB-PIF heterodimerization system [Levskaya et al., 2009], and we used a Gα-GAP and a βγ-sequestering protein to inhibit *Cxcr4b* signaling, as established previously in murine macrophages [O'Neill & Gautam, 2014].

3.1.2 PA-Rac1, PA-Cxcr4b and bOpsin are Aphenotypic in the pLLP

We generated stable transgenic fish lines expressing PA-Rac1, PA-Cxcr4b and bOpsin in the pLLP under control of a Gal4-UAS transactivation system. All three constructs were expressed and localized to the cell membranes as expected, although bOpsin expression was comparatively weak (Fig. 3.2a).

To test light-dependent activity, we kept embryos in the dark prior to imaging and then used confocal laser light at activating wavelengths to stimulate the entire tissue either briefly or overnight, or to briefly stimulate a small region of interest such as single cells or subcellular membrane regions. As readouts, we first followed either short-term membrane dynamics or long-term pLLP migration using a membrane label. However, we did not observe any measurable effect on either process for any of the three optogenetic constructs (Fig. 3.2b-c). Next, we visualized actin dynamics using the F-actin probe *UtrCH* [Burkel et al., 2007] during stimulation of all three constructs but found no discernible active response (Fig. 3.2d). Finally, we sought to test whether stimulation of PA-Cxcr4b or bOpsin would lead to the depletion of the G-protein Gγ9 from the cell membrane, a measure that has been reported as a readout for GPCR activity [Saini et al., 2007], including bOpsin activity specifically [Karunarathne et al., 2013a]. However, although Gγ9 does localize to membranes in the pLLP, we could not detect any response to PA-Cxcr4b or bOpsin stimulation (data not shown). In fact, transgenic expression of *mNeonGreen:Gγ9* itself caused a deceleration and in some cases an arrest of pLLP migration, possibly due to an inhibitory effect on *Cxcr4b* signaling or alternatively through an unknown toxic effect.

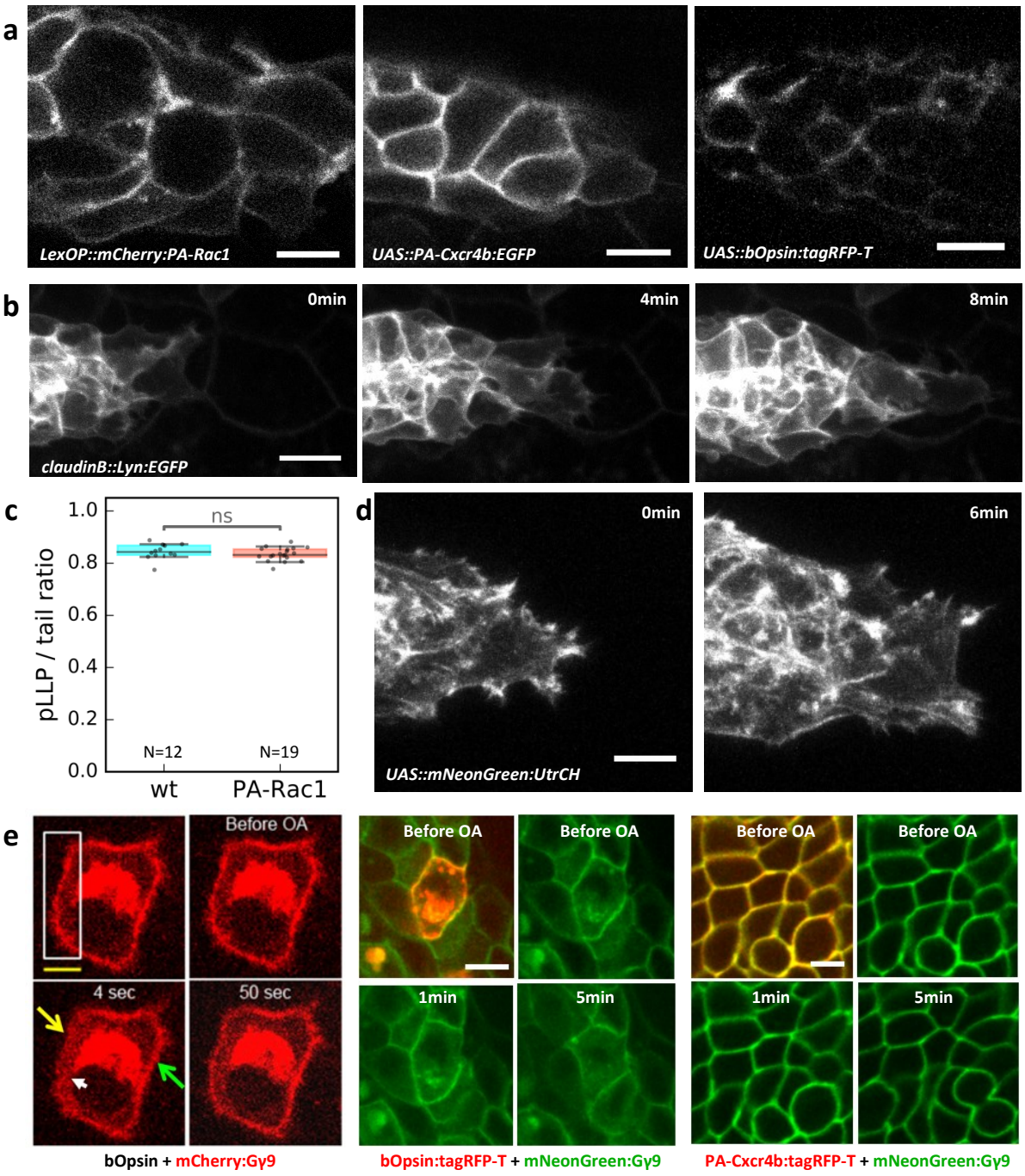

Figure 3.2: Example data from unsuccessful tests of PA-Rac1, PA-Cxcr4b and bOpsin optogenetics.

(a) Expression and localization of PA-Rac1 (left; scale bar: 10µm), PA-Cxcr4b (middle; scale bar: 7µm) and bOpsin (right; scale bar: 12µm) in subsections of the pLLP. (b) Stills from a movie of a pLLP expressing *bOpsin:tagRFP-T* [not shown] and the membrane marker *Lyn:EGFP*, displaying no evident alterations in leader cell membrane dynamics upon bOpsin stimulation. (c) At 48hpf, primordia expressing PA-Rac1 have migrated just as far as wild-type primordia relative to the total length of the embryo's tail (p=0.094), despite overnight illumination with activating light. (d) Stills from a movie of F-actin visualized by *mNG:UtrCH* during activation of *PA-Cxcr4b:tagRFP-T* [not shown], indicating no evident change of actin dynamics in response to activating light. Scale bar: 7µm. (e) G-protein Gγ9 immediately translocates away from the membrane upon bOpsin stimulation in HeLa cells and swiftly recovers thereafter (left) [reproduced from Karunarathne et al., 2013a] but does not respond even to continuous global activation of bOpsin (middle) or PA-Cxcr4b (right) in early zebrafish embryos (OA: Optical Activation). Time in left panel denotes time since a brief activation pulse, time in middle and right panels denotes continuous activation. Scale bars: 10µm. Data acquired jointly with Sruthi Raja.

To check whether the lack of observable optogenetic activity is specific to the lateral line primordium or to the stable transgenic lines we generated, we also injected mRNA of all three optogenetic constructs into zygotes and tested optical stimulation in early embryos undergoing gastrulation (approx. 8-10hpf). Consistent with our results in the pLLP, all constructs were capable of localizing to the membrane but we did not detect any distinctive phenotype in either membrane dynamics, actin dynamics, Gγ9 translocation, or long-term survival and development (see representative Gγ9 data in Fig. 3.2e; other data not shown).

3.1.3 The CRY2-CIBN System is Functional but CRY2-Effector Fusions are not Readily Expressible in Zebrafish

We engineered a set of membrane-bound CIBN constructs and tested whether they were capable of recruiting cytoplasmic *CRY2:tagRFP-T* or *tagRFP-T:CRY2* upon blue light stimulation. Both mRNA and DNA injections imaged in early embryos (approx. 10hpf) or in the pLLP showed high expression and membrane localization of CIBN constructs and lower but readily detectable cytoplasmic and nuclear expression of CRY2 constructs (Fig. 3.3a), although CRY2-positive clones in the pLLP were exceedingly rare.

Blue-light stimulation (but also ambient-light exposure, see Materials & Methods, section 2.2.2) induced the rapid recruitment of the cytoplasmic pool (but not the nuclear pool) of both CRY2 constructs to the cell membrane in cells expressing CIBN, both in early embryos (Fig. 3.3a-b) and in the pLLP (Fig. 3.3c). Stimulation was also possible with single-cell precision by illuminating only a small region of interest (Fig. 3.3d), whereas distinct sub-cellular recruitment to only one particular region of the membrane was not achievable, possibly due to the rapid cytoplasmic diffusion of activated CRY2 constructs.

Given the success of these test runs, we went on to design four effector-coupled CRY2 constructs, all based on effectors that had been used successfully in optogenetics studies in literature.

Two were designed to inhibit GPCR signaling and therefore chemokine signaling through *Cxcr4b*. *UAS::CRY2:tagRFP-T:GRK3ct* features the C-terminal domain of bovine *G protein–coupled receptor kinase 3* (GRK3), which selectively sequesters Gβγ complexes and inhibits their activity [Hollins et al., 2009; O'Neill & Gautam, 2014]. *UAS::CRY2:tagRFP-T:RGS4Δ* is based on *Regulator of G protein signaling 4* (RGS4), a GAP that inhibits $G\alpha_i$ and $G\alpha_q$ signaling and can be prevented from natively localizing to the membrane by truncating away the first 33 amino acids (hence RGS4Δ) [O'Neill & Gautam, 2014].

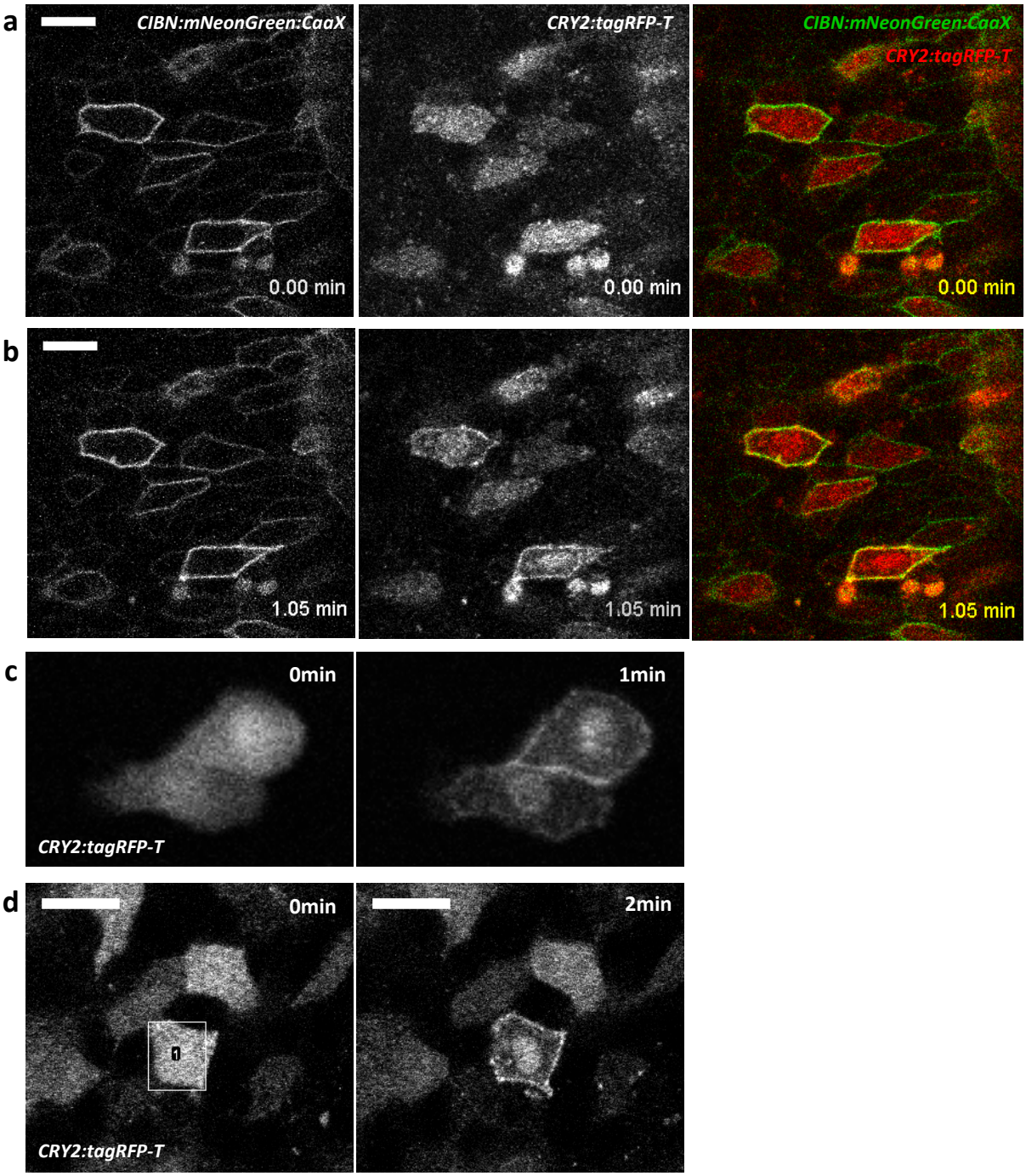

Figure 3.3: Optogenetic membrane recruitment with the CRY2-CIBN system.

(a) Expression and localization of *bAct::CIBN:mNeonGreen:CaaX* (left), *bAct::CRY2:tagRFP-T* (middle) or both (right) in co-injected early embryos (approx. 10hpf). Scale bar: 20µm. (b) The same cells after a minute of imaging; the 488nm laser used to image *CIBN:mNeonGreen:CaaX* has activated *CRY2:tagRFP-T* and induced its recruitment to the membrane. (c) pLLP cells positive for injected *UAS::CRY2:tagRFP-T* in embryos carrying *UAS::Lyn:CIBN* and the Gal4 driver as stable transgenic background. CRY2 is cytoplasmic prior to illumination with 488nm laser light (left) and gets recruited to the membrane afterwards (right). Scale bar: 10µm. (d) Early embryos (approx. 10hpf) injected with *bAct::Lyn:CIBN* and *bAct::CRY2:tagRFP-T*. The initially cytoplasmic CRY2 is illuminated with 488nm laser light within a specific region of interest (left, region 1), which induces cell-specific membrane translocation (right). Scale bars: 10µm.

The other two constructs were designed to activate *Rac1* and *RhoA* activity, respectively, and were based on the RacGEF *Tiam*, which has previously been used to control *Rac1* using the PhyB-PIF optogenetic recruitment system [Levskaya et al., 2009], and on the RhoGEF *ARHGEF25*, which has been used to activate *RhoA* in the context of a rapamycin-based chemical recruitment system [Van Unen et al., 2015]. For both, we identified the closest zebrafish homologues, predicted the catalytically active GEF domains, and cloned them to generate *UAS::CRY2:tagRFP-T:Tiam1a-Gd* and *UAS::CRY2:tagRFP-T:ARHGEF25b-Gd* (where Gd is short for GEF domain).

We generated stable transgenic lines of *Lyn:CIBN* and the four effector-coupled CRY2 constructs. However, we were unable to derive transgenic embryos that express any of the CRY2-effector fusion proteins at a detectable level in the lateral line. More specifically, we found that only an unusually small number of fish injected with a CRY2 effector construct produced any transgenic progeny (approx. 2% as opposed to 10-50% with *Lyn:CIBN* or other constructs), that those select few founder fish produced unusually few transgenic offspring per clutch (approx. 5-10%), and that those transgenic offspring showed unusually low expression of the co-injection marker (*clmc2::EGFP*; green heart marker). Crucially, although some of these rare positive embryos sporadically expressed very low levels of the CRY2 construct, none showed any detectable expression in the posterior lateral line primordium (data not shown).

As the same observations were made independently by other members of the lab[1] using CRY2 constructs with very different types of effectors, we concluded that CRY2 is not expressible in stable zebrafish transgenic lines.

3.1.4 A Constitutively Active Version of the RhoGEF ARHGEF25b Causes Cell Rounding and Sorting in the pLLP

Originally intended as positive controls for the CRY2 optogenetic approach, we engineered versions of the RhoA and Rac1 effectors that are directly targeted to the membrane (*Lyn:tagRFP-T:ARHGEF25b-Gd* and *Lyn:tagRFP-T:Tiam1a-Gd*) and therefore are expected to act in a constitutively active manner. We expressed these constructs in the pLLP under the chemically inducible LexPR-LexOP transactivation system [Emelyanov & Parinov, 2008], which allows tissue-specific and temporally controlled expression.

Contrary to the CRY2 fusions, we were readily able to generate transgenic zebrafish lines expressing these constructs in the lateral line primordium upon chemical induction of the LexPR transactivator. Whilst cells expressing *tagRFP-T:Tiam1a-Gd* seemed completely unaffected, expression of *Lyn:tagRFP-T:ARHGEF25b-Gd* caused cells to assume a more rounded shape (Fig. 3.4a). Interestingly, this increase in cell sphericity was accompanied by a positional phenotype: individual cells expressing *Lyn:tagRFP-T:ARHGEF25b-Gd* are predominantly localized to the primordium's center rather than its

[1] Specifically Mie Wong and Elisa Gallo.

periphery (Fig. 3.4a). As we had never before observed such a positional phenotype, we decided to investigate this effector more closely, terming the construct *chemoARHGEF* in reference to the chemical control of its expression.

Making use of the initial mosaicism of LexPR-driven expression, we quantitatively analyzed the effects of chemoARHGEF at the single-cell level. We acquired and automatically segmented (see section 3.3.2) high-resolution 3D volumes of primordia co-expressing a uniform membrane label (used for segmentation) alongside either chemoARHGEF or *LexOP::NLS:mIRFP*, an inert control construct expressed under the same promoter (Fig. 3.4b). Consistent with our initial qualitative observations, we found that high levels of chemoARHGEF expression but not *NLS:mIRFP* expression are associated with substantially increased cell sphericity (Fig. 3.4c) and with a location bias toward the center of the tissue (Fig. 3.4d).

Since a more central cell location is itself associated with increased cell sphericity in wild-type primordia (Fig. 3.4e), we considered the possibility that the location bias of chemoARHGEF might be sufficient to explain its effect on cell sphericity. However, by performing a resampling analysis of the single-cell segmentation data that compares centrally located chemoARHGEF cells to centrally located wild-type cells (see Materials & methods, section 2.3.3), we were able to exclude this possibility (Fig. 3.4f); the increased sphericity observed in chemoARHGEF-positive cells goes beyond what would be expected from wild-type cells located centrally in the pLLP.

We hypothesized that the observed central location bias of chemoARHGEF-positive cells might result from surface tension-based cell sorting [Brodland, 2002], although alternative explanations are also possible. For instance, chemoARHGEF could be specifically repressed or rapidly degraded in peripheral cells. We first attempted to clarify which mechanism is at play by performing time lapse microscopy, finding that the first cells showing a detectable *tagRFP-T* signal were generally already located in the center. However, there is a considerable delay between chemical induction and the point in time when the *tagRFP-T* signal becomes visible, likely due to the fluorophore's long maturation time [Shaner et al., 2008], so sorting may simply occur before expression is visible.

To unambiguously resolve this issue, we made use of the fact that the mosaic expression patterns of chemoARHGEF and the inert control construct *NLS:mIRFP* are largely overlapping in embryos carrying both constructs simultaneously (Fig. 3.4g). We found that in such crosses chemoARHGEF imparts both the central location bias (Fig. 3.4h) and the increase in cell roundness (data not shown) onto the cells positive for *NLS:mIRFP*. Thus, we conclude that the location bias of chemoARHGEF is not due to specific repression or degradation in peripheral cells but must be due to a cell sorting process that sorts chemoARHGEF-positive cells to the center of the tissue.

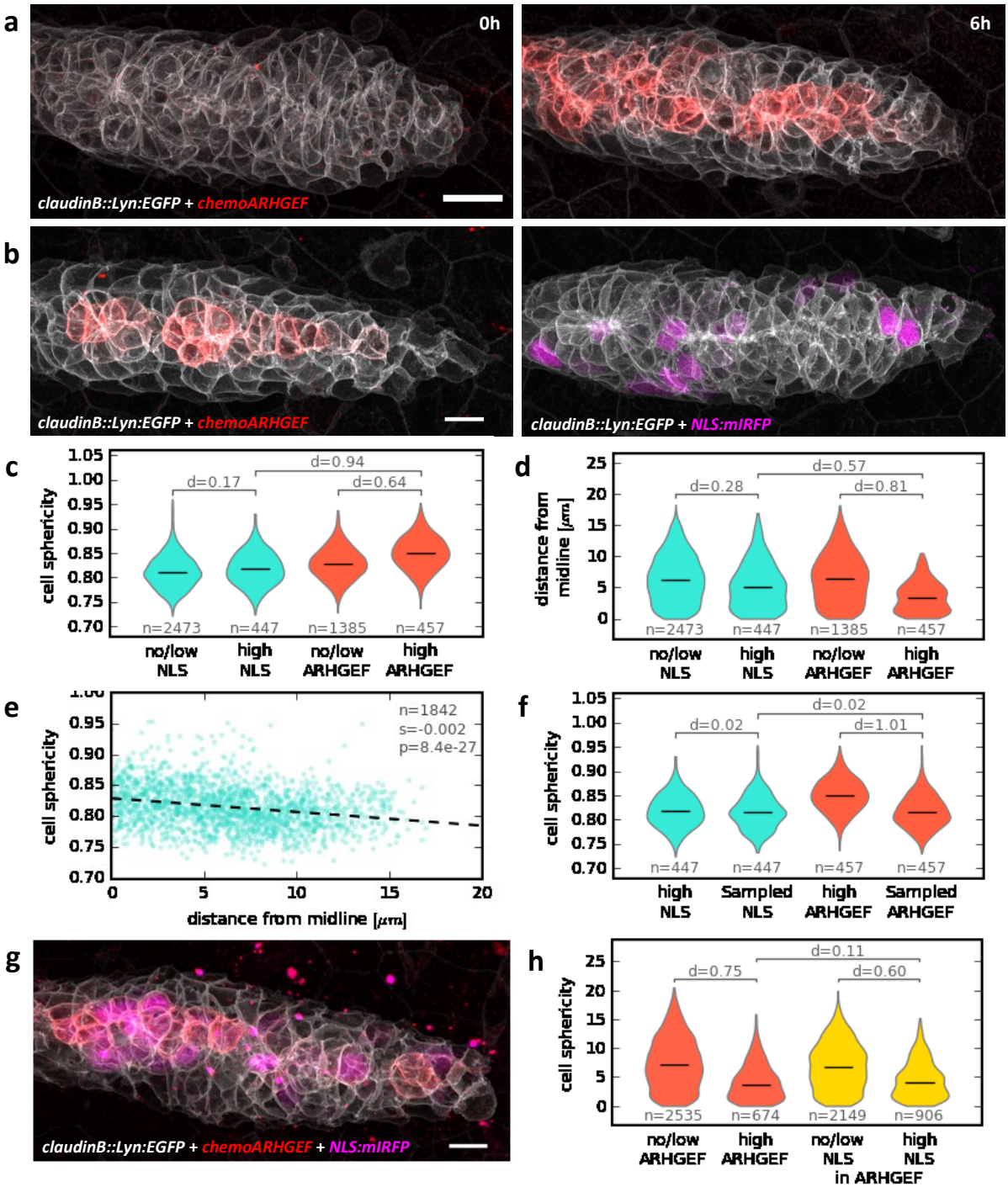

Figure 3.4: *Lyn:tagRFP-T:ARHGEF25b-Gd* causes cell rounding and sorting toward the pLLP's center.

(a) Stills from a movie of chemoARHGEF induction. Scale bar: 15µm. (b) Example images of chemoARHGEF and an inert construct, *NLS:mIRFP*, expressed under the same promoter. Scale bar: 10µm. (c) Comparison of cell sphericity in primordia expressing *NLS:mIRFP* (cyan, N=26) or chemoARHGEF (red, N=20), grouped by cells with no/low expression and high expression. (d) Comparison of cell location relative to the pLLP's front-to-back midline, showing the central location bias of cells expressing chemoARHGEF. (e) Wild-type embryos exhibit a weak correlation of cell sphericity and proximity to the center (s=slope, p=p-value for non-zero slope test). (f) Comparison of cells expressing chemoARHGEF (high ARHGEF) to wild-type cells sampled to match their spatial distribution (Sampled ARHGEF), showing that wild-type cells located just as centrally as chemoARHGEF-positive cells are still far less spherical. Resampling cells based on high *NLS:mIRFP* expression (cyan) has no effect. (g) Crossing chemoARHGEF and *NLS:mIRFP* lines yields mosaic co-expression in the same cells. Scale bar: 10µm. (h) The location bias of chemoARHGEF (red, N=23) is imposed on the *NLS:mIRFP* pattern (yellow, N=25) if the two are co-expressed. In all plots: n=number of cells, d=effect size (Cohen's d).

Taken together, our preliminary results indicate that chemoARHGEF is a useful tool for manipulating cell surface tension and that pLLP cells are readily able to undergo tension-based sorting. Further work aimed at elucidating the role of differential surface tension during rosette formation in wild-type primordia is currently ongoing.

3.2 Studying Tissue-Scale Feedback of Morphogenetic Remodeling on Collective Cell Migration in the pLLP

In order to generate the desired final organ arrangement, migrating collectives such as the pLLP need to coordinate their motility and directionality with follower cell morphogenesis and differentiation, as detailed in sections 1.4.3 and 1.4.4. However, little is known about how this coordination is achieved, both in the pLLP and beyond.

Given that the polarity of migrating cells can be controlled not only by external chemokine gradients (chemotaxis) but also by internal mechanical stress (plithotaxis) [Trepat & Fredberg, 2011] (see section 1.4.3) and given that the processes of rosette assembly and proneuromast deceleration during deposition are bound to have an effect on tissue mechanics, we hypothesized that follower cell morphogenesis feeds back on leader cell motility and directionality via plithotaxis (Fig. 3.5a-b).

Such a mechanism could not only serve as a means of coordination but might also explain previous observations of the remarkable robustness of migrational polarity in lateral line cells. For instance, cells of the pLLP can establish directional polarity even in the absence of the chemokine receptor *Cxcr4b*, so long as there are some *Cxcr4b*-positive cells at the leading edge [Haas & Gilmour, 2006] (see section 1.5.2). Plithotaxis based on the pulling forces generated by those few leader cells could be the mechanism by which the rest of the tissue is polarized. Similarly, laser ablation of the rearmost proneuromast, which harbors most of the primordium's supply of the scavenger receptor *Cxcr7* and thus should be essential in maintaining the chemokine gradient, does not lead to an arrest of the remaining primordium – unless all nascent rosettes are removed, in which case the leader cells cease directional migration [unpublished data[1]]. This outcome can easily be explained if the mechanical stress generated by the juxtaposition of motile leaders and rosette-assembling followers generates a mechanical cue that maintains leader cell polarity through plithotaxis (Fig. 3.5c-d).

Here, we sought to reproduce and further refine these laser ablation experiments and to directly test our hypothesis by specifically manipulating follower cell compliance and observing the response of leader cells both qualitatively and by quantitative computational image analysis.

3.2.1 pLLP Migration is Robust to the Acute Ablation of the *Cxcr7* Domain

To confirm that pLLP migration can indeed proceed even if the scavenger receptor *Cxcr7* is acutely removed (Fig. 3.5c-d), we used a previously established transcriptional reporter line based on *Cxcr7* BAC transgenesis (see Materials & Methods, section 2.2.4) to identify which cells in the primordium are expressing *Cxcr7*. We crossed this reporter to a uniform membrane marker and used pulsed UV laser ablation to precisely destroy the cells in the rear of the primordium, up to and including the

[1] These experiments were performed by two former members of the lab, Petra Haas and Sebastian Streichan.

frontal-most cells expressing the transcriptional reporter, which did not include the frontal-most rosette or the leader-follower transition zone (Fig. 3.5e).

Time course imaging of pLLP migration following laser ablation showed heterogeneous outcomes (Fig. 3.5f): in some cases the primordium continued migrating as if unperturbed whereas in others it temporarily slowed down to different degrees before picking up speed again. In rare cases, pLLP migration was abolished completely, likely due to damage to the surrounding tissue.

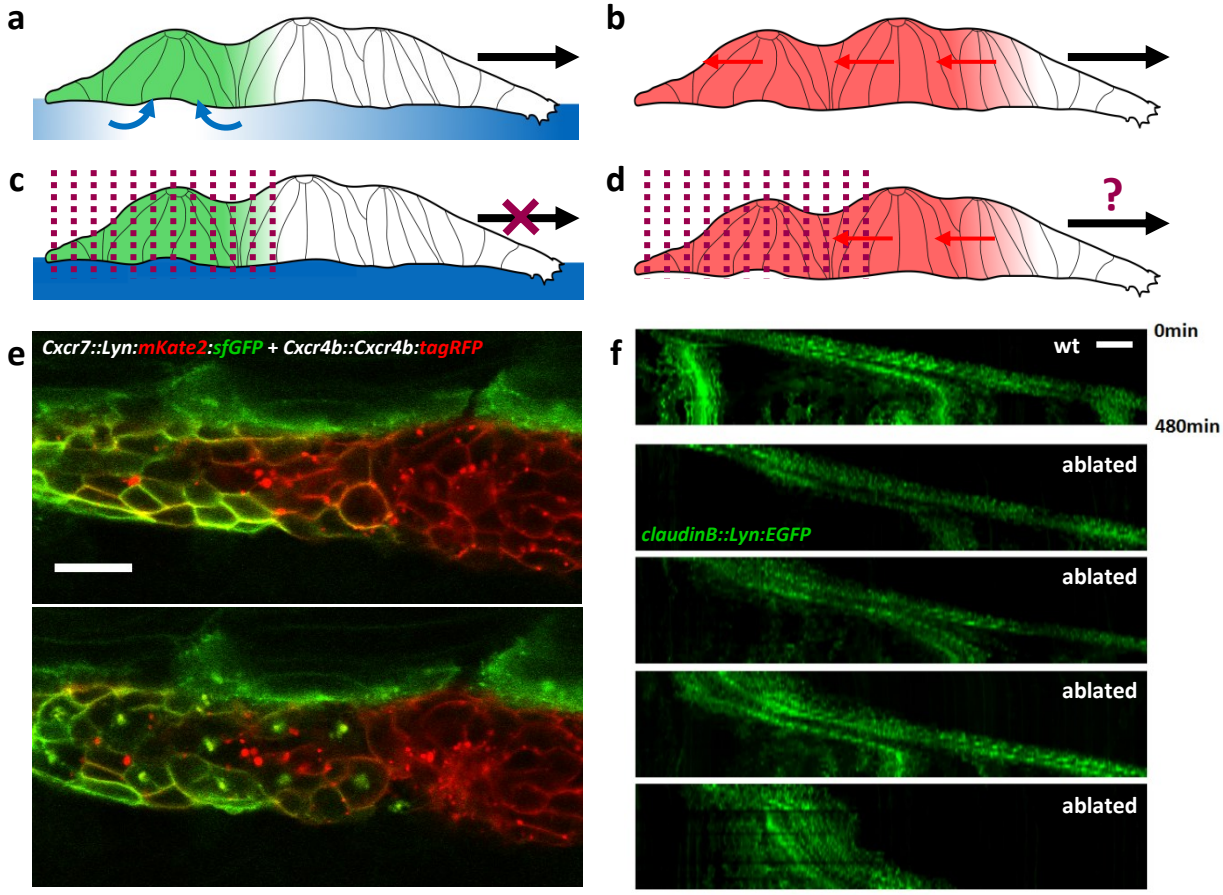

Figure 3.5: pLLP migration is robust to precise and acute ablation of the *Cxcr7* domain.

(a) Illustration of how leader cell migration (black arrow) is guided by a chemokine gradient generated through scavenging of extracellular *Cxcl12a* (blue) by follower cells expressing the decoy receptor *Cxcr7* (green) [Donà et al., 2013]. (b) Complementary model where directionality is (also) maintained through plithotaxis: rosette assembly in follower cells (red) provides a counter-force (red arrows) to leader cell motion, which aligns the axis of maximum mechanical stress to the front-back axis of the tissue and thus guides leader cell directionality. (c) Ablation of *Cxcr7*-expressing cells should lead to the abolishment of the chemokine gradient and – if the gradient was the only cue for leader cell migration – to an arrest of the primordium. (d) However, if plithotaxis is also capable of maintaining leader cell polarity, pLLP migration should be robust to this perturbation, so long as the remaining followers can provide sufficient counter-force. (e) 2D slice of the back of a primordium with all pLLP cells labeled in red and cells expressing *Cxcr7* specifically labeled in green, before (top) and after (bottom) pulsed UV laser ablation. Although UV ablation destroys the targeted cells (and sometimes their direct neighbors), damage to the surrounding tissue is usually minimal, as seen here. Scale bar: 15µm. (f) Representative kymographs showing wild-type migration (top) and the range of possible consequences of *Cxcr7*-domain ablation, from virtually no effect to temporary deceleration to complete arrest (top to bottom, ordered by severity). Scale bar: 40µm. The sketched outline of the pLLP used in (a-d) was adapted from [Donà et al., 2013].

Importantly, the observation that some primordia are capable of unimpeded migration despite ablation of the *Cxcr7*-expressing domain confirms that the pLLP is to some extent robust to acute perturbation the self-generated chemokine gradient.

3.2.2 Effects of Selective Follower Arrest on Leader Cell Behavior

If tissue-scale stress generated by follower cell deceleration feeds back on leader cell directional polarity, experimentally enhancing follower cell deceleration should lead to a noticeable response in leader cells (Fig. 3.6a-b).

We first attempted to perform this experiment using a caged version of blebbistatin, azido-blebbistatin [Képiro et al., 2015], but initial tests with the drug proved unsuccessful (see Materials & Methods, section 2.2.5). However, during these tests we serendipitously discovered that medium-intensity 2-Photon irradiation at 800nm induces an arrest of cell motility in the pLLP within tens of minutes without causing visible tissue damage. This phenomenon reliably occurs on completely wild-type primordia, independently of any drug treatment. Although irradiation eventually leads to cell death, the cells and the entire tissue remain intact until approximately 10 hours post irradiation (data not shown), which is long after migration has halted completely. We termed this approach *2P-Arrest* and used it to selectively arrest follower cells (Fig. 3.6c) and to examine the behavior of the unirradiated leader cells in response to this increased mechanical drag by fast time course images.

Initial qualitative observations indicated that tip cells produced elongated protrusions in response to 2P-Arrest of followers (Fig. 3.6d), which prompted us to conduct a quantitative analysis of tip cell behavior as well as actin dynamics in the leading region. We used automatically generated masks of primordia as well as some manual annotation to quantify tip speed and protrusion length based on the membrane marker *Lyn:EGFP* as well as F-actin abundance and polarized motion (optical flow) based on the F-actin probe *mNeonGreen:UtrCH*. However, although we could confirm the increased protrusion length of tip cells (Fig. 3.6e), we found no other indication of an active response of leaders to 2P-Arrest of followers (Fig. 3.6f-i).

Since increased protrusion length could be an entirely passive mechanical effect and appears to be the only detectable response of leaders to increased follower drag, we conclude that acute experimental perturbation of tissue mechanics does not lend strong support to a plithotaxis-based model of primordium migration.

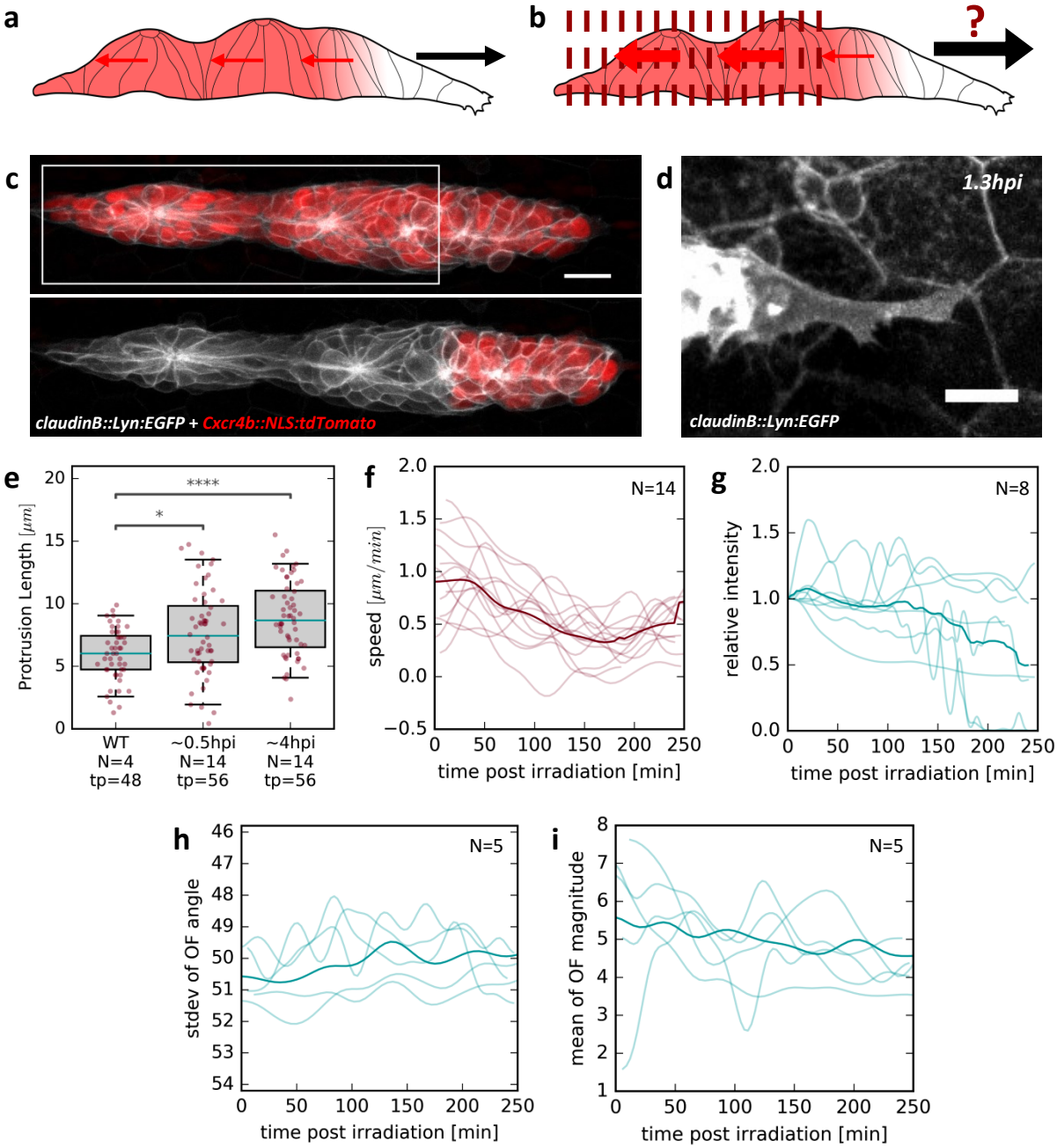

Figure 3.6: Leader cell response to 2P-Arrest of follower cells.

(a-b) Illustration of follower 2P-Arrest experiment. If follower cells (red) normally generate a counter-force to migration (red arrows) and thereby regulate leader cell polarity and motility (black arrow) through plithotaxis (a), then arresting follower cells by 2P-Arrest should increase this counter-force and consequently increase leader cells' migrational activity (b). (c) Maximum z-projection of pLLP before (top) and after (bottom) 2-photon irradiation. The white ROI in the top image indicates the area to be irradiated, which is directly reflected in the bleaching of the *NLS:tdTomato* channel. Importantly, whilst the nuclei are bleached in the bottom panel, the membrane labeling shows no indication of compromised tissue integrity. Scale bar: 15µm. (d) Maximum z-projection of a tip cell's unusually long protrusion following 2P-Arrest of the followers. *hpi* is hours post irradiation. Scale bar: 10µm. (e) Quantification of protrusion lengths in WT and after 2P-Arrest of follower cells. Because protrusion length is highly variable between time points, several time points are plotted for each sample, yielding a total of *tp* points. Measurements apart more than 10min were considered independent during statistical analysis (p[wt-0.5hpi]=9.99e-03, p[wt-4hpi]=3.37e-06).

Figure legend continued on the following page.

Figure 3.6 *(legend continued from previous page)*

(f-i) Different leader cell features measured over time. Individual time points have been smoothed to reduce noise, emphasizing overall behavior. Faint lines show individual samples, strong lines show running mean. None of these measures shows a consistent trend that might indicate an active leader cell response to follower arrest. (f) Tip cell speed measured using the *Lyn:EGFP* membrane marker. (g) Background-subtracted *mNeonGreen:UtrCH* intensity (a proxy of F-actin abundance) relative to the initial time point. The rapid loss of signal at time>150min in some samples is due to masking issues resulting from bleaching of the *NLS:tdTomato* channel. (h) The standard deviation of the F-actin optical flow angle. A decrease in this measure would indicate increased coaxial polarization in leader cells. (i) Mean of F-actin optical flow magnitude. An increase in this measure would indicate faster motion of actin structures such as lamellipodia.

3.3 Development of a Computational Framework for Image-Based Quantitative Single-Cell Analysis of Cellular Architecture in Living Tissues

Data-driven biology represents an important new avenue for the perturbation-free investigation of complex biological systems, as elaborated in section 1.3.3. However, whilst the omics technologies have enabled the large-scale collection and analysis of big data snapshots of the abundance and interactions of system components, equivalent approaches for capturing components' spatial distributions and temporal dynamics as well as their relation to biological function are still underdeveloped, especially at the multi-cellular level.

In principle, microscopy is readily capable of providing the context-rich 3D+time data required for this purpose, but more work is needed to advance the extraction, representation, visualization and analysis of the rich information encoded within images. Here, we set out to establish a framework for the computational analysis of cell shape, architecture and dynamics in living tissues, seeking ultimately to discover predictive relationships between various cell and tissue-scale properties that could inform new mechanistic studies.

To this end, we performed high-resolution live imaging of the pLLP, implemented a pipeline for single-cell segmentation based on a membrane marker, developed a method for casting the result into a computational representation amenable to data-scientific techniques such as machine learning, and finally applied such techniques for atlas mapping and quantitative analysis of the diversity of cellular architectures across the lateral line primordium.

3.3.1 Fast High-Quality 3D Imaging of the pLLP Using AiryScan Microscopy

In vivo live imaging is a balancing act between image quality, acquisition speed and phototoxicity. For automated single-cell segmentation to be possible in 3D, high signal-to-noise ratios and good axial resolution are required. This can be achieved with classical confocal microscopy and deconvolution (Fig. 3.7a) but long acquisition times lead to an axial distortion of the primordium due to its migratory motion and also massively reduce sample throughput or temporal resolution during time courses. Classical spinning disk confocal microscopy, by contrast, is sufficiently fast but does not produce images of high enough quality for single-cell segmentation (Fig. 3.7b).

We therefore tested and optimized the recently released Zeiss LSM880 AiryScan confocal microscope for high-resolution 3D imaging of the lateral line primordium using the AiryScan FAST mode [Huff, 2016]. We found that the line-scanning approach employed by the FAST mode coupled with AiryScan deconvolution enabled the acquisition of high-quality volumes (Fig. 3.7c) at an acceptable rate, although not quite as quickly as with a spinning disk.

All images shown and analyzed in the following sections have been acquired in this fashion, consistently using the membrane marker *claudinB::Lyn:EGFP* to enable single-cell segmentation and – when required – using two-color acquisition to simultaneously image an additional label of interest.

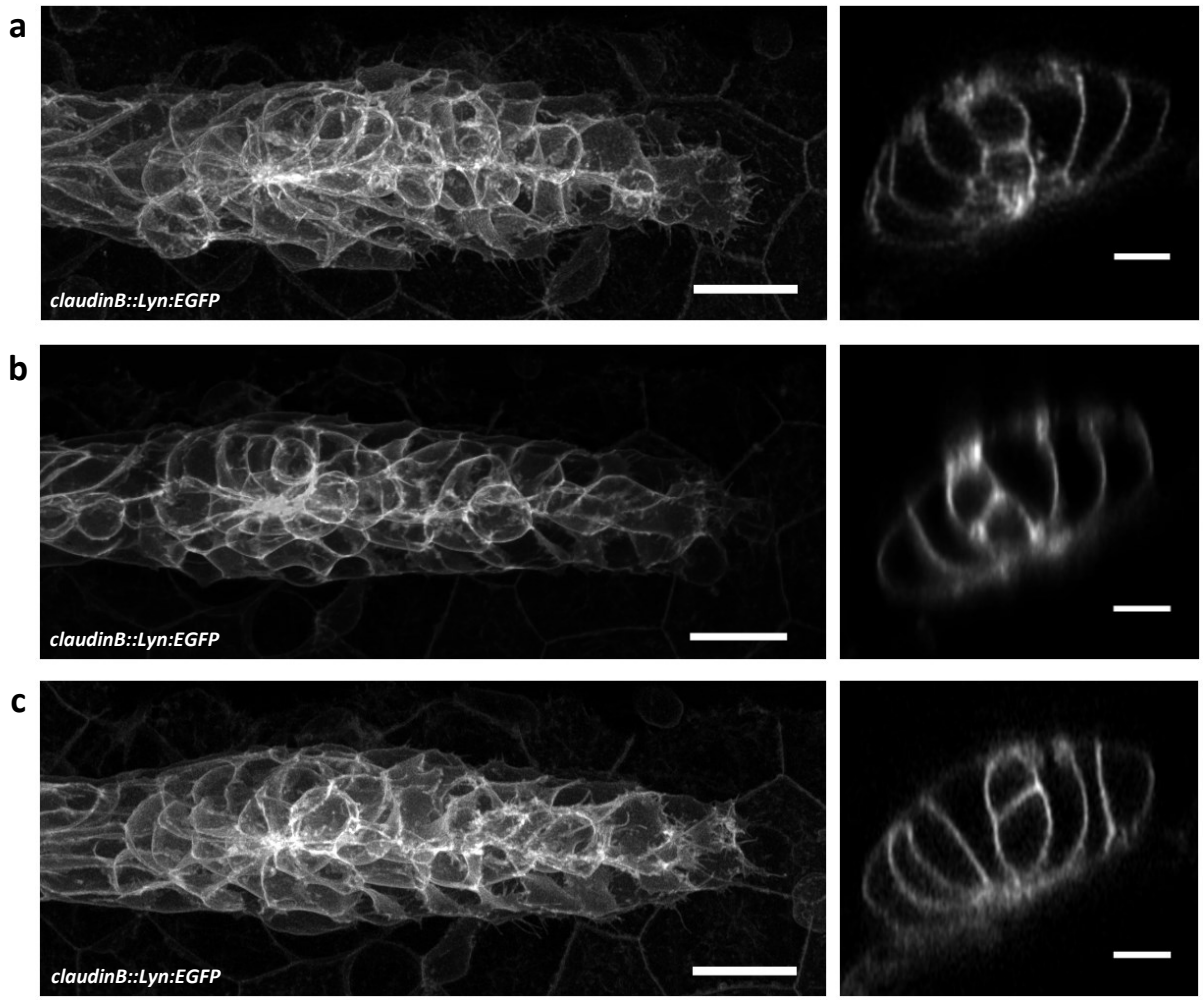

Figure 3.7: Comparison of scanning confocal, spinning disk confocal and AiryScan imaging of the pLLP.

(a-c) Maximum z-projections (left, scale bars: 15µm) and yz-reslices (right, scale bars: 5µm) of deconvolved 3D volumes of the same primordium acquired using different modes of confocal imaging with settings individually optimized for high image quality first and acquisition speed second. (a) Acquired on the LSM880 in conventional scanning confocal mode. Acquisition speed: 944ms/slice. (b) Acquired on the VoX spinning disk microscope. Acquisition speed: 200ms/slice. (c) Acquired on the LSM880 in AiryScan FAST mode. Acquisition speed: 322ms/slice.

3.3.2 Automated 3D Single-Cell Segmentation of the pLLP

We combined and adapted commonplace image analysis algorithms to implement an automated 3D single-cell segmentation pipeline based solely on the *claudinB::Lyn:EGFP* membrane marker in order to keep other channels free for imaging of other labels of interest. Briefly, we preprocessed images by smoothing, masked foreground signal (membranes) using automated threshold detection, inverted the mask and labeled connected components to obtain segmentation seeds from low-signal areas (i.e. cell bodies), filtered away very large and very small seeds (background), and performed watershed on the smoothed membrane image to expand seeds into full cell segmentations.

We found that this approach generally yields high-quality segmentations (Fig. 3.8). Importantly, erroneous cell fusion or over-segmentation are exceedingly rare, likely due to the high quality of the input images. Occasionally, a few cells are being missed and wrongly assigned to the background, which we expect not to affect our downstream analysis. To ensure consistent segmentation quality, we manually double-checked each segmented stack and excluded rare cases exhibiting substantial segmentation issues.

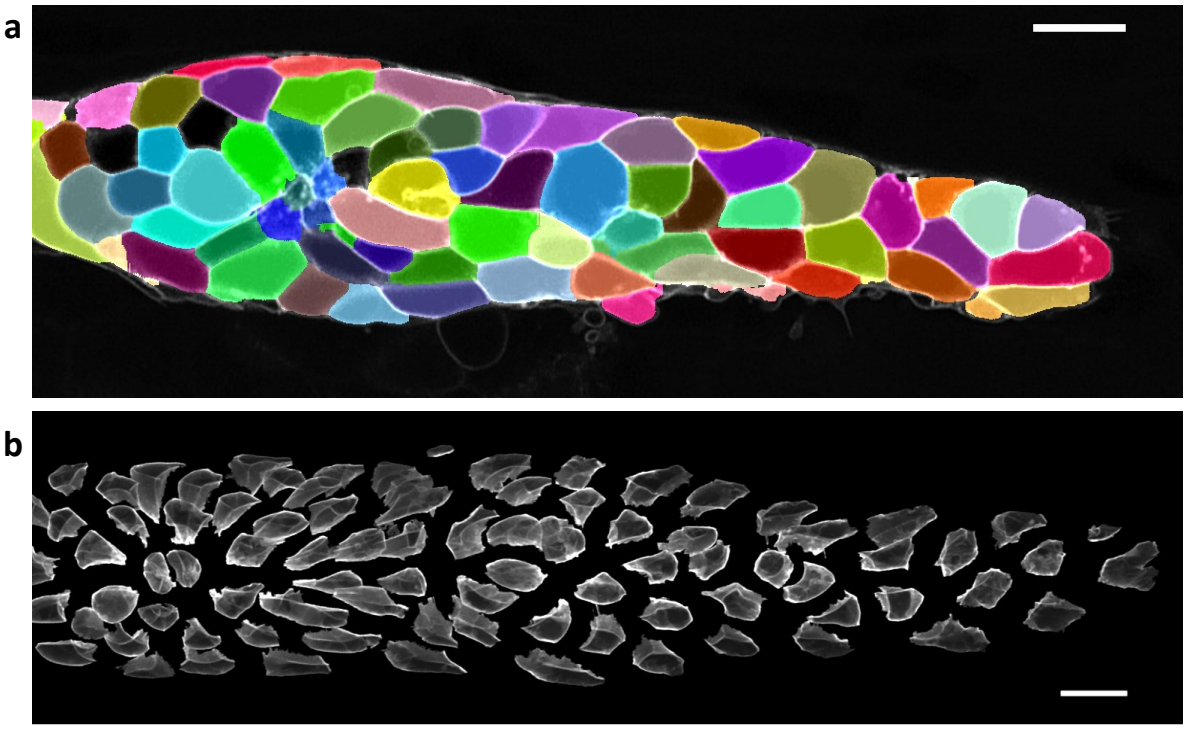

Figure 3.8: Automated 3D single-cell segmentation of the pLLP.

(a) A single xy-slice with segmented cells indicated by a semi-transparent color overlay. Scale bar: 10µm. (b) "Exploded view" of a segmented pLLP. Individual segmented cells have been shifted apart, revealing their individual shapes within the collective. Scale bar: 15µm.

The main drawback of our current solution is that fine cell protrusions, such as lamellipodia or thin apical extensions toward rosette microlumina, are often missed by the algorithm. This is unavoidable, however, as most of these fine structures are not even distinguishable by the human eye when the entire tissue is labeled – they only become apparent with mosaic labeling, which would not provide the throughput required for data-driven analysis. Thus, we decided to focus on overall cell shape and architecture, ignoring fine protrusive structures for the time being.

3.3.3 Intensity-Biased Stochastic Landmark Assignment (ISLA) and Cluster-Based Embedding (CBE) for Latent Feature Extraction from Arbitrary Fluorescence Distributions

Following segmentation, we measured a host of specifically selected or designed features (henceforth referred to as *engineered features*) for each cell individually, including size features such as volume or height, shape features such as aspect ratios, and fluorescence features such as the mean fluorescence of different imaging channels (see appendix A2 for a complete list).

However, whilst such specifically engineered features are often sufficient to address hypothesis-driven questions, they are problematic for exploratory approaches as they may miss important architectural properties and may be biased by the scientist's perception of what features might be relevant [Pincus & Theriot, 2007].

Here, we therefore additionally sought to develop an unbiased approach to embed fluorescence distributions into a feature space of *latent features*, a challenge that has been addressed to some extent for 2D images [Pincus & Theriot, 2007] but is not trivial to generalize to the third dimension. Importantly, all currently available solutions require cells to be spatially registered to account for rotational variance (but registration is an ill-posed problem in cases where very different cells are to be analyzed) and many are based on segmentation masks or outlines (which is not ideal for embedding fluorescence distributions where information is encoded in intensity variation, e.g. the actin cytoskeleton).

We took classical geometric morphometrics [Adams et al., 2004; Adams et al., 2013] as a starting point to solve the challenge of latent feature extraction in a robust, versatile and general fashion. In geometric morphometrics, distributions are usually represented as landmarks that form 3D point clouds. Feature space embedding is then commonly performed by spatial alignment of point clouds followed by re-representation of landmark coordinates as feature vectors, often by means of dimensionality reduction techniques such as Principal Component Analysis (PCA) [Adams et al., 2004].

In order to apply this approach to cells, we needed to solve two problems. First, point clouds must be extracted from fluorescence intensity images in such a way that they represent the fluorescence intensity distribution. Second, as points are not matched across cells in the way they would be e.g.

across facial features (the 'nose landmark' of face 1 would be matched with the 'nose landmark' of face 2), a different way of re-representing point clouds is needed, ideally one that allows rotational invariance to be achieved without registration.

We solved the first problem using a technique we termed *Intensity-biased Stochastic Landmark Assignment* (ISLA). To generate a sparse point cloud representation of a 3D volume, ISLA treats voxel intensities as a multinomial probability distribution (normalized such that the volume's total intensity equals 1) from which a defined number of points are selected by random sampling. In the resulting point cloud, the local density of points represents local fluorescence intensity (Fig. 3.9a). Increasing the total number of landmarks sampled in this way increases the accuracy of the point cloud representation (asymptotically toward 100%) at the cost of point cloud sparsity and thus computational efficiency; this trade-off can be tuned according to the user's requirements.

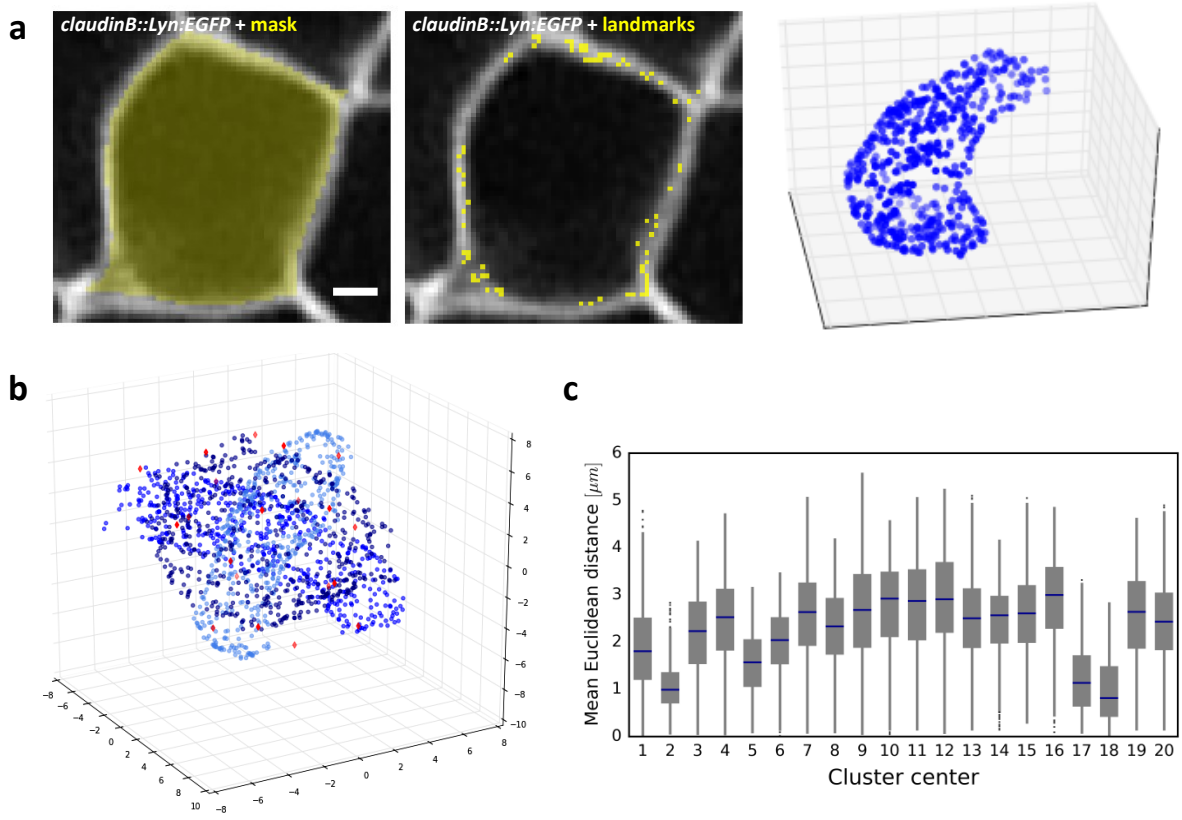

Figure 3.9: Illustrations of key aspects of ISLA and CBE.

(a) Illustration of Intensity-biased Stochastic Landmark Assignment (ISLA). Pixels intensities within a segmentation mask (left, yellow area) are considered probabilities for multinomial sampling of landmarks (right, yellow dots), which yields a 3D point cloud (right) that encodes much of the information about the original fluorescence intensity distribution. Scale bar in images: 1µm. Note that for illustration purposes the 2D points (middle) are more densely sampled than the 3D cloud (right). (b-c) Illustration of Cluster-Based Embedding (CBE). (b) Point clouds from different cells (different shades of blue) are overlaid in order to determine common reference points (red) by clustering. Axis scales are in µm. (c) Next, the mean Euclidean distance from each reference point to a set of its nearest neighbors in each cell's point cloud is computed, yielding a feature space such as the one exemplified here. In this boxplot, each observation is a cell, the x-axis are the common reference cluster centers and the y-axis is the mean Euclidean distance metric.

To solve the second problem, we reminded ourselves that classical point cloud re-representation usually either explicitly or implicitly represents landmarks not by their absolute coordinates but by coordinates relative to a common reference. When landmarks are matched, this common reference is simply the mean position of each landmark across samples. Here, landmarks are unmatched but we reasoned that even arbitrary points in space could in principle serve as common reference points, given that they are sensibly distributed.

Following this rationale, we developed *Cluster-Based Embedding* (CBE). To perform CBE, we first overlay all point clouds from all samples (centered on each point cloud's centroid) and subsequently use k-means clustering [MacQueen, 1967] to determine local landmark enrichments (i.e. clusters) and to provide each with a reference point (i.e. the cluster centers) (Fig. 3.9b). The number of clusters, which is user-defined, determines the trade-off between the fidelity and the computational efficiency of the embedding. Next, we re-represent each individual cell's point cloud relative to the common reference points by computing the mean Euclidean distance between each reference point and its nearest neighbors (Fig. 3.9c) (for additional details see Materials & Methods, section 2.3.6). Finally, we z-score the resulting features and transform them using PCA, which has been found repeatedly to yield biologically meaningful features [Pincus & Theriot, 2007]. The result is a latent feature space with as many features as there were common reference points, although higher-order Principal Components (PCs) may be discarded as noise.

Whilst CBE addresses the challenge of embedding arbitrary point cloud distributions, it does not natively handle rotational variance. We therefore made use of the point cloud shape representation generated by ISLA to also resolve this issue. For each cell's point cloud, we computed the set of the lengths of all Pairwise Distances (PDs) between that cloud's points. Although this information alone may not always enable a full reconstruction of the original point cloud, it is expected to encode most of the relevant information and – importantly – it is rotationally invariant. To cope with the high dimensionality of this representation, we downsampled the set of lengths by including only a certain subset for each point (see Materials & Methods, section 2.3.6). Because CBE can embed arbitrary point cloud distributions of arbitrary dimensionality, the pairwise distance representation can be embedded just as well as the originally coordinate representation, yielding a rotationally invariant latent feature space.

We ultimately combined the above to compute two embedded spaces, termed the *Tissue Frame of Reference* (TFOR) and the *Cell Frame of Reference* (CFOR) space. To compute the TFOR space, primordia were first registered in their entirety and CBE was performed without a PD transform. Thus, the TFOR space is expected to encode not only shape and size but also meaningful rotational information (i.e. the rotation of cells relative to the entire primordium). By contrast, the CFOR space was computed on point clouds that were first size-normalized and then PD-transformed, resulting in

a feature space that encodes exclusively information on the shape of the embedded fluorescence distribution.

Further details, including a complete overview of data flow through the ISLA-CBE pipeline (Fig. 2.1), can be found in Materials & Methods, section 2.3.6.

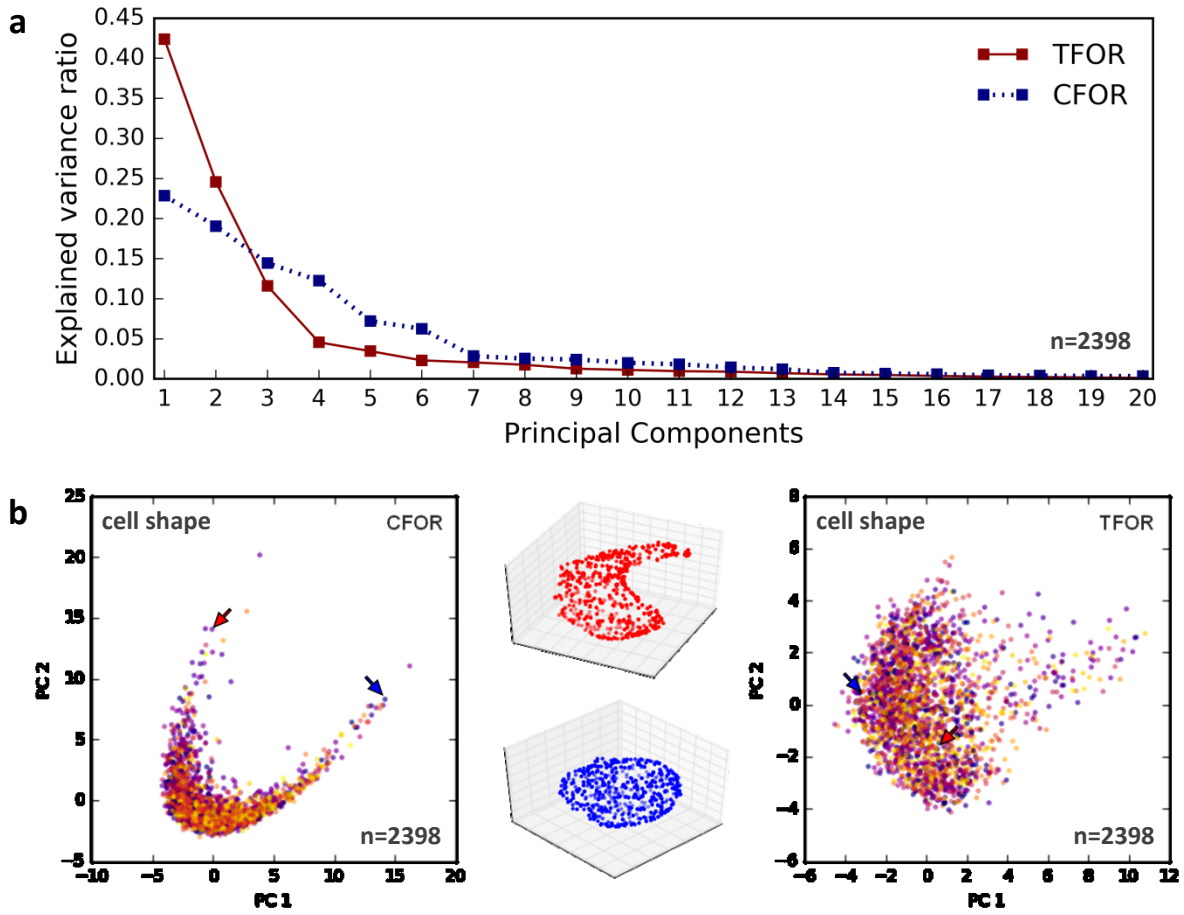

Figure 3.10: ISLA and CBE produce compact and expressive latent feature spaces.

(a) Ratios of total variance explained by individual PCs in an embedding constructed from pLLP cell shapes. Both in TFOR and CFOR, most of the population variance is explained by only a few principal components. (b) Visualizations of the first two principal components of CFOR (left) and TFOR (right) latent feature spaces created by embedding cell shape. Each dot is a cell and the different colors denote the primordia from which each cell derives, showing that the distribution of cell shapes within each primordium is similar. Different cell shapes (middle) are very clearly separated in CFOR (arrows). This is less obvious in TFOR, where cell orientation and size also play a major role. All axes are unitless.

3.3.4 ISLA and CBE Generate Meaningful Latent Feature Spaces

When applied to our dataset of pLLP cells, we found that the ISLA-CBE approach produces compact and expressive latent feature spaces. For example, an embedding of cell shape based on the *claudinB::Lyn:EGFP* membrane marker showed that most of the shape heterogeneity across the primordium can be explained by just a small number of principal components (Fig. 3.10a). Furthermore, the principal components clearly separate cells based on biologically relevant shape differences. In the CFOR cell shape embedding, for instance, very round cells (dividing cells or cells at the center of rosettes) fall onto one tip of the distribution and very curved cells (cells in the periphery of rosettes) fall onto the other (Fig. 3.10b).

We employed machine learning to quantitatively assess how much information is shared between our latent feature spaces and the explicitly engineered features. To do so, we trained Support Vector Regressors (SVRs) using CBE-embedded cell shape features to predict different engineered descriptors of cell shape. Consistent with expectations, we found that the TFOR shape space encodes much of the information contained in simple shape descriptors (Fig. 3.11a) whereas the CFOR shape space only encodes information that is unrelated to the absolute size or orientation of cells (Fig. 3.11b). We also compared CBE to an alternative embedding method based on the extraction of moments from the ISLA point cloud (for details see Materials & Methods, section 2.3.6). However, although this approach performed similarly well in CFOR (Fig. 3.11d), it could not compete with CBE in TFOR (Fig. 3.11c), prompting us to proceed with CBE as our primary strategy for latent feature extraction.

Finally, we sought to test whether latent features faithfully detect single-cell phenotypes in response to perturbation. We initially attempted to do so by treating embryos with EDTA, inducing cell rounding in the lateral line. However, due to cell-internal accumulation of *lyn:EGFP* following EDTA treatment, segmentation of these samples was not possible (data not shown). We therefore turned to the mosaic expression of *LexOP::Lyn:tagRFP-T:ARHGEF25b* (chemoARHGEF), which causes cells to round up (see section 3.14). As expected, such cell rounding is clearly detectable in a CBE-embedding of cell shape, where cells expressing higher levels of chemoARHGEF also show higher values of PC 1 in CFOR (Fig. 3.12).

Taken together, these results demonstrate that our morphometric pipeline based on ISLA and CBE is capable of embedding 3D fluorescence intensity distributions into meaningful latent feature spaces.

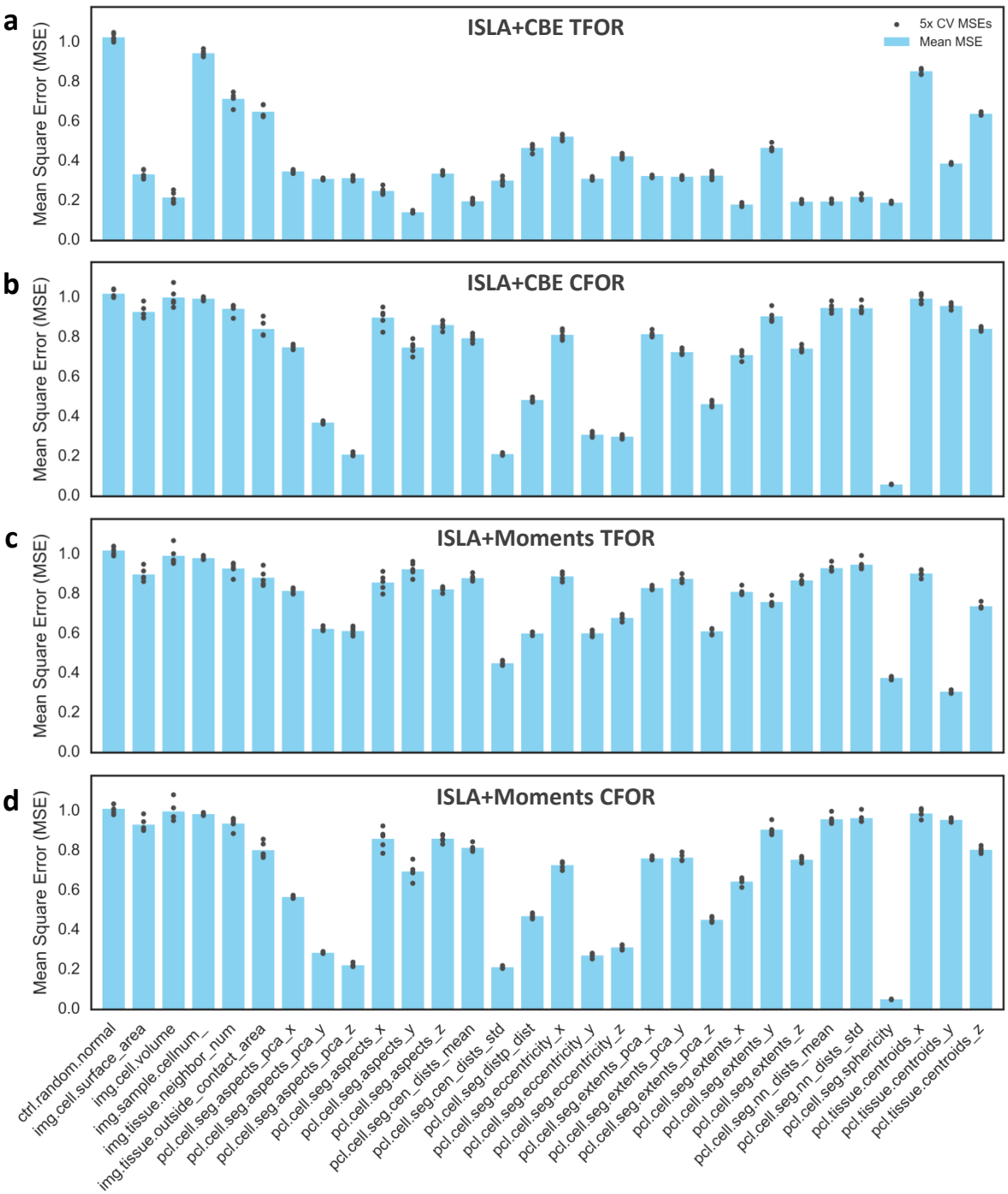

Figure 3.11: Recovering engineered shape descriptors from the latent feature space.

Mean Square Errors (MSE) for SVR regressions predicting z-scored engineered features from different latent feature representations of cell shape. Lower MSEs indicate that the SVR was able to predict the feature in question to a greater extent. Note that the first column, *ctrl.random.normal*, is randomly sampled from a normal distribution and thus represents a negative control where a meaningful regression is impossible. The ISLA+CBE TFOR (a) produces an embedding that encodes much of the information included in engineered features, whereas the corresponding CFOR (b) does not encode information that is related to cell size (e.g. *img.cell.volume*) or where orientation is relevant (e.g. the cell's length along the front-rear axis, *pcl.cell.seg.extents_x*) but is even better at encoding pure shape features (e.g. *pcl.cell.seg.sphericity*). The alternative embedding strategy based on point cloud moments behaves similarly for CFOR (d) but does not perform nearly as well as CBE for TFOR (c). A complete list of explanations of the different engineered features can be found in appendix A2. Y-axes are unitless due to z-scoring.

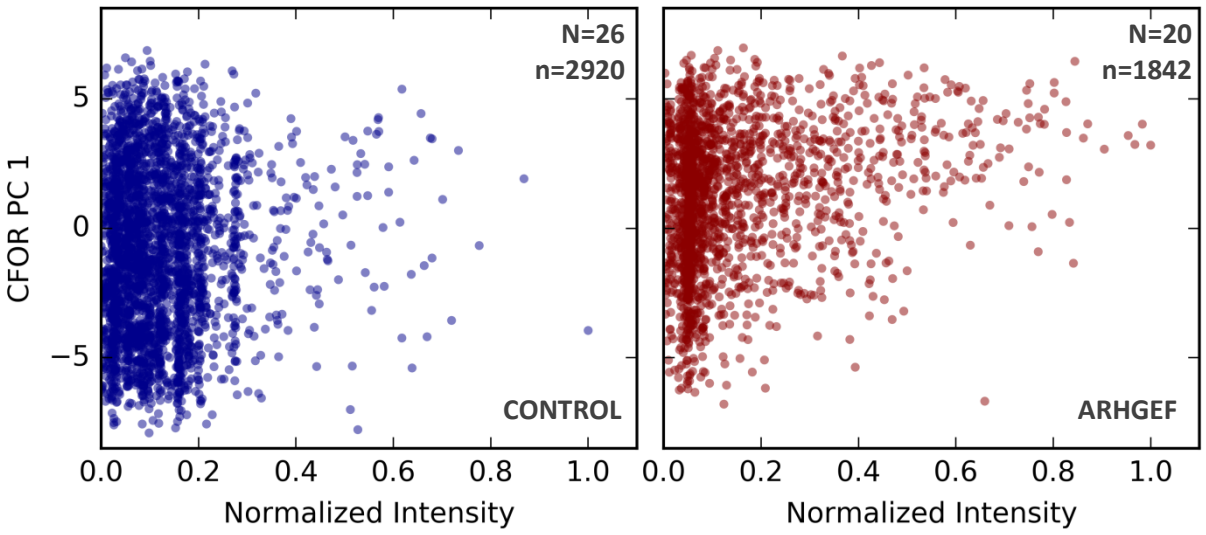

Figure 3.12: Latent features can detect phenotypes at the single-cell level.

Mosaic expression of chemoARHGEF leads to cell rounding, which is captured by a CFOR embedding of cell shape. Cells with a high intensity of *Lyn:tagRFP-T:ARHGEF25b* are enriched among higher values of CFOR PC 1 (right), unlike cells with a high intensity of the inert control construct *NLS:mIRFP* (left). Intensity was normalized as described in section 2.3.3. All axes are unitless.

3.3.5 Preliminary Analysis of the pLLP Cellular Shape Space

Cell shape is a key architectural and functional property of pLLP cells. It influences and is influenced by both migration and rosette formation and it is controlled both internally by the cytoskeleton and externally by the forces and constraints experienced by each cell in the tissue context. Even cursory visual inspection reveals that cell shapes differ non-trivially between leader cells, cells at the center of rosettes and cells in the periphery of rosettes. We therefore employed our novel method for latent feature extraction to embed the shapes of 16'974 cells from 190 wild-type primordia based on the cell membrane marker *claudinB::Lyn:EGFP*, which allows us to perform a comprehensive analysis of the pLLP's cellular shape space, the preliminary results of which are described here.

First, we sought to identify the primary factors of shape heterogeneity in the pLLP. As mentioned in section 3.3.4, we found that just a few PCs are sufficient to explain shape heterogeneity, with >95% of variance in the shape space being captured by 11 PCs in TFOR or 9 PCs in CFOR. We focused our analysis on the first 6 PCs, which each explain more than 5% of variation on their own (Fig. 3.10a). To determine the nature of these principal components, we correlated them with our engineered features (Fig. 3.13ab) and visualized resulting associations as bigraphs (Fig. 3.13cd).

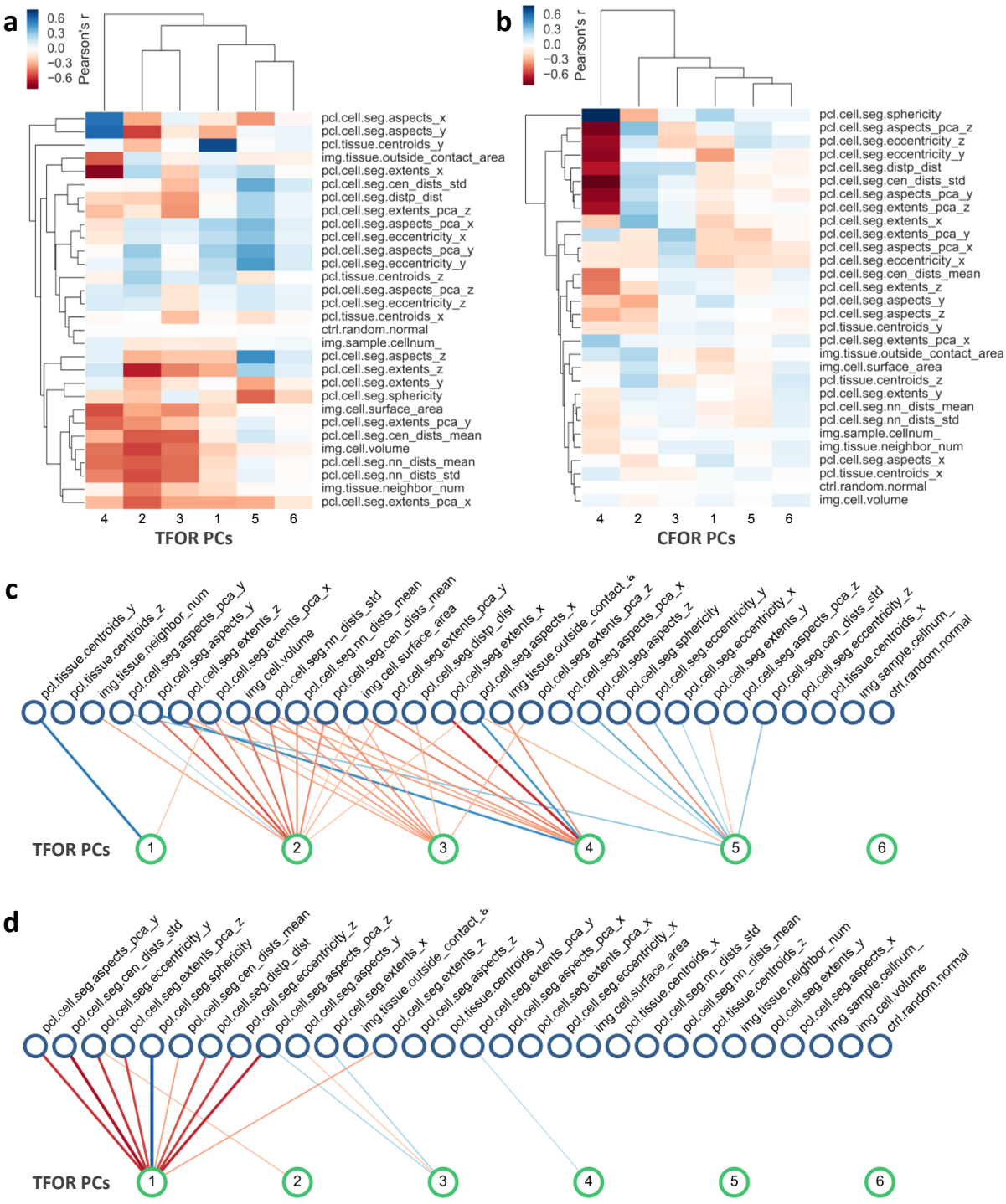

Figure 3.13: Identification of latent feature meaning by correlation with engineered features.

(a-b) Clustermaps showing Pearson correlation coefficients between shape space components and various engineered features for both TFOR (a) and CFOR (b) . (c-d) Bigraphs showing the association of engineered features with shape space components based on the strength of correlation. Edge widths reflect the strength and edge colors the sign (red: negative, blue: positive) of the Pearson correlation coefficient between two samples (edges with *r<abs(0.3)* are omitted). TFOR PCs (c) are associated with the centroid position along the dorso-ventral axis (*pcl.tissue.centroids_y*) (PC 1), with various measures of cell size and height (PC 2), with measures of protrusion length such as the distance from the centroid to the most distant landmark (*pcl.cell.seg.distp_dist*) (PC 3), with the length of a cell along the pLLP's front-rear axis (*pcl.cell. seg.extents_x*) (PC4), and with measures of sphericity (PC5). Among the CFOR PCs (d), only PC 1 has a very clear association, namely to measures of sphericity. See appendix A2 for an index of engineered features.

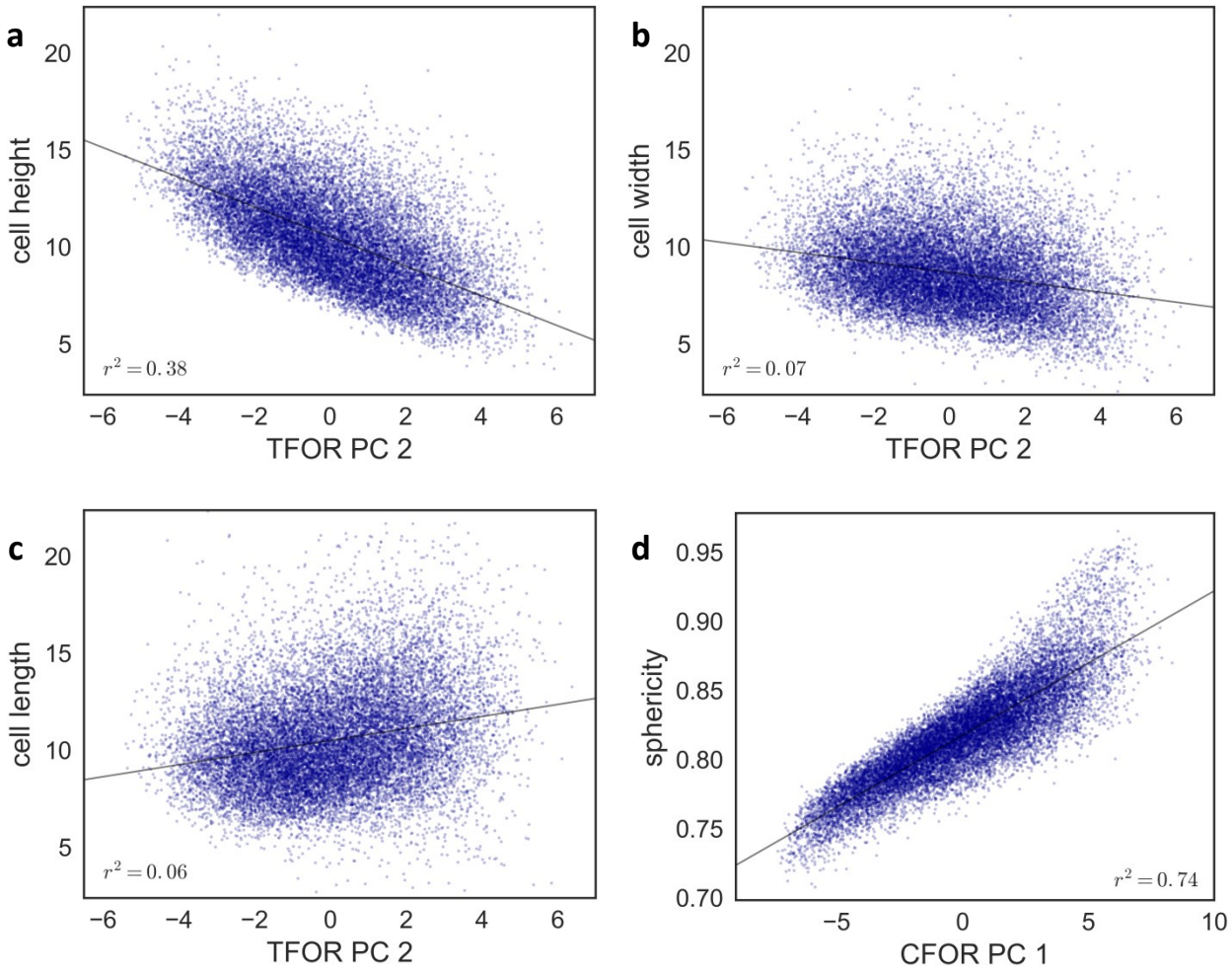

Figure 3.14: Key examples of relationships between latent and engineered features.

(a-c) TFOR PC 2 is associated with cell size, including cell volume (r^2=*0.29*) (not shown). The main contribution to size variation across the primordium appears to stem from cell height along the apico-basal axis (r^2=*0.38*) (a) rather than from cell width (r^2=*0.07*) (b) or from cell length along the front-rear axis (r^2=*0.06*) (c). (d) When size and orientation information are removed, the most important remaining aspect of shape variation is cell sphericity, which is very closely associated with CFOR PC 1 (r^2=*0.74*). Cell height, width and length are given in µm, sphericity and principal components are unitless.

In doing so, we found that TFOR PC 1 is clearly associated with lateral orientation of primordium cells, reflecting the mirror symmetry of the pLLP along its midline. TFOR PC 2 is associated with cell size and volume, which interestingly appears to be mainly a function of cell height along the apico-basal axis (Fig. 3.14a-c). TFOR PC 5 is associated with cell roundness and TFOR PC 4 could be associated with front-rear polarity of the cell, given its relationship with the cell's aspect ratio in the plane of the tissue. In CFOR, associations with size and orientation are missing, as expected. Instead, CFOR reveals a striking relationship of PC 1 and cell sphericity (Fig. 3.14d), indicating that sphericity is a central feature of cellular architecture in the pLLP. Interestingly, this would not be obvious based on the TFOR space alone, which shows that the inclusion of size and orientation during latent feature extraction may obscure important aspects of the shape space and highlights the success of the pairwise distance approach at removing these confounders. The remaining latent features could not

be clearly identified by this correlative approach. They may encode shape information that is not included in the engineered features and would have been missed without latent feature extraction.

Next, we mapped the shape space components back onto the spatial location of cells in the primordium (Fig. 3.15). Consistent with the observations above, TFOR PC 1 is patterned along the y-axis of the primordium (Fig. 3.15a), implying that it encodes cell orientation. Interestingly, TFOR PC 2 (cell height) shows a center-to-periphery pattern, indicating that central follower cells specifically have an extended apico-basal axis (Fig. 3.15b,d), although this does not take into account that the apico-basal axis in peripheral follower cells is curved. TFOR PC 3 appears to be increased in followers (Fig. 3.15c), indicating that it may encode biologically relevant information, the nature of which however remains unknown because it is not clear as of now what aspect of cell shape leads to high values in this feature. The distribution of CFOR features is less clear. CFOR PC 1 (sphericity) appears to be enriched along the center of the primordium but also in leader cells and at the very back (presumably chain cells), which in the first instance runs counter to expectations and demands additional validation (Fig. 3.15e,h). CFOR PC 2 does not seem to have a clear spatial pattern and may be uniquely enriched in certain rare cells (Fig. 3.15f), whereas CFOR PC 3 shows some indications of a center-to-periphery pattern (Fig. 3.15g). Further investigation will be required to determine the exact nature and the biological significance of these distributions.

We also asked whether pLLP cells naturally cluster into distinct groups of similar architectural "cell types" based on cell shape. However, we found no evidence for distinct clusters when visualizing the distribution of cells across principal components (for an example, see Fig. 3.10b) or when generating two-dimensional representations by t-Stochastic Neighbor Embedding (tSNE) [Van der Maaten & Hinton, 2008] (not shown). Instead, the cells of the pLLP appear to occupy a continuous domain of the shape space. Therefore, we decided not to proceed with any unsupervised clustering analysis for the time being and instead performed supervised archetype classification (see section 3.3.7).

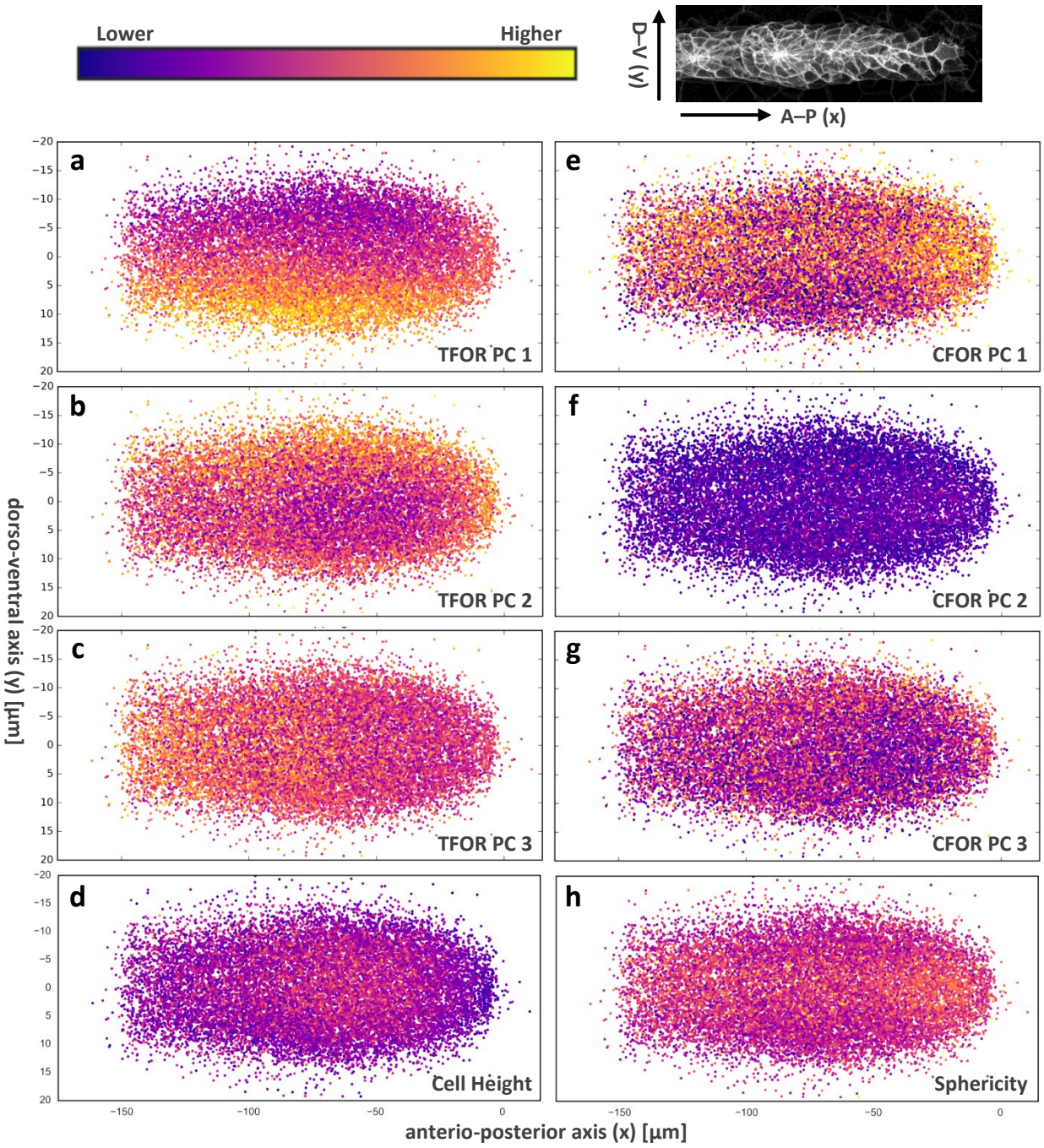

Figure 3.15: Distribution of shape features across the cells of the pLLP.

Top views composed of centroid xy-positions (each dot) of 16'974 single cells from 190 wild-type primordia, with colors indicating higher or lower values of the respective latent or engineered features, extracted from cell shapes based on *claudinB::Lyn:EGFP*. The small maximum z-projected image at the top right clarifies the viewing perspective. (a-c) Distribution of the three most expressive principal components in TFOR, exhibiting a lateral, an inside-outside and a front-back pattern, respectively. (d) Cell height, which closely corresponds to TFOR PC 2 and thus exhibits a similar inside-outside pattern (note that PC 2 is inverted; the sign of principal components is arbitrary). (e-g) The three most expressive PCs in CFOR. PC 1 is enriched in leaders and along the center of the primordium. PCs 2 and 3 do not show a clear spatial distribution. (h) Cell sphericity shows a similar pattern to CFOR PC 1, confirming that cell roundness follows this unexpected pattern, the source of which remains to be determined.

3.3.6 Preliminary Analysis of the Effects of Chemokine Signaling Mutants on the pLLP Shape Space

Having established a first overview of the cellular shape space of wild-type primordia, we next investigated mutant conditions, aiming to relate cell shape to molecular functions and potentially to tissue-scale pLLP behavior. To do so, we acquired stacks of *claudinB::Lyn:EGFP* primordia in three different mutant backgrounds, *Cxcr4b-/-* (guidance receptor mutant), *Cxcl12a-/-* (chemokine mutant) and *Cxcr7-/-* (scavenger receptor mutant). All of these mutants share the same macroscopic phenotype, namely that the primordium doesn't migrate or migrates very little. However, in the case of *Cxcr4b-/-* and *Cxcl12a-/-*, this is due to an absence of chemokine signaling in the primordium, whereas in the case of *Cxcr7-/-* it is due to uniformly high instead of graded chemokine signaling across the primordium.

Unfortunately, acquisition of high-quality stacks proved more challenging with mutant embryos because the primordia remain within the less accessible head region of the embryo. Consequently, segmentation of these samples was far more error-prone and in many cases a substantial number of cells were missed, which may introduce bias into the analysis. The following results are therefore to be regarded as preliminary.

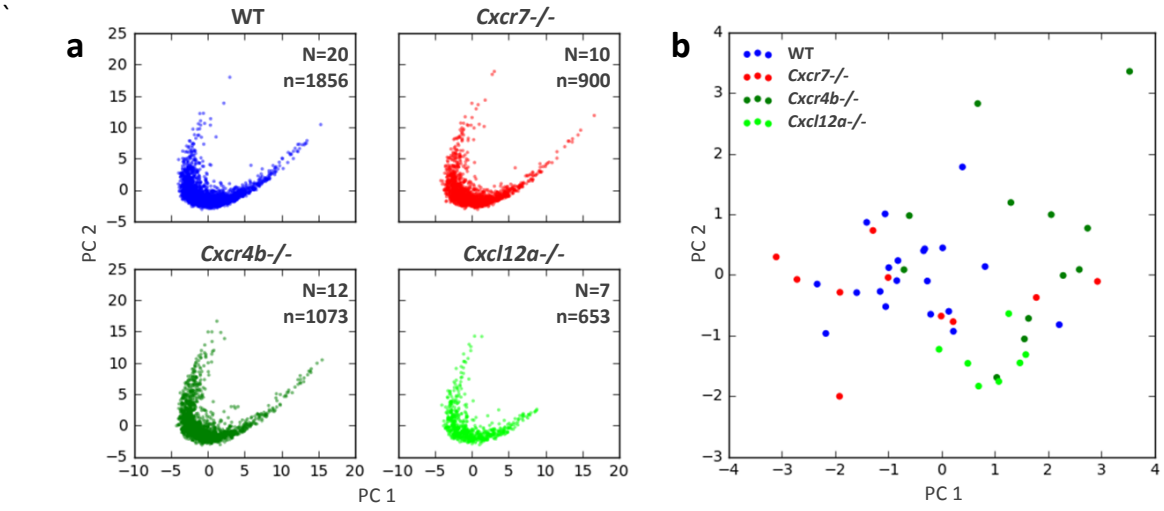

Figure 3.16: Comparison of wild-type and mutant shape spaces.

(a) An example of the lack of striking differences between wild-type and mutants; the first two principal components of the CFOR shape space do not show a clear loss or gain of subpopulations. (b) Results of a comparison of shape spaces on a per-primordium basis. Individual points in this plot are single primordia, not single cells. This comparison was accomplished by embedding each primordium's distribution of cells in the shape space using CBE followed by another PCA. Primordia whose cells are similarly distributed in the shape space therefore cluster together in this tissue-scale embedding. Whilst *Cxcr4b-/-* and *Cxcl12a-/-* primordia inhabit relatively distinct domains and are distinguishable from wild-type primordia along PC 1, *Cxcr7-/-* primordia show no clear pattern.

Surprisingly, we found no striking differences between these mutants and the wild-type among the most relevant principal components of the pLLP shape space (Fig. 3.16a), indicating that none of the different cellular architectures present in wild-type primordia is completely lost and no new and unique shapes are gained in the mutants.

Detecting more fine-grained differences such as changes in the relative frequency of different cell populations in the shape space is more challenging, which is why we decided to first test in an unbiased way whether the shape space allows wild-types and mutants to be distinguished at all. Considering the cells of each primordium as a point cloud in shape space, we used CBE to create a tissue-scale embedding that represents each primordium's entire cell population. Interestingly, this representation allows *Cxcr4b-/-* and *Cxcl12a-/-* mutants (but not *Cxcr7-/-* mutants) to be distinguished from wild-type primordia (Fig. 3.16b), showing that these mutants do in fact alter tissue architecture.

What exactly those alterations are and how they relate to the loss of chemokine signaling or the inability of the primordium to move directionally remains to be determined, although doing so will require higher-quality data or more robust segmentation.

3.3.7 Adding Biological Context to the pLLP's Cellular Shape Space through Morphological Archetype Classification and Visualization

Since interpreting high-dimensional data is a challenging task, any representation that provides additional biological context stands to be very useful. In order to annotate our dataset with such contextual information, we manually classified a small subset of cells into simple and biologically meaningful categories based on their location and architectural context within the pLLP. Using supervised machine learning, we then extended these labels to the entire dataset and used them to represent the single-cell dataset in a more interpretable fashion.

More specifically, we manually classified cells from 26 primordia into four groups: leader cells, inner rosette cells, outer rosette cells, and between-rosette cells (Fig. 3.17a). We only labeled archetypical cells, i.e. cells that clearly and unambiguously belong to a particular group. Next, we used this manual gold standard annotation to train a Support Vector Classifier (SVC) to classify cells across our entire dataset into these four categories based on the TFOR shape space features. We found that our classifier was reliably able to distinguish leader cells, inner rosette cells and outer rosette cells but frequently mistook between-rosette cells for outer rosette cells, which indicates that they are morphologically very similar to normal outer rosette cells despite their more internal location (Fig. 3.17b). Overall, this classification reproduced the spatial distribution of the selected cell archetypes across the primordium (Fig. 3.17c).

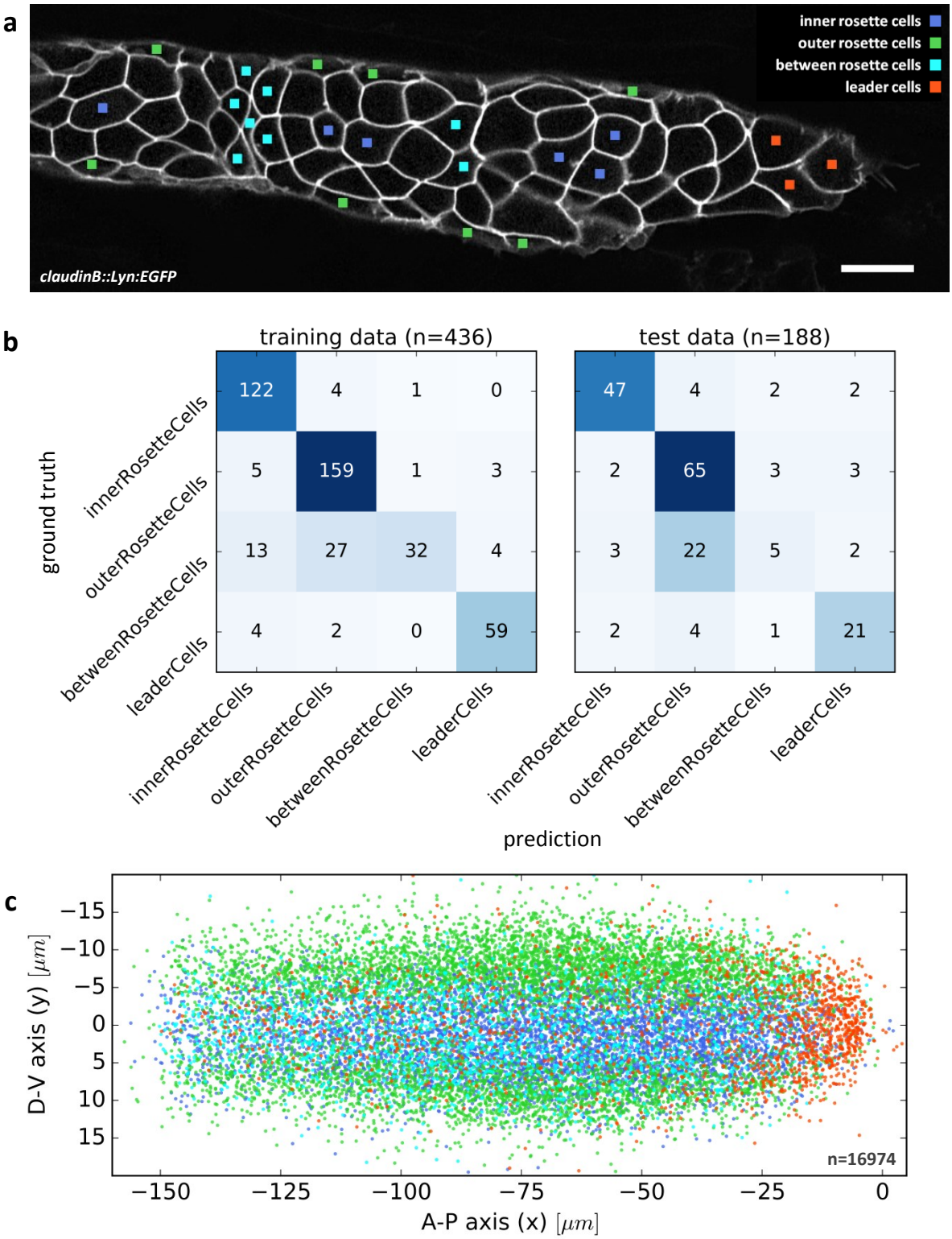

Figure 3.17: Archetype analysis of the pLLP cellular shape space.

(a) A single xy-slice from a 3D volume showing cells manually annotated as belonging to one of the four indicated archetypes. Scale bar: 10µm. (b) Confusion matrices showing the correspondence of ground truth and prediction produced by the archetype classifier on training data (left) and on previously unseen test data (right). Overall accuracy is very high in both cases but between-rosette cells are very commonly confused for outer rosette cells, especially in test data. (c) Mapping of archetype classification labels onto cell centroid positions in a registered primordium space (as in figure 3.15), showing that overall the SVC predictor recovers the spatial distribution of morphological archetypes very well, despite not being given any explicit positional information through its input features.

Having classified all cells into these particular groups makes it possible to analyze other data in an archetype-specific and thus more context-aware manner, for example by comparing the distribution of various other measures between archetypes. However, we wanted to go one step further and also resolve intermediate states between archetypes. To do so, we performed PCA on the classification probabilities inferred by the SVC archetype classifier, creating an embedding of archetype-likeness (the *archetype space*), where different archetypes fall onto the different corners of a high-dimensional polygon and intermediate states fall onto the edges or planes between them. In this case, the archetype space can be fully represented in three dimensions (Fig. 3.18a), but given the similarity of outer rosette cells and between-rosette cells a two-dimensional representation is also sufficient to represent the cell distribution in an interpretable way (Fig. 3.18b, top-left panel).

The distribution of pLLP cells within the archetype space can be interpreted with respect to the pLLP's architectural organization. For instance, the very low density of intermediate states between leader and inner rosette cells indicates that there is no direct transition pathway between the two. Cells must first adopt an outer rosette cell-like (or between-rosette cell-like) morphology in order to eventually transition into inner rosette cells. On the other hand, the triangle created by outer, inner and between-rosette cells features many intermediate cells, reflecting the continuous variation of cell shape from the center of a rosette to its outside.

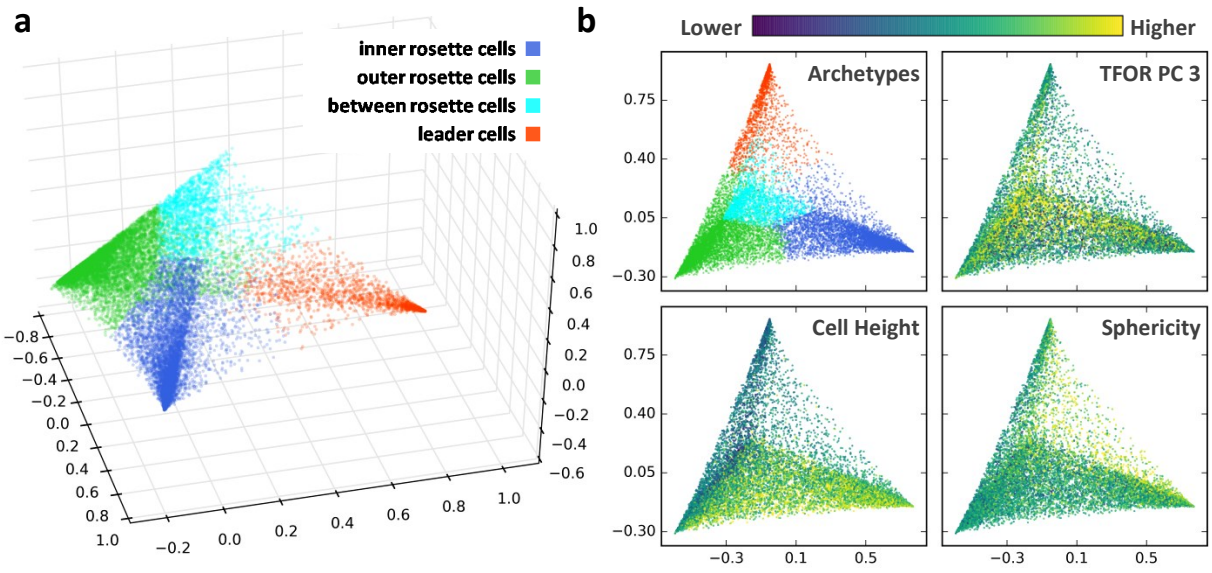

Figure 3.18: Archetype-based visualization of single-cell data.

(a) 3D visualization of the pLLP's archetype space based on cell shapes. Each dot represents a cell (n=16'974). The closer a cell to one of the corners the more clearly its shape resembles that of the corresponding morphological archetype. (b) 2D visualization of the same archetype space with different color overlays showing the predicted archetype label (top left), shape space TFOR PC 3 (top right), cell height along the apico-basal axis (bottom left) and cell sphericity (bottom right). All axes are unitless.

Importantly, the archetype space can serve as a canvas to map other types of information into a more interpretable representation, as shown in figure 3.18b. Both latent and engineered features can be mapped, showing for instance the increase in cell height along the apico-basal axis across follower cells (Fig. 3.18b, bottom-left panel) and high levels of sphericity both in inner rosette cells but also in particular among the sporadic cells that fall between inner rosette cells and leader cells (Fig. 3.18b, bottom-right panel), which could potentially be cells rounding up due to mitosis. It also shows an enrichment for the as of yet unidentified TFOR PC 3 latent feature in between-rosette cells (Fig. 3.18b, top-right panel), indicating that this principal component may in fact be useful in identifying such cells or a particular subset among them.

Overall, these data show that archetype classification and archetype space visualization constitute a simple and intuitive tool for adding biological context to high-dimensional datasets, making them a suitable starting point for a more in-depth investigation of the pLLP's tissue architecture and its cellular shape space.

3.3.8 Machine Learning Enables an Atlas Overlay of Multiple Experiments Based on Cell Shape as a Common Reference

The concept of data integration across multiple experiments by constructing an "atlas" based on a common reference measurement was introduced in section 1.3.3. Here, we relied on cell shape as a reference both because any segmentation-based single-cell analysis inherently provides the required shape information (encoded in the segmentation mask) and because we already established that pLLP cells exhibit diverse but non-random shapes closely related to their state and function.

Two approaches are conceivable for the construction of an atlas based on cell shape. One would be to register segmented cells in 3D space, thus overlaying whatever other channels were acquired alongside the membrane marker. However, because cell shapes in the primordium are not stereotypical, registration across all cells would yield a nonsensical result. This issue can be resolved by selectively registering only cells that have a very similar shape – in other words, cells within close proximity of each other in the shape space. The other option presents itself because the ISLA-CBE pipeline is capable of extracting latent feature spaces from arbitrary fluorescence distributions. Once latent features have been extracted both from cell shape and from a channel to be mapped, machine learning can be used to learn the relationship between the two and thus to predict the latent feature values of the latter based on those of the former.

Here, we imaged a range of markers as secondary channels (red or far-red) alongside the green membrane marker *claudinB::Lyn:EGFP* and used both selective registration and machine learning on embedded spaces to generate atlas mappings.

More specifically, we extracted latent shape features from the segmentation hulls of all available cells and performed a nearest-neighbor search in shape space to determine the 10 morphologically

most similar cells for each secondary channel. We then used the Iterative Closest Point (ICP) algorithm [Chen & Medioni, 1992] to register the ISLA-generated point clouds of these cells, keeping only the 5 best matches based on ICP loss values. Finally, we reconstructed image volumes from the resulting point cloud overlays by Gaussian Kernel Density Estimation (KDE) [Scott, 1992] and mapped the individual cellular predictions back into the corresponding primordia. In this way, we were able to generate predictions of the fluorescence distribution of any of the secondary channels for any of the primordia in our dataset, including those where only the membrane marker alone had been imaged at the microscope (Fig. 3.19).

To create an atlas of embedded spaces, we extracted latent features both from the segmentation hull and from the fluorescence distribution of the secondary channel. Next, we evaluated different machine learning models for their capability to predict the latent features of the secondary channel based on cell shape. We found that Support Vector Regressors (SVRs) performed reasonably well (Fig. 3.20a,b) and were computationally efficient across different conditions. We also tested the alternative moment-based point cloud embedding approach (see Materials & Methods, section 2.3.6), finding that it was also capable of predicting secondary markers but not to the same degree as CBE features (Fig. 3.20a).

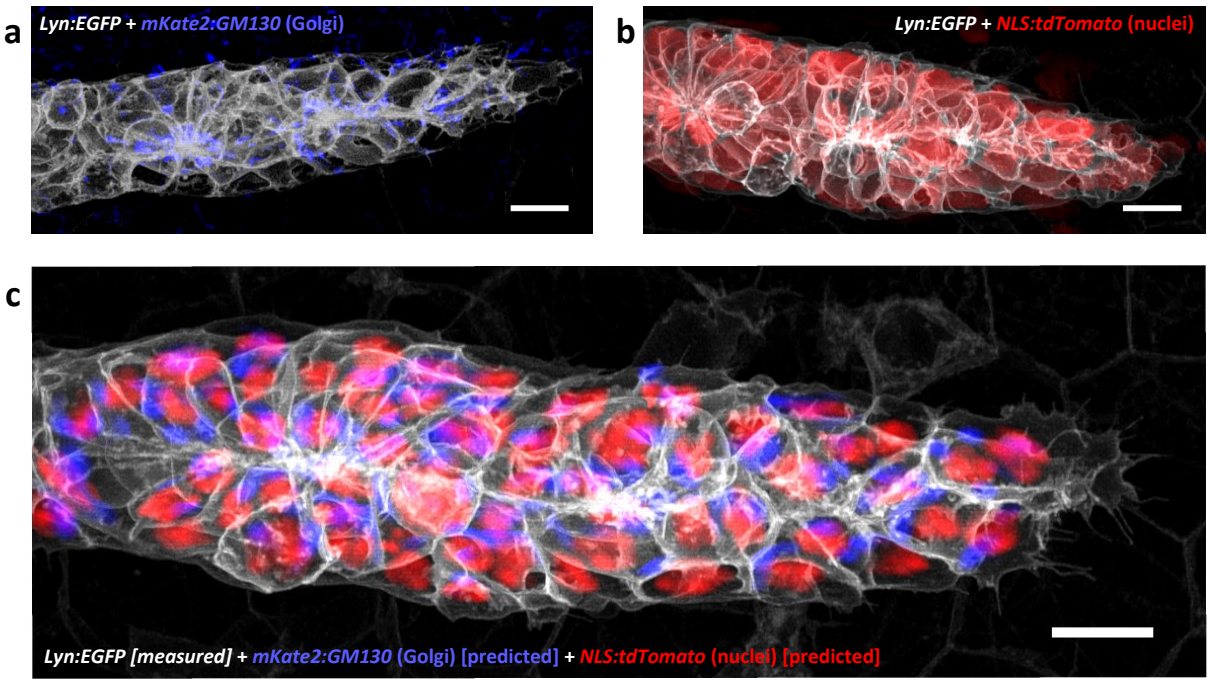

Figure 3.19: Atlas visualization by selective registration and image reconstruction.

(a) Maximum z-projection of a primordium where the Golgi marker *mKate2:GM130* was imaged in addition to the membranes. Scale bar: 10µm. Data acquired by Mie Wong. (b) Maximum z-projection of a primordium where the nuclear marker *NLS:tdTomato* was imaged in addition to the membrane. Scale bar: 10µm. (c) Maximum z-projection of a primordium where only the green membrane marker was imaged. Selective point cloud registration of morphologically similar cells followed by image reconstruction using Gaussian KDE was used to predict an approximate fluorescence distribution of the Golgi and nuclear markers. Scale bar: 10µm.

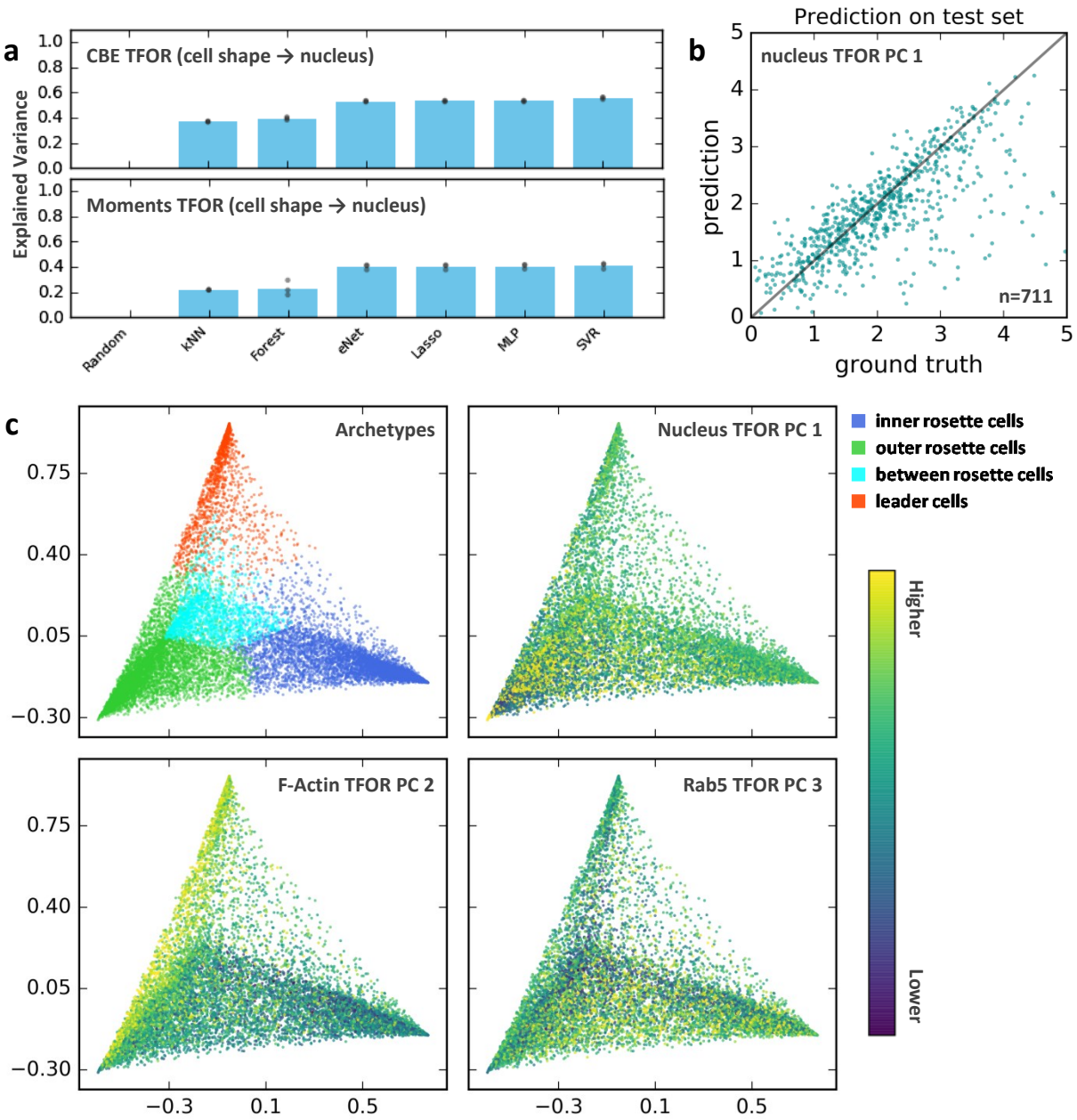

Figure 3.20: Atlas mapping of different markers based on cell shape as a reference.

(a) Example output from the performance evaluation of different machine learning models, showing how well each model can predict nuclear marker latent features based on cell shape latent features, both embedded using CBE (top) or moments (bottom). The different models evaluated are: random sampling from training values as a negative control (random), k-nearest neighbors (kNN), random forest regression (forest), elastic net regression (eNet), Lasso regression (Lasso), a multi-layer perceptron (MLP), and a support vector regressor with an RBF-kernel (SVR). Cyan bars indicate the mean explained variance, dots indicate the results of individual cross-validation runs. (b) Example comparison of ground truth and prediction for a test set of nuclear marker latent features, predicted from cell shape using an SVR regressor. Axes are unitless. (c) Various examples of latent features predicted from different channels and mapped onto the archetype space, showing that a range of potentially biologically relevant patterns can be uncovered through atlas analysis. Predictions were performed based on images of *NLS:tdTomato* (Nucleus), *tagRFP-T:UtrCH* (F-actin) and *mKate2:Rab5* (Rab5). Axes are unitless. Rab5 raw data acquired by Mie Wong.

Training an SVC to predict each of the secondary channels based on cell shape allowed us to generate a complete atlas, meaning that the latent features representing all secondary channels are known for all cells in the entire dataset. This atlas can now be mined for relationships between different marker distributions, between shape and marker distributions, and between biological context and marker distributions. As an example, figure 3.20c shows latent features from different markers mapped onto the archetype space.

3.3.9 Toward Single-Cell Dynamics with Automated Feedback Microscopy

Dynamics are key to cell and tissue self-organization and could be integrated in a myriad ways with single-cell analysis and atlas mapping. However, whilst microscopes like the LSM880 AiryScan can acquire high-quality stacks of migrating tissues with reasonable throughput, they are not fast enough to image multiple locations along an embryo at the high rate required to unambiguously match cells across multiple time points. Thus, the standard approach of stitching multiple locations in order to follow the pLLP's migration along the embryo's tail is not a viable option for obtaining single-cell dynamics.

We developed a solution for this problem using adaptive feedback microscopy, which we termed *prim tracker 880* (pt880). By enabling the microscope to continuously and autonomously track the pLLP and adjust its stage position to follow the migrating tissue (Fig. 3.21) (for details see Materials & Methods, section 2.2.8), pt880 makes it possible to acquire full overnight time courses of primordium migration at a temporal resolution approaching that of imaging just a single static position. Since this high rate of acquisition is beyond what is required even for single-cell tracking and may in fact induce some phototoxicity, imaging in practice is done at a slightly lower rate and the surplus in the time budget is used to track multiple primordia simultaneously, increasing throughput.

At present, the pt880 adaptive feedback microscopy pipeline is fully functional and a number of time course datasets have been generated (not shown). The next step will be to implement single-cell tracking and to integrate the results into the pLLP's cellular shape space.

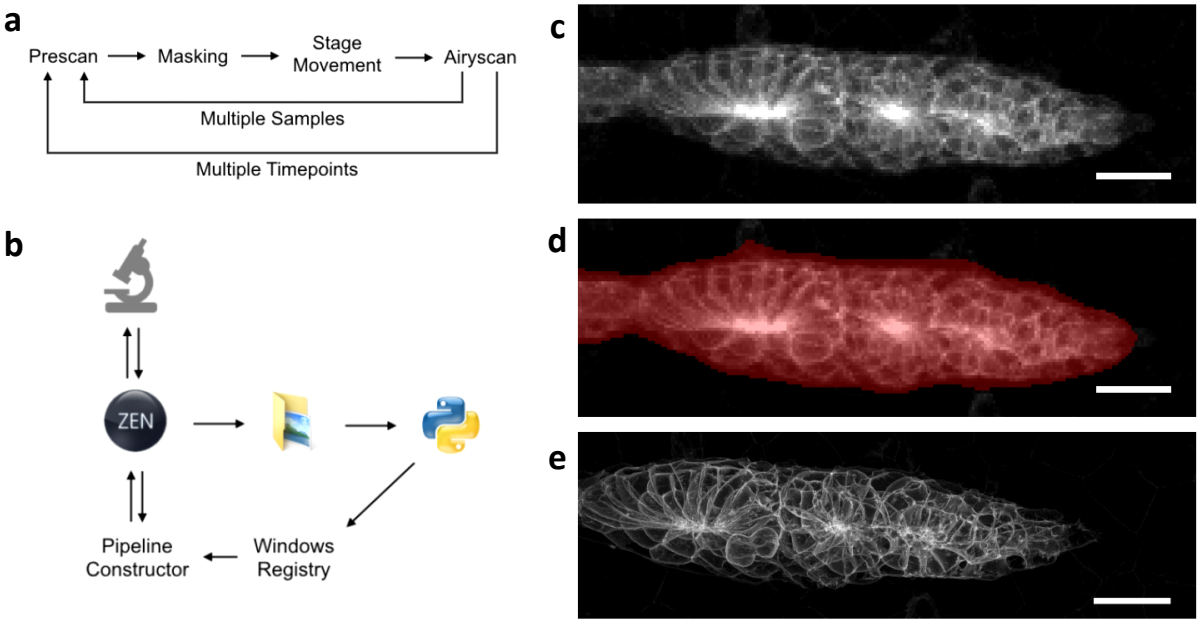

Figure 3.21: Tracking of the pLLP with adaptive feedback microscopy.

(a) Illustration of the basic workflow during acquisition. First, a very fast but low-resolution prescan image is acquired, which is then automatically masked in order to determine how much the primordium has migrated. This migration is then compensated by stage movement and a high-resolution AiryScan stack acquisition is triggered. This process can be repeated across multiple samples in the same dish and over time to generate a time course. (b) Software architecture used to implement pt880. The Zeiss ZEN Black software controls the microscope and is in turn controlled by the MyPiC pipeline constructor macro [Politi et al., 2018]. Saved prescan images are detected by an externally running python script, which performs the image analysis and writes the new coordinates to the Windows registry. The registry is monitored by the pipeline constructor, which forwards the new coordinates to ZEN and triggers the acquisition of a high-resolution stack. (c) An example of a prescan (maximum z-projected). (d) The same prescan with the overlaid pLLP mask in red. (e) The corresponding maximum z-projected high-quality AiryScan stack, acquired after adjustment of the stage. Note that the primordium is shifted backward compared to the prescan. All scale bars are 15µm.

4 Discussion

4.1 Optogenetics Remain Challenging to Adapt to *in vivo* Models

Highly precise and minimally disruptive perturbation tools such as optogenetics promise deep insights into complex biological systems by allowing them to be studied within the region of linear biology. Here, we sought to adapt a range of existing optogenetic tools to gain precise control over chemokine signaling and cytoskeletal organization in the zebrafish posterior lateral line primordium, with the aim of studying the interplay of migration and morphogenesis. However, despite our best efforts, we were so far unsuccessful in deriving a working optogenetic tool.

The conformation-based single-component systems we tested – PA-Rac1, PA-Cxcr4b and bOpsin – could be expressed in the zebrafish pLLP and localized to cell membranes as expected, but did not yield phenotypes in response to either short-term or long-term optical activation. There are a number of possible reasons for this outcome.

PA-Cxcr4b and bOpsin might be hampered by a shortage or absence of their obligatory cofactor, retinal [Xu et al., 2014; Karunarathne et al., 2013a]. Furthermore, our modifications to the published version of PA-CXCR4 [Xu et al., 2014] to convert it to the zebrafish homolog PA-Cxcr4b could potentially have rendered the protein inactive or alternatively may not have been sufficient to allow interaction with zebrafish G-proteins. Similarly, the human bOpsin used here may be unable to activate zebrafish $G\alpha_i$ due to species-specific differences.

The most perplexing case is PA-Rac1, which has already been used successfully in zebrafish neutrophils [Yoo et al., 2012] and neural crest cells [Scarpa et al., 2015]. It is possible but unlikely that its obligatory cofactor, flavin mononucleotide (FMN) [Christie et al., 1999; Herrou & Crosson, 2011], would be far less abundant in the lateral line. Another explanation could be that PA-Rac1 is in fact functional but unable to overcome the natural buffering of the Rac1 regulatory network at play in the pLLP, for example due to a compensatory response of RhoA [Rottner et al., 1999; Williams et al., 2007; Chauhan et al., 2011]. This interpretation would also explain why the constitutively membrane-localized RacGEF *Lyn:tagRFP-T:Tiam1a-Gd* does not cause a phenotype when expressed in lateral line cells, unlike its RhoGEF counterpart *Lyn:tagRFP-T:ARHGEF25b-Gd*.

We also tested the recruitment-based heterodimeric CRY2-CIBN system and found that it is functional in principle: light-dependent recruitment of transiently expressed CRY2 to the cell membrane can be observed in embryos where membrane-anchored CIBN is present. However, in practice it proved impossible to generate CRY2-effector fusion proteins that were expressed at detectable levels in the pLLP. In fact, even generating stable transgenic fish lines of CRY2-effector

constructs was challenging, with low rates of transmission and unusually weak expression of the independent transfection marker.

It is possible that CRY2 exhibits some degree of toxicity in zebrafish, perhaps mediated by scavenging of the cellular pool of its obligatory cofactor flavin adenine dinucleotide (FAD) [Liu et al., 2008], which is also a cofactor for several metabolic enzymes [Mansoorabadi et al., 2007]. As a consequence, cells expressing CRY2 constructs might die or might be outcompeted by others during development. Such toxicity could explain how the initially high transfection rate seen in early embryos following DNA injections (see figure 3.3a) is drastically reduced by the time the lateral line can be imaged. However, it is not clear how toxicity could explain the low number of transgenic founder fish and the low rates of transmission in cases where CRY2 was placed under UAS control and thus should not be expressed in absence of the Gal4 driver. Indeed, the ability to generate transgenic lines with toxic constructs was among the original motivations for the development of the Gal4-UAS transactivation system [Brand & Perrimon, 1993]. Thus, some of the problems we encountered appear to be independent of the CRY2 protein itself.

The potential toxicity of CRY2 in zebrafish appears not to have been addressed in the literature, apart from being mentioned briefly in a study presenting an optogenetic tool for the control of gene expression based on the bacterial LOV protein EL222 [Motta-Mena et al., 2014]. There, the authors claim that a similar system based on CRY2, which had been published two years prior [Liu et al., 2012], exhibits substantial toxicity in zebrafish, unlike their EL222-based system. However, to our reading the earlier study makes no direct mention of toxicity [Liu et al., 2012] and the latter study only presents results on the toxicity of their own construct compared to a GFP control [Motta-Mena et al., 2014]. Either way, it is perhaps telling that neither study presents stable transgenic lines of their tools and that the literature otherwise appears to be devoid of applications of the CRY2-CIBN system in zebrafish to date.

Taken together, our results illustrate that the adaptation of optogenetic tools to an *in vivo* system of choice is still a non-trivial and failure-prone endeavor, even if said tools have been used successfully in other systems. A potential culprit for many of these issues could be the availability of cofactors, although other known and unknown variables may also play a role.

Fortunately, the field of optogenetics continues to progress toward becoming more readily applicable to *in vivo* studies [Johnson & Toettcher, 2018]. For instance, the PhyB-PIF hetero-dimerization system has recently been applied successfully to control nuclear localization [Beyer et al., 2015] and precise subcellular membrane localization [Buckley et al., 2016] in zebrafish embryos, although cofactors are once again a crucial limitation, given that the small molecule cofactor PCB has to be exogenously delivered for PhyB to function. Newly discovered or engineered optogenetic constructs such as the near-infrared tool BphP1-QPAS1 [Redchuck et al., 2017] or the LOV2-derived iLID-SSPB system [Guntas et al., 2015; Johnson et al., 2017] could help alleviate this dependency.

Even CRY2 itself can be a highly effective tool in some *in vivo* models, as evidenced by its successful application in *Drosophila* [Guglielmi et al., 2015; Izquierdo et al., 2018; Krueger et al., 2018].

Despite the roadblocks encountered here, it is likely only a matter of time before optogenetics mature into a "plug and play" part of the developmental biologist's molecular toolbox.

4.2 chemoARHGEF Reveals a Potential Role for Cortical Tension in Rosette Patterning and Morphogenesis

We engineered a constitutively active form of the RhoGEF *ARHGEF25b* under chemical control of expression, termed chemoARHGEF. When mosaically expressed in the pLLP, chemoARHGEF induces cell rounding, consistent with an increase in cortical tension induced through a *RhoA-ROCK-MLC-MyoII* cascade [Lecuit & Lenne, 2007; Schwartz, 2004]. Interestingly, chemoARHGEF-positive cells are also sorted to the center of the tissue, indicating that surface tension-based cell sorting plays a role in rosette organization.

Cell sorting due to differential interfacial tension is a physical process that is thought to serve as a fundamental patterning motif in developing tissues [Brodland, 2002; Lecuit & Lenne, 2007; Fagotto, 2014] and has been invoked to explain a range of processes, including patterning of the *Drosophila* retina [Käfer et al., 2007], germ-layer organization in zebrafish gastrulation [Krieg et al., 2008], and inside-outside patterning of mouse blastocysts [Maître et al., 2016]. It is closely related to sorting based on differential adhesion [Steinberg, 2007]; indeed, the two can be thought of as two sides of the same coin, with cortical tension acting to reduce cell-cell contact interfaces (to minimize the cell surface area by making cells more spherical) and adhesion acting to increase them (to maximize the area of adhesion) [Brodland, 2002; Tepass et al., 2002], although the actual physical process at the subcellular scale may not entirely conform to this simplification [Maître et al., 2012].

Tension-based cell-cell interactions have also been linked to more than just cell sorting. In mouse blastocysts, asymmetric cell division results in an imbalanced distribution of contractile factors, which not only leads to sorting of the more contractile daughter cell to the inside of the embryo but simultaneously affects cell fate, possibly through the *Yap* mechanosensing pathway [Maître et al., 2016]. In the *Drosophila* wing disk, interface contractility between two different tissue domains is classically associated with boundary formation [Landsberg et al., 2009] but has also been shown to intrinsically arise at the interface of any two groups of cells committed to different fates, driving either extrusion and apoptotic elimination of individual cells or cyst formation of small cell groups [Bielmeier et al., 2016].

In the lateral line primordium, rosette formation is thought to be driven by FGF-induced apical constriction [Lecaudey et al., 2008] and the inside-outside patterning of future hair cells at the center

versus the surrounding support and mantle cells is thought to be mediated by a *Notch*-based lateral inhibition mechanism downstream of FGF signaling [Itoh & Chitnis, 2001; Sarrazin et al., 2006; Matsuda & Chitnis, 2010]. Our finding that pLLP cells in the periphery of forming rosettes readily sort to the center in response to increased cortical contractility raises the intriguing possibility that differential tension may also play a role in rosette formation and/or patterning. Based on the above-mentioned wide range of processes known to involve differential tension mechanisms, several such roles (and combinations thereof) are conceivable.

Firstly, differential tension could be involved in the initial establishment of an inside-outside pattern by way of tension-based sorting of a heterogeneous population of cells in the leader-follower transition zone. *Notch*-based lateral inhibition [Matsuda & Chitnis, 2010] would then act only to refine this initial pattern. Similar to what was found in mouse blastomeres [Maître et al., 2016], asymmetric cell division may contribute to the generation of transition zone heterogeneity, since cell divisions of the highly polarized leader cells may yield daughter cells with unequal surface tension.

Secondly, differential tension may act in tandem with apical constriction to implement robust rosette morphogenesis. Highly contractile cells at the center of the rosette may effectively pull the softer peripheral cells inward through a combination of tensile and adhesive forces [Maître et al., 2012], thus contributing to the progression of apical constriction. Such a mechanism could potentially even generate rosette-like structures without the need for apical constriction, making it a putative morphogenetic mechanism that could be at play independently in other systems.

Thirdly, differential tension may be important for the maintenance of physically stable rosette configurations in a dynamic environment. A graded increase in surface tension from the periphery to the center results in a rosette configuration that is mechanically stable whilst still allowing the more fluid-like peripheral cells to absorb environmental forces exerted for example by the tissue's migration. In simple terms, the peripheral cells could act as a "lubricant" that allows the compact rosette core to slide through the embryo.

Finally, differential tension could interact with established patterning systems to modify or enhance them. For instance, the increased surface tension of central rosette cells renders them spherical and thus minimizes their contact area with neighbors. Correspondingly, the average total contact area of a cell with its neighbors would be expected to progressively increase for cells located more toward the periphery of the tissue. This may impact on cell contact signaling, where integration over the total interaction surface is relevant [Shaya et al., 2017]. Furthermore, the physical forces rosette cells exert on each other might directly modify signaling; *Notch* signaling activation in particular is thought to have a mechanical component [Kopan & Ilagan, 2009; Wang & Ha, 2013].

At present, all of these potential functions are hypothetical. Before they can be addressed, further work remains to be done in order to fully establish that chemoARHGEF indeed activates RhoA and

that this indeed induces increased cell surface tension. Complementarily, force measurement experiments (e.g. laser cutting [Smutny et al., 2014]) are required to confirm that cell surface tension is patterned in wild-type rosettes. If this is indeed the case the source of this pattern must be determined, for instance by testing whether RhoA signaling itself is patterned (e.g. using a RhoA activity probe [Stephenson & Miller, 2017]).

Following confirmation of these core assumptions, the functional importance of differential cell surface tension in rosette formation can be investigated. To this end, one important avenue will be to link surface tension to relevant downstream outcomes, such as hair cell fate. Another direction to pursue will be to disentangle the functions of apical constriction and differential tension in rosette morphogenesis, which could be achieved by abolishing apical constriction (e.g. through inhibition of FGF signaling [Lecaudey et al., 2008]) whilst maintaining differential tension (e.g. through chemoARHGEF expression).

In summary, cell surface tension heterogeneity can mediate diverse processes ranging from cell sorting to morphogenesis to fate specification. It may thus also play an important role in the integration of these processes when they take place simultaneously, such as during rosette formation in the pLLP. Here, we established chemoARHGEF as a powerful tool for manipulating cortical tension and we presented preliminary results indicating that tension heterogeneity could indeed be involved in rosette formation. As we continue to elucidate the exact nature of this involvement, we expect to learn more about the integrative functions that interface tension phenomena may play in developing tissues.

4.3 No Conclusive Evidence for Tissue-Scale Feedback of Rosette Morphogenesis on Leader Cell Migration

When cells build organs by simultaneously moving, changing shape and making fate decisions, they have to tightly and robustly coordinate these processes or risk cascade failure and thus rapid divergence from the intended developmental trajectory. Here, we hypothesized that collective migration and rosette assembly in the zebrafish pLLP might be coordinated mechanically, with the deceleration of assembling organs providing an anchoring counterforce to leader cell migration and thereby establishing anisotropic tension across the primordium that could contribute to cell guidance through plithotaxis (see figure 3.5a-b).

A similar mechanism is thought to be at play during *in vitro* monolayer migration [Vitorino & Meyer, 2008; Tambe et al., 2011; Trepat & Fredberg, 2011; Zaritsky et al., 2015] and tension-based interactions in general have been implicated in the organization of migrating collectives in a number of systems, including *Drosophila* border cell migration [Somogyi & Rørth, 2004] and tracheogenesis [Han et al., 2004]. However, what has remained largely unaddressed is the link of such mechanisms

with morphogenesis and differentiation – and thus their potentially important role as coordinators of multiple different developmental processes.

To explore this idea, we first reproduced unpublished prior evidence[1] and confirmed that pLLP migration is partially robust to acute loss of graded chemokine signaling. Specifically, ablation of the follower cell domain expressing *Cxcr7*, the scavenger receptor required to set up the local self-generated chemokine gradient that is thought to guide pLLP migration [Donà et al., 2013], does not in all cases lead to an arrest of the primordium.

The hypothesized tension-based feedback mechanism could explain this robustness of migratory cell polarity, as the plithotactic component of guidance can persist so long as some rosette-assembling followers are left intact. A second possible explanation would be cell-autonomous persistence of migratory polarity, maintained by a self-stabilizing configuration of polarity components and the cytoskeleton [Pegtel et al., 2007; Krause & Gautreau, 2014]. However, in previous experiments where all follower cells had been ablated, leaders immediately ceased to migrate [unpublished data1], which indicates that they do not possess such autonomous persistence. There is also a third explanation for this outcome, namely that our ablation is in fact incomplete and a small pool of *Cxcr7* remains in follower cells that do not show detectable levels of our *Cxcr7* transcriptional reporter. This pool could be sufficient to uphold a shallow gradient and thus maintain leader cell polarity – which is why we next sought to test our hypothesis with experiments that alter tissue-tension directly.

To do so, we made use of 2P-Arrest, a phenomenon we serendipitously discovered and subsequently optimized for this purpose. A few minutes of irradiation with 2-photon laser light at levels above what is ordinarily used for 2-photon imaging but below what would cause ablation reliably induces an arrest of migration in pLLP cells. Interactions of laser light with biological tissues are complicated [Niemz, 2013], so determining the exact mechanism underlying this treatment is non-trivial. The effect is reminiscent of laser cauterization, where the same kind of irradiation is used to "glue" Drosophila cells to the adjacent vitelline membrane [Collinet et al., 2015; Rauzi et al., 2015]. However, since 2P-Arrest does not immediately stop cells but rather causes a gradual deceleration of migration, a different effect must be at play. Consistent with our observation that irradiated cells eventually enter apoptosis (but not until several hours after the treatment), we suspect that laser-induced DNA damage triggers pLLP cells to terminate their migratory prograem. Further experiments such as testing for p53 and caspase-3 activation [Lee et al., 2007] or performing 2P-Arrest in embryos mutant for p53 [Berghmans et al., 2005] will be required to support this hypothesis. Regardless of its specific mechanism of action, 2P-Arrest can serve as a light-based precision perturbation tool to arrest migratory cells or cell groups *in vivo* and may be useful beyond the lateral line primordium.

[1] These experiments were performed by two former members of the lab, Petra Haas and Sebastian Streichan.

By specifically arresting pLLP follower cells and observing the behavior of leader cells in response, we were able to directly test the hypothesis that follower cell deceleration impacts leader cell behavior. We quantitatively assessed membrane and actin dynamics of leader cells during follower 2P-Arrest but found no clear evidence of an active response. The only effect we could detect is an elongation of leader cells' lamellipodial projections. However, this outcome does not necessarily indicate an active response, since lamellipodium extension is driven by branched actin assembly at the leading edge as a sort of "front-wheel drive" [Ridley, 2011]; leader cell stretching and lamellipodium elongation would thus also be consistent with a passive physical effect resulting from the increased rearward anchoring force.

In conclusion, we found no clear evidence supporting a model of feedback coordination between rosette assembly and primordium migration. Although further experiments are possible (for example, we have not looked into leader cell myosin dynamics during follower arrest) and although the robustness of pLLP migration to *Cxcr7*-domain ablation remains to be explained, our current results do not support the proposal that plithotactic coupling helps guide the lateral line primordium.

4.4 Image-Based Quantitative Analysis of Cellular Architecture in a Developing Tissue

For data-driven approaches to succeed in accelerating the progress of biology, comprehensive quantitative descriptions of biological systems are an essential starting point. However, capturing not only the constituent biochemical components of multi-cellular tissues but also their intricate multi-scale architecture and their complex dynamics represents a major challenge.

We developed a computational framework for image-based quantitative characterization of cellular architecture in living tissues. Making use of AiryScan microscopy [Huff, 2016], we acquired high-quality 3D confocal volumes of membrane-labeled lateral line primordia, which we then automatically segmented into individual cells. Next, we employed a novel computational approach to embed cell segmentations or arbitrary subcellular fluorescence intensity distributions into compact and expressive latent feature spaces ready to be explored by data-scientific methods.

Re-representing image data in a more useful form is a non-trivial problem. Ideally, potential solutions would meet most or all of the following criteria: they capture all relevant variance in the input image, they allow rejection of non-relevant noise, they allow rotational variance and size variance to be factored out, they work for arbitrary shapes or intensity distributions, they can be reversed (a matching image can be reconstructed from any point in the latent feature space), they consist of mathematically tractable steps, and they are computationally efficient. To date, no existing approach readily meets these criteria, especially not for 3D volumetric data.

PCA has been established as a powerful tool for performing reversible and tractable data re-representation into features that are expressive and ordered by their contribution to the variance in the population, allowing noise to be rejected [Adams et al., 2004; Pincus & Theriot, 2007; Weight et al., 2008; Barnhart et al., 2011]. Although PCA can be applied directly to segmentation outlines or images, it does not natively account for rotational and size variance and it is unclear how well it could handle arbitrary fluorescence intensity distributions [Pincus & Theriot, 2007].

Here, we implemented a novel approach inspired by geometric morphometrics [Adams et al., 2004; Adams et al., 2013] that combines the advantages of using a point cloud representation with the aforementioned strengths of PCA. We use a tunable sampling algorithm (ISLA) to convert images to sparse but information-preserving point clouds and subsequently re-represent these clouds relative to common reference points (CBE). The resulting feature space is then transformed by PCA, which yields expressive latent features ordered by their contribution to the population's variance.

Our pipeline can take arbitrary intensity distributions as well as cell segmentations as inputs and can explicitly remove rotational variance through an intermediate pairwise distance re-representation step performed on the point cloud. It does not, however, natively provide reversibility from the latent feature space back to a point cloud or image representation. Further validation will be required to determine the extent of the advantages and drawbacks of our proposed method more clearly, in particular with respect to the implications of using a pairwise distance re-representation to achieve rotational invariance.

We applied our morphometric approach to characterize cell shape across a large dataset of 190 segmented primordia. If rotational symmetry and size differences were not removed (TFOR), they together explained over 70% of variance in the dataset, followed by shape parameters such as protrusion length and sphericity. Whilst the importance of the rotational component is easily explained by the mirror symmetry of the pLLP, it is interesting to note that cell size seems to mainly vary along the apico-basal axis (the height of the cell); this may be a consequence of the more flat and protrusive (mesenchyme-like) state of the leaders being contrasted with the more packed and apically polarized (epithelial-like) state of the followers [Pastor-Pareja, 2004; Lecaudey et al., 2008; Fischer et al., 2009; Friedl & Gilmour, 2009].

After correcting for size and orientation (CFOR), cell sphericity clearly dominates the population's variation within the shape space. As sphericity is closely linked with cell surface tension and adhesion [Matzke, 1946; Tepass et al., 2002; Lecuit & Lenne, 2007], this finding reinforces the need for a closer investigation of effective surface tension as a potential key parameter for tissue self-organization in the pLLP, as discussed in section 4.2 based on the chemoARHGEF phenotype. One concern to be addressed in this context is that leader cells or transition zone cells are ascribed relatively high values of sphericity in the shape space analysis, which is unexpected given the sorting bias of chemoARHGEF-positive cells to the center of the tissue. This observation could potentially be an

artifact caused by the impossibility of capturing leaders' thin protrusions projected forward directly underneath their neighbors, which may cause leader cell sphericity to be overestimated. The use of more specific measures of sphericity, such as circularity along different axes, will likely help resolve this issue.

In summary, we have developed a framework for imaging, single-cell segmentation, latent feature space embedding, and quantitative analysis of developing tissues *in vivo* and have applied it in a proof-of-concept study to chart the cellular shape space of the pLLP. We anticipate that image-based large-scale characterization of multi-cellular systems will provide a useful basis for data integration and data-driven biological inference (see section 4.5) and we expect that further analysis of the pLLP shape space – in particular with regard to cortical tension and in tandem with perturbation experiments using chemoARHGEF – will yield new insights into the principles underlying the tissue architecture of migrating cell collectives.

4.5 The Cellular Shape Space as a Reference for Data Integration and Context-Sensitive Analysis

Deriving meaningful mechanistic and human-interpretable conclusions and predictions from big data is extremely challenging [Holzinger et al., 2014]. Moving from data that is *big* to data that is *rich* (i.e. multi-modal data that integrates different types of information) as well as *exploitable* (i.e. data that can be queried and visualized in a variety of useful ways) is a crucial step in overcoming this challenge.

Data integration has been an active field of research for some time but has recently intensified in the life sciences as a consequence of the advent of multi-omics technologies [Gomez-Cabrero et al., 2014; Bersanelli et al., 2016]. In these cases, integration is usually based on matched nodes in a network, for instance matched patient samples that were analyzed with different omics approaches [Argelaguet et al., 2018] or matched genes for whom interactions have been quantified in different organisms [Szklarczyk et al., 2017].

By contrast, image data integration is usually performed by spatial registration of multiple specimens, which is a special case of data integration based on a common reference measurement, an approach we refer to as *atlas mapping*. Although registration has enabled a number of impressive successes [Vergara et al., 2017; Cai et al., 2018], its requirement for stereotypically shaped specimens presents a considerable limitation. Thus, other approaches to atlas mapping are now being explored, including the use of generative deep learning models [Johnson et al., 2017].

Here, we present a machine learning-based method for atlas mapping of single-cell data that is based on an embedded cellular shape space as a reference measurement. We move away from attempting

to perform atlas mapping at the level of actual image data and instead restrict ourselves to mapping between feature spaces, which is ultimately more important for downstream data analysis techniques. This decision enabled us to use straightforward classical machine learning techniques (in particular Support Vector Regression) to generate a proof-of-concept atlas of several different fluorescent marker, including markers for F-actin, cell nuclei, and different endomembrane compartments – all based on as few as 15 to 20 primordia per marker, which allows our approach to be applied to developing tissues *in vivo*, despite the limitations in throughput associated with such model systems.

Interestingly, this approach is not limited to integrating fluorescence intensity distributions. In future, different types of information could be included into the atlas by learning their relationship with the cellular shape space.

A very simple example would be to perform single-molecule fluorescence in-situ hybridization (smFISH), which allows the exact and spatially resolved quantification of gene expression levels [Raj et al., 2008]. Since a membrane marker can be imaged simultaneously, smFISH spots can easily be assigned to segmented cells and a regressor can be trained to predict spot counts from shape space information. In a second step, this approach could be extended to entire transcriptomes by means of single-cell RNA-seq [Wu et al., 2013], the results of which can be mapped back into the primordium based on a selected set of references genes whose distribution has been fully characterized by smFISH. A similar approach has already been employed elsewhere [Satija et al., 2015; Achim et al., 2015], albeit again based on image registration rather than latent feature regression.

Another very interesting possibility is to map cortical tension across the primordium. This could be achieved either by combining membrane imaging and segmentation with Brillouin microscopy [Scarcelli & Yun, 2007] or by performing laser-cutting experiments across many samples [Smutny et al., 2015]. In both cases, an acquired pilot dataset would be used to train a machine learning model to learn the relationship of interface tension with the combined shape space of the two interfacing cells. The trained model could then be used to predict tension for all interfaces and for any primordium for which cellular shape space information is available. This idea represents a data-driven alternative to established physics-based approaches for inferring cell surface tension from cell shape alone [Chiou et al., 2012; Brodland et al., 2014].

Perhaps the most important dimension to integrate into the dataset is time. Based on a set of fully tracked time courses, the cellular shape space – along with any other properties mapped onto it – could be extended to include the temporal trajectories of cell states throughout pLLP development. Here, we presented the *prim tracker 880*, an adaptive feedback microscopy tool that allows the acquisition of high-quality AiryScan time lapses with high temporal resolution. It should be relatively straightforward to perform automated single-cell tracking on such data, thus adding an arrow of time to the pLLP's shape space.

On its own, simply making big data even bigger by integrating additional datasets may not be sufficient to help derive biologically meaningful conclusions. We propose three ways in which a cell atlas such as the one described here could also be made more exploitable.

Firstly, supervised machine learning can be used to introduce human-interpretable biological context to the dataset and re-represent the data accordingly. We showed an example of this in the form of morphological archetypes. By manually categorizing pLLP cells into groups that make sense from a human researcher's perspective, the entire dataset could be transformed to reflect its relationship with these user-defined classes, making it immediately more interpretable. A similar approach could for instance be used to investigate the behavior of dividing cells within the pLLP, which would be especially interesting if time course data is included in the atlas, as it may for instance reveal information about possible asymmetric cell divisions in the primordium's transition zone (see section 4.2).

Secondly, context-sensitive analysis can help unpack complicated relationships between different aspects of the dataset. When a specific question is asked, such as whether two features of the atlas correlate, a classical correlation analysis will often be confounded by a multitude of independent sources of variation. In such cases, information within the rest of the dataset can provide the context required to distill the relationship of interest. For instance, one might ask whether two proteins of interest co-localize in the cells of the pLLP. However, a simple co-localization analysis across all cells might fail to reveal that the two proteins in fact co-localize exclusively within outer rosette cells. By looking for correlations within specific contexts of the atlas, for instance within each morphological archetype, such context-specific relationships become visible. It may even be possible to automate such queries, returning both the detected relationships between two features of interest and the relevant context.

Finally, there is little doubt that a first-principle mathematical model of the lateral line (based e.g. on a GGH framework [Swat et al., 2012]) would be invaluable for exploring and understanding the interplay of physics and cell signaling that underlies its self-organization. However, modeling and simulation of complex multi-cellular systems is technically and conceptually challenging, especially when limited information is available to guide choices as to how various sub-cellular processes should be simplified [Brodland, 2015]. A comprehensive cell atlas provides ample quantitative information and would thus be a useful aide during model generation, parameter estimation, and model evaluation. In essence, the cellular atlas represents a data-driven model of the tissue, which can be approached from first principles in order to arrive at a mechanistic and interpretable mathematical model. In the (far?) future, this process may eventually be automated in its entirety.

Overall, our computational efforts to quantitatively characterize the pLLP's cellular architecture have led to a number of proof-of-concept tools for the integration and analysis of rich biological datasets. As the cell atlas of the lateral line primordium continues to grow, these tools as well as our other

proposals will be put to the test. Hopefully, they will enable the data-driven discovery of new insights into tissue self-organization.

4.6 Concluding Remarks

Recently, discussions have unfolded about whether molecular and developmental biology are close to being "complete", leaving nothing but scraps for basic researchers and implying the necessity to increasingly refocus our efforts toward engineering and design [Woese, 2004; Cohen, 2017; Grewal, 2018].

Historically, speculations along these lines have come up repeatedly in various fields, including physics [Michelson, 1903; Hawking, 1993] and molecular biology itself [Stent, 1969]. They usually fell by the wayside rather quickly as new and exciting discoveries continued to propel those fields forward.

With respect to the current iteration of this debate, we feel that our modern understanding of biology is in fact far from complete. Even if we are approaching a point where most of the fundamental units of life – including genes and biochemical mechanisms – are known and relatively well understood, the secrets of life lie not in those components alone but in their dynamic interplay across several scales. Life is not just substance; it is information.

Here, we sought to tackle some of the challenges that arise from trying to understand multi-cellular systems from an integrated rather than a reductionist perspective. Complexity and its methodological consequence is one such challenge. The fact that big data alone does not readily lead to mechanistic models, let alone to human understanding, is another. The work presented in this thesis identifies a number of possible paths toward overcoming these challenges, both in terms of biological concepts and in terms of methodological approaches. However, it also reflects how long the road still is and how small our steps along the road tend to be.

There are several additional challenges we have so far left unaddressed, including the need for improved theoretical frameworks to quantitatively yet intuitively model and predict higher-order biological systems, or the need for more integration between mechanistic and evolutionary biology to bring more generality to the former and more applicability to the latter. Even when it comes to engineering and design, we anticipate that emergent biological phenomena will be both a potential hindrance (for attempts to program biological systems as one would program a computer) and an incredible opportunity (when the self-organizing principles of biology can be harnessed successfully).

In conclusion, we ask those worried about the imminent completeness of our field to turn their eyes toward higher-order living systems. We would be surprised to find that they continue to worry.

References

Achim, K., Pettit, J.-B., Saraiva, L. R., Gavriouchkina, D., Larsson, T., Arendt, D., Marioni, J. C. (2015): *High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin*, Nature Biotechnology 33(5): 503-509.

Adams, D. C., Rohlf, F. J., Slice, D. E. (2004): *Geometric morphometrics: Ten years of progress following the "revolution"*, Italian Journal of Zoology 71(1): 5-16.

Adams, D. C., Rohlf, F. J., Slice, D. E. (2013): *A field comes of age: geometric morphometrics in the 21st century*, Hysterix 24 (1): 7-14.

Affolter, M., Weijer, C. J. (2005): *Signaling to Cytoskeletal Dynamics during Chemotaxis*, Developmental Cell 9(1): 19-34.

Aman, A., Piotrowski, T. (2008): *Wnt/β-Catenin and Fgf Signaling Control Collective Cell Migration by Restricting Chemokine Receptor Expression*, Developmental Cell 15(5): 749-761.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002): *Cell Communication*, Molecular Biology of the Cell: 4th edition, Garland Science.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J., (1997): *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*, Nucleic Acids Research 25: 3389-3402.

Alverez-Melis, D., Jaakkola, T. S. (2018): *Towards Robust Interpretability with Self-Explaining Neural Networks*, arXiv: 1806.07538v1 [preprint].

Aman, A., Piotrowski, T. (2008): *Wnt/β-Catenin and Fgf Signaling Control Collective Cell Migration by Restricting Chemokine Receptor Expression*, Developmental Cell 15(5): 749-761.

Amy, R. L., Storb, R. (1965): *Selective Mitochondrial Damage by a Ruby Laser Microbeam: An Electron Microscopic Study*, Science 150(3697): 756-758.

Angerer, P., Simon, L., Tritschler, S., Wolf, F. A., Fischer, D., Theis, F. J. (2017): *Single cells make big data: New challenges and opportunities in transcriptomics*, Current Opinion in Systems Biology 4: 85-91.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., Stegle, O. (2018): *Multi-Omics Factor Analysis — a framework for unsupervised integration of multi-omics data sets*, Molecular Systems Biology 14(6): e8124.

Artavanis-Tsakonas, S., Rand, M. D., Lake, R. J. (1999): *Notch Signaling: Cell Fate Control and Signal Integration in Development*, Science 284(5415): 770-776.

Artegiani, B., Clevers, H. (2018): *Use and application of 3D-organoid technology*, Human Molecular Genetics 27/R2: 99-107.

Ashe, H. L., Briscoe, J. (2006): *The interpretation of morphogen gradients*, Development 133(3): 385-394.

Auer, T. O., Del Bene, F. (2014): *CRISPR/Cas9 and TALEN-mediated knock-in approaches in zebrafish*, Methods 69(2): 142-150.

Averbukh, I., Gavish, A., Shilo, B.-Z., Barkai, N. (2017): *Dealing with noise: The challenge of buffering biological variability*, Current Opinion in Systems Biology 1: 69-74.

Badrinarayanan, V., Kendall, A., Cipolla, R. (2017): *SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation*, IEEE Transactions on Pattern Analysis and Machine Intelligence 39(12): 2481-2495.

Barnhart, E. L., Lee, K.-C., Keren, K., Mogilner, A., Theriot, J. A. (2011): *An Adhesion-Dependent Switch between Mechanisms That Determine Motile Cell Shape*, PLoS Biology 9(5): e1001059.

Barrett, A. B., Barnett, L., Seth, A. K. (2010): *Multivariate Granger causality and generalized variance*, Physical Review E 81(4).

Belmonte, J. M., Leptin, M., Nédélec, F. (2017): *A theory that predicts behaviors of disordered cytoskeletal networks*, Molecular Systems Biology 13(9): 941.

Belousov, B. P. (1958): *A periodic reaction and its mechanism,* in *Collection of short papers on radiation medicine for 1958*, Med. Publ.

Berghmans, S., Murphey, R. D., Wienholds, E., Neuberg, D., Kutok, J. L., Fletcher, C. D. M., Morris, J. P., Liu, T. X., Schulte-Merker, S., Kanki, J. P., Plasterk, R., Zon, L. I., Look, A. T. (2005): *tp53 mutant zebrafish develop malignant peripheral nerve sheath tumors*, Proceedings of the National Academy of Sciences 102(2): 407-412.

Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., Milanesi, L. (2016): *Methods for the integration of multi-omics data: mathematical aspects*, BMC Bioinformatics 17(S2): 15.

Bertet, C., Sulak, L., Lecuit, T. (2004): *Myosin-dependent junction remodelling controls planar cell intercalation and axis elongation*, Nature 429(6992): 667-671.

Betz, C., Lenard, A., Belting, H.-G., Affolter, M. (2016): *Cell behaviors and dynamics during angiogenesis*, Development 143(13): 2249-2260.

Beyer, H. M., Juillot, S., Herbst, K., Samodelov, S. L., Müller, K., Schamel, W. W., Römer, W., Schäfer, E., Nagy, F., Strähle, U., Weber, W., Zurbriggen, M. D. (2015): *Red Light-Regulated Reversible Nuclear Localization of Proteins in Mammalian Cells and Zebrafish*, ACS Synthetic Biology 4(9): 951-958.

Bielmeier, C., Alt, S., Weichselberger, V., La Fortezza, M., Harz, H., Jülicher, F., Salbreux, G., Classen, A.-K. (2016): *Interface Contractility between Differently Fated Cells Drives Cell Elimination and Cyst Formation*, Current Biology 26(5): 563-574.

Blanchoin, L., Boujemaa-Paterski, R., Sykes, C., Plastino, J. (2014): *Actin Dynamics, Architecture, and Mechanics in Cell Motility*, Physiological Reviews 94(1): 235-263.

Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G., Deisseroth, K. (2005): *Millisecond-timescale, genetically targeted optical control of neural activity*, Nature Neuroscience 8(9): 1263-1268.

Brand, A. H., Perrimon, N. (1993): *Targeted gene expression as a means of altering cell fates and generating dominant phenotypes*, Development 118(2): 401-415.

Brenner, S. (1974); *The genetics of Caenorhabditis elegans*, Genetics 77(1): 71-94.

Briscoe, J., Small, S. (2015): *Morphogen rules: design principles of gradient-mediated embryo patterning*, Development 142(23): 3996-4009.

Brodland, G. W. (2002): *The Differential Interfacial Tension Hypothesis (DITH): A Comprehensive Theory for the Self-Rearrangement of Embryonic Cells and Tissues*, Journal of Biomechanical Engineering 124(2): 188-197.

Brodland, G. W. (2015): *How computational models can help unlock biological systems*, Semin. Cell Dev. Biol. 47-48: 62-73.

Brodland, G. W., Veldhuis, J. H., Kim, S., Perrone, M., Mashburn, D., Hutson, M. S. (2014): *CellFIT: A Cellular Force-Inference Toolkit Using Curvilinear Cell Boundaries*, PLoS ONE 9(6): e99116.

Buckley, C. D., Tan, J., Anderson, K. L., Hanein, D., Volkmann, N., Weis, W. I., Nelson, W. J., Dunn, A. R. (2014): *The minimal cadherin-catenin complex binds to actin filaments under force*, Science 346(6209): 1254211-1254211.

Buckley, C. E., Moore, R. E., Reade, A., Goldberg, A. R., Weiner, O. D., Clarke, J. D. W. (2016): *Reversible Optogenetic Control of Subcellular Protein Localization in a Live Vertebrate Embryo*, Developmental Cell 36(1): 117-126.

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C., Stegle, O. (2015): *Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells*, Nature Biotechnology 33(2): 155-160.

Buldyrev, S. V., Parshani, R., Paul, G., Stanley H. E., Havlin, S. (2010): *Catastrophic cascade of failures in interdependent networks*, Nature 464: 1025-1028.

Burger, A., Lindsay, H., Felker, A., Hess, C., Anders, C., Chiavacci, E., Zaugg, J., Weber, L. M., Catena, R., Jinek, M., Robinson, M. D., Mosimann, C. (2016): *Maximizing mutagenesis with solubilized CRISPR-Cas9 ribonucleoprotein complexes*, Development 143(11): 2025-2037.

Burkel, B. M., von Dassow, G., Bement, W. M. (2007): *Versatile fluorescent probes for actin filaments based on the actin-binding domain of utrophin*, Cell Motility and the Cytoskeleton 64(11): 822-832.

Burns, J. M., Summers, B. C., Wang, Y., Melikian, A., Berahovich, R., Miao, Z., Penfold, M. E. T., Sunshine, M. J., Littman, D. R., Kuo, C. J., Wei, K., McMaster, B. E., Wright, K., Howard, M. C., Schall, T. J. (2006): *A novel chemokine receptor for SDF-1 and I-TAC involved in cell survival, cell adhesion, and tumor development*, The Journal of Experimental Medicine 203(9): 2201-2213.

Cabrera, V. C. (1990): *Lateral inhibition and cell fate during neurogenesis in Drosophila: the interactions between scute, Notch and Delta*, Development 109: 733-742.

Cai, Y., Hossain, M. J., Hériché, J.-K., Politi, A. Z., Walther, N., Koch, B., Wachsmuth, M., Nijmeijer, B., Kueblbeck, M., Martinic-Kavur, M., Ladurner, R., Alexander, S., Peters, J.-M., Ellenberg, J. (2018): *Experimental and computational framework for a dynamic protein atlas of human cell division*, Nature.

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., Collins, J. J. (2018): *Next-Generation Machine Learning for Biological Networks*, Cell 173(7): 1581-1592.

Casey, P. J., Seabra, M. C. (1996): *Protein Prenyltransferases*, Journal of Biological Chemistry 271(10): 5289-5292.

Caussinus, E., Colombelli, J., Affolter, M. (2008): *Tip-Cell Migration Controls Stalk-Cell Intercalation during Drosophila Tracheal Tube Elongation*, Current Biology 18(22): 1727-1734.

Cearns, M. D., Escuin, S., Alexandre, P., Greene, N. D. E., Copp, A. J. (2016): *Microtubules, polarity and vertebrate neural tube morphogenesis*, Journal of Anatomy 229(1): 63-74.

Charras, G., Paluch, E. (2008): *Blebs lead the way: how to migrate without lamellipodia*, Nature Reviews Molecular Cell Biology 9(9): 730-736.

Charras, G., Yap, A. S. (2018): *Tensile Forces and Mechanotransduction at Cell–Cell Junctions*, Current Biology 28(8): R445-R457.

Chauhan, B. K., Lou, M., Zheng, Y., Lang, R. A. (2011): *Balanced Rac1 and RhoA activities regulate cell shape and drive invagination morphogenesis in epithelia*, Proceedings of the National Academy of Sciences 108(45): 18289-18294.

Chen, N. X., Geist, D. J., Genetos, D. C., Pavalko, F. M., Duncan, R. L. (2003): *Fluid shear-induced NFκB translocation in osteoblasts is mediated by intracellular calcium release*, Bone 33(3): 399-410.

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M., Xie, W., Rosen, G. L., Lengerich, B. J., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A. E., Shrikumar, A., Xu, J., Cofer, E. M., Lavender, C. A., Turaga, S. C., Alexandari, A. M., Lu, Z., Harris, D. J., DeCaprio, D., Qi, Y., Kundaje, A., Peng, Y., Wiley, L. K., Segler, M. H. S., Boca, S. M., Swamidass, S. J., Huang, A., Gitter, A., Greene, C. S. (2018): *Opportunities and obstacles for deep learning in biology and medicine*, Journal of The Royal Society Interface 15(141): 20170387.

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., Blaschke, T. (2018): *The rise of deep learning in drug discovery*, Drug Discovery Today 23(6): 1241-1250.

Chen, Y., Medioni, G. (1992): *Object modelling by registration of multiple range images*, Image and Vision Computing 10(3): 145-155.

Chiou, K. K., Hufnagel, L., Shraiman, B. I. (2012): *Mechanical Stress Inference for Two Dimensional Cell Arrays*, PLoS Computational Biology 8(5): e1002512.

Christie, J. M., Salomon, M., Nozue, K., Wada, M., Briggs, W. R. (1999): *LOV (light, oxygen, or voltage) domains of the blue-light photoreceptor phototropin (nph1): Binding sites for the chromophore flavin mononucleotide*, Proceedings of the National Academy of Sciences 96(15): 8779-8783.

Chrzanowska-Wodnicka, M., Burridge, K. (1996): *Rho-stimulated contractility drives the formation of stress fibers and focal adhesions*, The Journal of cell biology 133(6): 1403-1415.

Cohen, B. A. (2017): *How should novelty be valued in science?*, eLife 6: e28699.

Cohen, J. (1988): *Statistical Power Analysis for the Behavioral Sciences: 2nd edition*, Lawrence Erlbaum Associates.

Collinet, C., Rauzi, M., Lenne, P.-F., Lecuit, T. (2015): *Local and tissue-scale forces drive oriented junction growth during tissue extension*, Nature Cell Biology 17(10): 1247-1258.

Conrad, C., Wünsche, A., Tan, T. H., Bulkescher, J., Sieckmann, F., Verissimo, F., Edelstein, A., Thomas, W., Liebel, U., Pepperkok, R., Ellenberg, J. (2011): *Micropilot: automation of fluorescence microscopy–based imaging for systems biology*, Nature Methods 8(3): 246-249.

Cordelières, F., Schindelin, J., Hiner, M., Kazimiers, T. (2017): *Manual Tracking Fiji Plugin*, imagej.net/Manual_Tracking [online; accessed 2018-10-29].

Dalle Nogare, D., Somers, K., Rao, S., Matsuda, M., Reichman-Fried, M., Raz, E., Chitnis, A. B. (2014): *Leading and trailing cells cooperate in collective migration of the zebrafish posterior lateral line primordium*, Development 141(16): 3188-3196.

Dambly-Chaudière, C., Cubedo, N., Ghysen, A. (2007): *Control of cell migration in the development of the posterior lateral line: antagonistic interactions between the chemokine receptors CXCR4 and CXCR7/RDC1*, BMC Developmental Biology 7(1).

Dask Development Team (2016): *Dask: Library for dynamic task scheduling*, dask.org [online; accessed 2018-10-29].

David, N. B., Sapède, D., Saint-Etienne, L., Thisse, C., Thisse, B., Dambly-Chaudière, C., Rosa, F. M., Ghysen, A. (2002): *Molecular basis of cell migration in the fish lateral line: role of the chemokine receptor CXCR4 and of its ligand, SDF1*, Proceedings of the National Academy of Sciences 99(25): 16297-16302.

Davies, J. (2017): *Using synthetic biology to explore principles of development*, Development 144: 1146-1158.

De Curtis, I., Meldolesi, J. (2012): *Cell surface dynamics - how Rho GTPases orchestrate the interplay between the plasma membrane and the cortical cytoskeleton*, Journal of Cell Science 125(19): 4435-4444.

Dean, W. (2016): *Computational Complexity Theory*, The Stanford Encyclopedia of Philosophy (Winter 2016 Edition), plato.stanford.edu/entries/computational-complexity.

Deng, Y., Bao, F., Dai, Q., Wu, L., Steven Altschuler (2018): *Massive single-cell RNA-seq analysis and imputation via deep learning*, bioRxiv: 10.1101/315556 [preprint].

Dhamala, M., Rangarajan, G., Ding, M. (2008): *Analyzing information flow in brain networks with nonparametric Granger causality*, NeuroImage 41(2): 354-362.

Diego, X., Marcon, L., Müller, P., Sharpe, J. (2018): *Key Features of Turing Systems are Determined Purely by Network Topology*, Phys. Rev. X 8.

Dijkgraaf, S. (1963): *The functioning and significance of the lateral-line organs*, Biological Reviews 38(1): 51-105.

Distel, M., Wullimann, M. F., Koster, R. W. (2009): *Optimized Gal4 genetics for permanent gene expression mapping in zebrafish*, Proceedings of the National Academy of Sciences 106(32): 13365-13370.

Donà, E., Barry, J. D., Valentin, G., Quirin, C., Khmelinskii, A., Kunze, A., Durdu, S., Newton, L. R., Fernandez-Minan, A., Huber, W., Knop, M., Gilmour, D. (2013): *Directional tissue migration through a self-generated chemokine gradient*, Nature 503(7475): 285-289.

Dormann, D., Weijer, C. J. (2003): *Chemotactic cell movement during development*, Current Opinion in Genetics & Development 13(4): 358-364.

Doshi-Velez, F., Kim, B. (2017): *Towards A Rigorous Science of Interpretable Machine Learning*, arXiv: 1702.08608v2 [preprint].

Dupont, S., Morsut, L., Aragona, M., Enzo, E., Giulitti, S., Cordenonsi, M., Zanconato, F., Le Digabel, J., Forcato, M., Bicciato, S., Elvassore, N., Piccolo, S. (2011): *Role of YAP/TAZ in mechanotransduction*, Nature 474(7350): 179-183.

Durdu, S., Iskar, M., Revenu, C., Schieber, N., Kunze, A., Bork, P., Schwab, Y., Gilmour, D. (2014): *Luminal signalling links cell communication to tissue architecture during organogenesis*, Nature 515(7525): 120-124.

Eichler, M. (2012): *Causal Inference in Time Series Analysis*, Wiley Series in Probability and Statistics: 327-354.

Ellis-Davies, G. C. R. (2007): *Caged compounds: photorelease technology for control of cellular chemistry and physiology*, Nature Methods 4(8): 619-628.

Emelyanov, A., Parinov, S. (2008): *Mifepristone-inducible LexPR system to drive and control gene expression in transgenic zebrafish*, Developmental Biology 320(1): 113-121.

Evans, M. (2018): *The Machine Learning Revolution: How Artificial Intelligence Could Transform Your Business*, Forbes, forbes.com/sites/allbusiness/2018/10/20/machine-learning-artificial-intelligence-could-transform-business [online, accessed: 2018-11-26].

Fagotto, F. (2014): *The cellular basis of tissue separation*, Development 141(17): 3303-3318.

Farge, E. (2003): *Mechanical Induction of Twist in the Drosophila Foregut/Stomodeal Primordium*, Current Biology 13(16): 1365-1377.

Farnebäck, G. (2003): *Two-Frame Motion Estimation Based on Polynomial Expansion*, Lecture Notes in Computer Science: 363-370.

Ferrell, J. E. (2013): *Feedback loops and reciprocal regulation: recurring motifs in the systems biology of the cell cycle*, Current Opinion in Cell Biology 25(6): 676-686.

Finkle, J. D., Wu, J. J., Bagheri, N. (2018): *Windowed Granger causal inference strategy improves discovery of gene regulatory networks*, Proc. Natl. Acad. Sci. USA 115(9): 2252-2257.

Fischer, R. S., Gardel, M., Ma, X., Adelstein, R. S., Waterman, C. M. (2009): *Local Cortical Tension by Myosin II Guides 3D Endothelial Cell Branching*, Current Biology 19(3): 260-265.

Friedl, P., Gilmour, D. (2009): *Collective cell migration in morphogenesis, regeneration and cancer*, Nature Reviews Molecular Cell Biology 10(7): 445-457.

Friston, K., Moran, R., Seth, A. K. (2013): *Analysing connectivity with Granger causality and dynamic causal modelling*, Current Opinion in Neurobiology 23(2): 172-178.

Gayathri Vegesna N. V., Ronchi P., Durdu S., Terjung S., Pepperkok R. (2017): *Targeted Ablation Using Laser Nanosurgery*, in: *Light Microscopy*, Methods in Molecular Biology 1563.

Ghabrial, A. S., Krasnow, M. A. (2006): *Social interactions among epithelial cells during tracheal branching morphogenesis*, Nature 441(7094): 746-749.

Ghysen, A., Dambly-Chaudière, C. (2004): *Development of the zebrafish lateral line*, Current Opinion in Neurobiology 14(1): 67-73.

Ghysen, A., Dambly-Chaudière, C. (2007): *The lateral line microcosmos*, Genes & Development 21(17): 2118-2130

Gierer, A., Meinhardt, H. (1972): *A Theory of Biological Pattern Formation*, Kybernetik 12: 30-39.

Giesen, C., Wang, H. A. O., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., Schüffler, P.J., Grolimund, D., Buhmann, J. M., Brandt, S., Varga, Z., Wild, P.J., Günther, D., Bodenmiller, B. (2014): *Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry*: Nature Methods 11(4): 417-422.

Gilmour, D., Knaut, H., Maischein, H.-M., Nüsslein-Volhard, C. (2004): *Towing of sensory axons by their migrating target cells in vivo*, Nature Neuroscience 7(5): 491-492.

Gilmour, D. T., Maischein, H.-M., Nüsslein-Volhard, C. (2002): *Migration and Function of a Glial Subtype in the Vertebrate Peripheral Nervous System*, Neuron 34(4): 577-588.

Gilmour, D., Rembold, M., Leptin, M. (2017): *From morphogen to morphogenesis and back*, Nature 541(7637): 311-320.

Gohlke, C. (2016): *tifffile version 0.11.1*, pypi.org/project/tifffile [online; accessed 2018-10-29].

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A., Tegnér, J. (2014): *Data integration in the era of omics: current and future challenges*, BMC Systems Biology 8(S2): I1.

Gompel, N., Cubedo, N., Thisse, C., Thisse, B., Dambly-Chaudière, C., Ghysen, A. (2001): *Pattern formation in the lateral line of zebrafish*, Mechanisms of Development 105(1-2): 69-77.

Gong, X., Mills, K. L. (2018): *Large-scale patterning of single cells and cell clusters in hydrogels*, Scientific Reports 8/1.

Gramacy, R. B., Lee, H. K. H., Macready, W. G. (2004): *Parameter space exploration with Gaussian process trees*, ICML '04.

Granger, C. W. J. (1969): *Investigating Causal Relations by Econometric Models and Cross-spectral Methods*, Econometrica 37(3): 424-438.

Grant, K. A., Raible, D. W., Piotrowski, T. (2005): *Regulation of Latent Sensory Hair Cell Precursors by Glia in the Zebrafish Lateral Line*, Neuron 45(1): 69-80.

Grewal, S. (2018): *An interview with James Sharpe*, Development 145.

Guglielmi, G., Barry, J. D., Huber, W., De Renzis, S. (2015): *An Optogenetic Method to Modulate Cell Contractility during Tissue Morphogenesis*, Developmental Cell 35(5): 646-660.

Guglielmi, G., Falk, H. J., De Renzis, S. (2016): *Optogenetic Control of Protein Function: From Intracellular Processes to Tissue Morphogenesis*, Trends in Cell Biology 26(11): 864-874.

Guntas, G., Hallett, R. A., Zimmerman, S. P., Williams, T., Yumerefendi, H., Bear, J. E., Kuhlman, B. (2014): *Engineering an improved light-induced dimer (iLID) for controlling the localization and activity of signaling proteins*, Proceedings of the National Academy of Sciences 112(1): 112-117.

Gut, G., Herrmann, M. D., Pelkmans, L. (2018): *Multiplexed protein maps link subcellular organization to cellular states*, Science 361(6401): eaar7042.

Haas, P., Gilmour, D. (2006): *Chemokine Signaling Mediates Self-Organizing Tissue Migration in the Zebrafish Lateral Line*, Developmental Cell 10(5): 673-680.

Hagberg, A. A., Schult, D. A., Swart, P. J. (2008): *Exploring network structure, dynamics, and function using NetworkX*, Proceedings of the 7th Python in Science Conference (SciPy2008): 11-15.

Haghverdi, L., Buettner, F., Theis, F. J. (2015): *Diffusion maps for high-dimensional single-cell analysis of differentiation data*, Bioinformatics 31(18): 2989-2998.

Han, Z., Li, X., Wu, J., Olson, E. N. (2004): *A Myocardin-Related Transcription Factor Regulates Activity of Serum Response Factor in Drosophila*, Proceedings of the National Academy of Sciences 101(34): 2567-12572.

Harding, M. J., Nechiporuk, A. V. (2012): *Fgfr-Ras-MAPK signaling is required for apical constriction via apical positioning of Rho-associated kinase during mechanosensory organ formation*, Development 139(18): 3467-3467.

Harris, A. R., Jreij, P., Fletcher, D. A. (2018): *Mechanotransduction by the Actin Cytoskeleton: Converting Mechanical Stimuli into Biochemical Signals*, Annual Review of Biophysics 47(1): 617-631.

Hawkings, S. W. (1993): *Black Holes and Baby Universes and other Essays*, Bantam Dell Publishing Group.

Hernández, P. P., Olivari, F. A., Sarrazin, A. F., Sandoval, P. C., Allende, M. L. (2007): *Regeneration in zebrafish lateral line neuromasts: Expression of the neural progenitor cell marker sox2 and proliferation-dependent and-independent mechanisms of hair cell renewal*, Developmental Neurobiology 67(5): 637-654.

Hellström, M., Phng, L.-K., Hofmann, J. J., Wallgard, E., Coultas, L., Lindblom, P., Alva, J., Nilsson, A.-K., Karlsson, L., Gaiano, N., Yoon, K., Rossant, J., Iruela-Arispe, L., Kale, M., Gerhardt, H., Hetsholtz, C. (2007): *Dll4 signalling through Notch1 regulates formation of tip cells during angiogenesis*, Nature 445(7129): 776-780.

Herrou, J., Crosson, S. (2011): *Function, structure and mechanism of bacterial photosensory LOV proteins*, Nature Reviews Microbiology 9(10): 713-723.

Ho, D. (2016): *Notepad++*, notepad-plus-plus.org [online; accessed 2018-10-29].

Hollins, B., Kuravi, S., Digby, G. J., Lambert, N. A. (2009): *The c-terminus of GRK3 indicates rapid dissociation of G protein heterotrimers*, Cellular Signalling 21(6): 1015-1021.

Holovatch, Y., Kenna, R., Thurner, S. (2017): *Complex systems: physics beyond physics*, Eur. J. Phys. 38.

Holzinger, A., Dehmer, M., Jurisica, I. (2014): *Knowledge Discovery and interactive Data Mining in Bioinformatics - State-of-the-Art, future challenges and research directions*, BMC Bioinformatics 15(S6): I1.

Huang, S., Chaudhary, K., Garmire, L. X. (2017): *More Is Better: Recent Progress in Multi-Omics Data Integration Methods*, Frontiers in Genetics 8.

Huff, J. (2016): *The Fast mode for ZEISS LSM 880 with Airyscan: high-speed confocal imaging with super-resolution and improved signal-to-noise ratio (Application Note)*: Nature Methods 13(11): i-ii.

Hughes, R. M., Bolger, S., Tapadia, H., Tucker, C. L. (2012): *Light-mediated control of DNA transcription in yeast*, Methods 58(4): 385-391.

Hunter, J. D. (2007): *Matplotlib: A 2D Graphics Environment*, Computing in Science & Engineering 9: 90-95.

Ilina, O., Friedl, P. (2009): *Mechanisms of collective cell migration at a glance*, Journal of Cell Science 122(18): 3203-3208.

Irvine, K. D., Wieschaus, E. (1994): *Cell intercalation during Drosophila germband extension and its regulation by pair-rule segmentation genes*, Development 120(4): 827-841.

Itoh, M., Chitnis, A. B. (2001): *Expression of proneural and neurogenic genes in the zebrafish lateral line primordium correlates with selection of hair cell fate in neuromasts*, Mechanisms of Development 102(1-2): 263-266.

Izquierdo, E., Quinkler, T., De Renzis, S. (2018): *Guided morphogenesis through optogenetic activation of Rho signalling during early Drosophila embryogenesis*, Nature Communications 9(1).

Jaeger, J., Manu, Reinitz, J. (2012): *Drosophila blastoderm patterning*, Current Opinion in Genetics & Development 22(6): 533-541.

Johnson, G. R., Donovan-Maiye, R. M., Maleckar, M. M. (2017): *Building a 3D Integrated Cell*, bioRxiv 10.1101/238378 [preprint].

Johnson, H. E., Goyal, Y., Pannucci, N. L., Schüpbach, T., Shvartsman, S. Y., Toettcher, J. E. (2017): *The Spatiotemporal Limits of Developmental Erk Signaling*, Developmental Cell 40(2): 185-192.

Johnson, H. E., Toettcher, J. E. (2018): *Illuminating developmental biology with cellular optogenetics*, Current Opinion in Biotechnology 52: 42-48.

Jones, E., Oliphant, E., Peterson, P., and the SciPy Development Team (2001): *SciPy: Open Source Scientific Tools for Python*, www.scipy.org [online; accessed 2018-10-29].

Jung, H.-S., Francis-West, P.H., Widelitz, R. B., Jiang, T.-J., Ting-Berreth, S., Tickle, C., Wolpert, L., Chuong, C.-M. (1998): *Local Inhibitory Action of BMPs and Their Relationships with Activators in Feather Formation: Implications for Periodic Patterning*, Developmental Biology 196: 11-23.

Käfer, J., Hayashi, T., Marée, A. F. M., Carthew, R. W., Graner, F. (2007): *Cell adhesion and cortex contractility determine cell patterning in the Drosophila retina*, Proceedings of the National Academy of Sciences 104(47): 18549-18554.

Karunarathne, W. K. A., Giri, L., Kalyanaraman, V., Gautam, N. (2013a): *Optically triggering spatiotemporally confined GPCR activity in a cell and programming neurite initiation and extension*, Proceedings of the National Academy of Sciences 110(17): E1565-E1574.

Karunarathne, W. K. A., Giri, L., Patel, A. K., Venkatesh, K. V., & Gautam, N. (2013b): *Optical control demonstrates switch-like PIP3 dynamics underlying the initiation of immune cell migration*, Proceedings of the National Academy of Sciences 110(17): E1575-E1583.

Karunarathne, W. K. A., O'Neill, P. R., Gautam, N. (2014): *Subcellular optogenetics - controlling signaling and single-cell behavior*, Journal of Cell Science 128(1): 15-25.

Keller, R., Davidson, L., Edlund, A., Elul, T., Ezin, M., Shook, D., Skoglund, P. (2000): *Mechanisms of convergence and extension by cell intercalation*, Philosophical Transactions of the Royal Society B: Biological Sciences 355(1399): 897-922.

Kennedy, M. J., Hughes, R. M., Peteya, L. A., Schwartz, J. W., Ehlers, M. D., Tucker, C. L. (2010): *Rapid blue-light–mediated induction of protein interactions in living cells*, Nature Methods 7(12): 973-975.

Képiró, M., Várkuti, B. H., Bodor, A., Hegyi, G., Drahos, L., Kovacs, M., Málnási-Csizmadia, A. (2012): *Azidoblebbistatin, a photoreactive myosin inhibitor*, Proceedings of the National Academy of Sciences 109(24): 9402-9407.

Képiró, M., Várkuti, B. H., Rauscher, A. A., Kellermayer, M. S. Z., Varga, M., Málnási-Csizmadia, A. (2015): *Molecular Tattoo: Subcellular Confinement of Drug Effects*, Chemistry & Biology 22(4): 548-558.

Kettleborough, R. N. W., Busch-Nentwich, E. M., Harvey, S. A., Dooley, C. M., de Bruijn, E., van Eeden, F., Sealy, I., White, R. J., Herd, C., Nijman, I. J., Fényes, F., Mehroke, S., Scahill, C., Gibbons, R., Wali, N., Carruthers, S., Hall, A., Yen, J., Cuppen, E., Stemple, D. L. (2013): *A systematic genome-wide analysis of zebrafish protein-coding gene function*, Nature 496(7446): 494-497.

Kim, J.-M., Hwa, J., Garriga, P., Reeves, P. J., RajBhandary, U. L., Khorana, H. G. (2005): *Light-Driven Activation of β2-Adrenergic Receptor Signaling by a Chimeric Rhodopsin Containing the β2-Adrenergic Receptor Cytoplasmic Loops*, Biochemistry 44(7): 2284-2292.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., and the Jupyter Development Team (2016): *Jupyter Notebooks – a publishing format for reproducible computational workflows*, in *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, 87-90.

Knaut, H., Werz, C., Geisler, R., Nüsslein-Volhard, C. (2002): *A zebrafish homologue of the chemokine receptor Cxcr4 is a germ-cell guidance receptor*, Nature 421(6920): 279-282.

Kondo, S., Miura, T. (2010): *Reaction-Diffusion Model as a Framework for Understanding Biological Pattern Formation*, Science 329: 1616-1620.

Konermann, S., Brigham, M. D., Trevino, A. E., Hsu, P. D., Heidenreich, M., Cong, L., Platt, R. J., Scott, D. A., Church, G. M., Zhang, F. (2013): *Optical control of mammalian endogenous transcription and epigenetic states*, Nature 500(7463): 472-476.

Konishi, S., Kitagawa, G. (1996): *Generalised information criteria in model selection*, Biometrika 83/4: 875-890.

Kopan, R., Ilagan, M. X. G. (2009): *The Canonical Notch Signaling Pathway: Unfolding the Activation Mechanism*, Cell 137(2): 216-233.

Krause, M., Gautreau, A. (2014): *Steering cell migration: lamellipodium dynamics and the regulation of directional persistence*, Nature Reviews Molecular Cell Biology 15(9): 577-590.

Krieg, M., Arboleda-Estudillo, Y., Puech, P.-H., Käfer, J., Graner, F., Müller, D. J., Heisenberg, C.-P. (2008): *Tensile forces govern germ-layer organization in zebrafish*, Nature Cell Biology 10(4): 429-436.

Kron, S. J., Spudich, J. A. (1986): *Fluorescent Actin Filaments Move on Myosin Fixed to a Glass Surface*, Proc. Natl. Acad. Sci. USA 83, 6272-6276.

Krueger, D., Tardivo, P., Nguyen, C., De Renzis, S. (2018): *Downregulation of basal myosin-II is required for cell shape changes and tissue invagination*, The EMBO Journal: e100170.

Kwan, K. M., Fujimoto, E., Grabher, C., Mangum, B. D., Hardy, M. E., Campbell, D. S., Parant, J. M., Yost, H. J., Kanki, J. P., Chien, C.-B. (2007): *The Tol2kit: A multisite gateway-based construction kit forTol2 transposon transgenesis constructs*, Developmental Dynamics 236(11): 3088-3099.

Landsberg, K. P., Farhadifar, R., Ranft, J., Umetsu, D., Widmann, T. J., Bittig, T., Said, A., Jülicher, F., Dahmann, C. (2009): *Increased Cell Bond Tension Governs Cell Sorting at the Drosophila Anteroposterior Compartment Boundary*, Current Biology 19(22): 1950-1955.

Le Clainche, C., Carlier, M.-F. (2008): *Regulation of Actin Assembly Associated With Protrusion and Adhesion in Cell Migration*, Physiological Reviews 88(2): 489-513.

Lecaudey, V., Cakan-Akdogan, G., Norton, W. H. J., Gilmour, D. (2008): *Dynamic Fgf signaling couples morphogenesis and migration in the zebrafish lateral line primordium*, Development 135(16): 2695-2705.

Lecaudey, V., Gilmour, D. (2006): *Organizing moving groups during morphogenesis*, Current Opinion in Cell Biology 18(1): 102-107.

Lecuit, T., Lenne, P.-F. (2007): *Cell surface mechanics and the control of cell shape, tissue patterns and morphogenesis*, Nature Reviews Molecular Cell Biology 8(8): 633-644.

Lee, K., Elliott, H. L., Oak, Y., Zee, C.-T., Groisman, A., Tytell, J. D., Danuser, G. (2015): *Functional Hierarchy of Redundant Actin Assembly Factors Revealed by Fine-Grained Registration of Intrinsic Image Fluctuations*, Cell Systems 1(1): 37-50.

Lee, K.-C., Goh, W. L. P., Xu, M., Kua, N., Lunny, D., Wong, J. S., Coomber, D., Vojtesek, B., Lane, E. B., Lane, D. P. (2007): *Detection of the p53 response in zebrafish embryos using new monoclonal antibodies*, Oncogene 27(5): 629-640.

Lele, Z., Krone, P. H. (1996): *The zebrafish as a model system in developmental, toxicological and transgenic research*, Biotechnology Advances 14(1): 57-72.

Letunic, I., Bork, P. (2017): *20 years of the SMART protein domain annotation resource*. Nucleic Acids Research 46(D1): D493-D496.

Levine, M., Davidson, E. H. (2005): *Gene regulatory networks for development*, Proceedings of the National Academy of Sciences 102(14): 4936-4942.

Levskaya, A., Weiner, O. D., Lim, W. A., Voigt, C. A. (2009): *Spatiotemporal control of cell signalling using a light-switchable protein interaction*, Nature 461(7266): 997-1001.

Lewis-Kraus, G. (2016): *The Greate A.I. Awakening*, The New York Times Magazine, www.nytimes. com/2016/12/14/magazine/the-great-ai-awakening.html [online, accessed: 2018-11-26].

Libbrecht, M. W., Noble, W. S. (2015): *Machine learning applications in genetics and genomics*, Nature Reviews Genetics 16(6): 321-332.

Lin, J.-R., Fallahi-Sichani, M., Sorger, P. K. (2015): *Highly multiplexed imaging of single cells using a high-throughput cyclic immunofluorescence method*, Nature Communications 6(1).

Lipton, Z. C. (2017): *The Mythos of Model Interpretability*, arXiv: 1606.03490v3 [preprint].

Liu, A. P., Fletcher, D. A. (2009): *Biology under construction: in vitro reconstitution of cellular function*, Nature Reviews Molecular Cell Biology 10: 644-650.

Liu, H., Gomez, G., Lin, S., Lin, S., Lin, C. (2012): *Optogenetic Control of Transcription in Zebrafish*, PLoS ONE 7(11): e50738.

Liu, H., Yu, X., Li, K., Klejnot, J., Yang, H., Lisiero, D., Lin, C. (2008): *Photoexcited CRY2 interacts with CIB1 to regulate transcription and floral initiation in Arabidopsis*, Science 322(5907): 1535-1539.

Lo, C.-M., Wang, H.-B., Dembo, M., Wang, Y. (2000): *Cell Movement Is Guided by the Rigidity of the Substrate*, Biophysical Journal 79(1): 144-152.

Lock, J. G., Mamaghani, M. J., Shafqat-Abbasi, H., Gong, X., Tyrcha, J., Strömblad, S. (2014): *Plasticity in the Macromolecular-Scale Causal Networks of Cell Migration*, PLoS ONE 9(2): e90593.

Lucitti, J. L., Jones, E. A. V., Huang, C., Chen, J., Fraser, S. E., Dickinson, M. E. (2007): *Vascular remodeling of the mouse yolk sac requires hemodynamic force*, Development 134(18): 3317-3326.

Lundberg, S. M., Lee, S.-I. (2017): *A Unified Approach to Interpreting Model Predictions*, arXiv: 1705.07874v2 [preprint].

Ma, E. Y., Raible, D. W. (2009): *Signaling Pathways Regulating Zebrafish Lateral Line Development*, Current Biology 19(9): R381-R386.

MacQueen, J. (1967): *Some methods for classification and analysis of multivariate observations*, Proc. Fifth Berkeley Symp. on Math. Statist. and Prob. 1: 281-297.

Maître, J.-L., Berthoumieux, H., Krens, S. F. G., Salbreux, G., Julicher, F., Paluch, E., Heisenberg, C.-P. (2012): *Adhesion Functions in Cell Sorting by Mechanically Coupling the Cortices of Adhering Cells*, Science 338(6104): 253-256.

Maître, J.-L., Turlier, H., Illukkumbura, R., Eismann, B., Niwayama, R., Nédélec, F., Hiiragi, T. (2016): *Asymmetric division of contractile domains couples cell positioning and fate specification*, Nature 536(7616): 344-348.

Malet-Engra, G., Yu, W., Oldani, A., Rey-Barroso, J., Gov, N. S., Scita, G., Dupré, L. (2015): *Collective Cell Motility Promotes Chemotactic Prowess and Resistance to Chemorepulsion*, Current Biology 25(2): 242-250.

Mammoto, T., Ingber, D. E. (2010): *Mechanical control of tissue and organ development*, Development 137(9): 1407-1420.

Manson, S. M. (2001): *Simplifying complexity: a review of complexity theory*, Geoforum 32: 405-414.

Mansoorabadi, S., Thibodeaux, C. J., Liu, H. (2007): *The Diverse Roles of Flavin Coenzymes — Nature's Most Versatile Thespians*, ChemInform 38(48).

Marcon, L., Diego, X., Sharpe, J., Müller, P. (2016): *High-throughput mathematical analysis identifies Turing networks for patterning with equally diffusing signals*, eLife 2016/5: e14022.

Marcus, G. (2018): *Deep Learning: A Critical Appraisal*, arXiv: 1801.00631 [preprint].

Martin, A. C., Goldstein, B. (2014): *Apical constriction: themes and variations on a cellular mechanism driving morphogenesis*, Development 141(10): 1987-1998.

Matsuda, M., Chitnis, A. B. (2010): *Atoh1a expression must be restricted by Notch signaling for effective morphogenesis of the posterior lateral line primordium in zebrafish*, Development 137(20): 3477-3487.

Matzke, E. B. (1946): *The three-dimensional shape of bubbles in foam — an analysis of the rôle of surface forces in three-dimensional cell shape deermination*, American Journal of Botany 33: 58-80.

Mazzocchi, F. (2008): *Complexity in Biology*, EMBO reports 9: 10-14.

McBeath, R., Pirone, D. M., Nelson, C. M., Bhadriraju, K., Chen, C. S. (2004): *Cell Shape, Cytoskeletal Tension, and RhoA Regulate Stem Cell Lineage Commitment*, Developmental Cell 6(4): 483-495.

McCauley, H. A., Wells, J. M. (2017): *Pluripotent stem cell-derived organoids: using principles of developmental biology to grow human tissues in a dish*, Development 144: 958-962.

McKay, M. M., Morrison, D. K. (2007): *Integrating signals from RTKs to ERK/MAPK*, Oncogene 26(22): 3113-3121.

McKinney, W. (2010): *Data Structures for Statistical Computing in Python*, Proceedings of the 9th Python in Science Conference, 51-56.

Meinhardt, H. (1993): *A Model for Pattern Formation of Hypostome, Tentacles, and Foot in Hydra: How to Form Structures Close to Each Other, How to Form Them at a Distance*, Developmental Biology 157(2): 321-333.

Meinhardt, H. (1998): *The Algorithmic Beauty of Sea Shells*, Springer.

Meinhardt, H., Gierer, A. (2000): *Pattern formation by local self-activation and lateral inhibition*, BioEssays 22: 753-760.

Meinhardt, H. (2006): *Gierer-Meinhardt model*, Scholarpedia 1(12): 1418.

Metcalfe, W. K. (1985): *Sensory neuron growth cones comigrate with posterior lateral line primordial cells in zebrafish*, The Journal of Comparative Neurology 238(2): 218-224.

Michelson, A. A. (1903): *Light waves and their uses*, The Decennial Publications: Second Series, Volume III, The University of Chicago Press.

Montell, D. J. (2003): *Border-cell migration: the race is on*, Nature Reviews Molecular Cell Biology 4(1): 13-24.

Motta-Mena, L. B., Reade, A., Mallory, M. J., Glantz, S., Weiner, O. D., Lynch, K. W., Gardner, K. H. (2014): *An optogenetic gene expression system with rapid activation and deactivation kinetics*, Nature Chemical Biology 10(3): 196-202.

Munjal, A., Lecuit, T. (2014): *Actomyosin networks and tissue morphogenesis*, Development 141(9): 1789-1793.

Nabeshima, K., Inoue, T., Shimao, Y., Okada, Y., Itoh, Y., Seiki, M., & Koono, M. (2000): *Front-cell-specific expression of membrane-type 1 matrix metalloproteinase and gelatinase A during cohort migration of colon carcinoma cells induced by hepatocyte growth factor/scatter factor*, Cancer research 60(13): 3364-3369.

Newman, S. A., Tilmann, G., Bhat, R. (2018): *The Vertebrate Limb: An Evolving Complex of Self-Organizing Systems*, Progress in Biophysics and Molecular Biology [accepted manuscript].

Niemz, M. H. (2013): *Laser-tissue interactions: fundamentals and applications*, Springer Science & Business Media.

Niessen, C. M., Leckband, D., Yap, A. S. (2011): *Tissue Organization by Cadherin Adhesion Molecules: Dynamic Molecular and Cellular Mechanisms of Morphogenetic Regulation*, Physiological Reviews, 91(2): 691-731.

Nüsslein-Volhard, C., Wieschaus, E. (1980): *Mutations affecting segment number and polarity in Drosophila*, Nature 287: 795-801.

Oates, A. C., Morelli, L. G., Ares, S. (2012): *Patterning embryos with oscillations: structure, function and dynamics of the vertebrate segmentation clock*, Development 139(4): 625-639.

Occhetta, P., Centola, M., Tonnarelli, B., Redaelli, A., Martin, I., Rasponi, M. (2015): *High-Throughput Microfluidic Platform for 3 D Cultures of Mesenchymal Stem Cells, Towards Engineering Developmental Processes*, Scientific Reports 5/1.

O'Neill, P. R., Gautam, N. (2014): *Subcellular optogenetic inhibition of G proteins generates signaling gradients and cell migration*, Molecular Biology of the Cell 25(15): 2305-2314.

Parloff, R. (2016): *Why deep learning is suddenly changing your life*, Fortune, fortune.com/ai-artificial-intelligence-deep-machine-learning [online, accessed: 2018-11-26].

Pastor-Pareja, J. C., Grawe, F., Martín-Blanco, E., García-Bellido, A. (2004): *Invasive Cell Behavior during Drosophila Imaginal Disc Eversion Is Mediated by the JNK Signaling Cascade*, Developmental Cell 7(3): 387-399.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É. (2011): *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research 12: 2825-2830.

Pegtel, D. M., Ellenbroek, S. I. J., Mertens, A. E. E., van der Kammen, R. A., de Rooij, J., Collard, J. G. (2007): *The Par-Tiam1 Complex Controls Persistent Migration by Stabilizing Microtubule-Dependent Front-Rear Polarity*, Current Biology 17(19): 1623-1634.

Persson, P. B., Wagner, C.D. (1996): *General principles of chaotic dynamics*, cardiovascular Research 31: 332-341.

Phng, L.-K., Gerhardt, H. (2009): *Angiogenesis: A Team Effort Coordinated by Notch*, Developmental Cell 16(2): 196-208.

Pincus, Z., Theriot, J. A. (2007): *Comparison of quantitative methods for cell-shape analysis*, Journal of Microscopy 227(2): 140-156.

Politi, A. Z., Cai, Y., Walther, N., Hossain, M. J., Koch, B., Wachsmuth, M., Ellenberg, J. (2018): *Quantitative mapping of fluorescently tagged cellular proteins using FCS-calibrated four-dimensional imaging*, Nature Protocols 13(6): 1445-1464.

Preibisch, S., Saalfeld, S., Tomancak, P. (2009): *Globally optimal stitching of tiled 3D microscopic image acquisitions*, Bioinformatics 25(11): 1463-1465.

Purnick, E. M., Weiss, R. (2009): *The second wave of synthetic biology: from modules to systems*, Nature Reviews Molecular Cell Biology 10: 410-422.

Qiu, P., Simonds, E. F., Bendall, S. C., Gibbs, K. D., Bruggner, R. V., Linderman, M. D., Sachs, K., Nolan, G. P., Plevritis, S. K. (2011): *Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE*, Nature Biotechnology 29(10): 886-891.

Rabut, G., Ellenberg, J. (2004): *Automatic real-time three-dimensional cell tracking by fluorescence microscopy*, Journal of Microscopy 216(2): 131-137.

Radovic, A., Williams, M., Rousseau, D., Kagan, M., Bonacorsi, D., Himmel, A., Aurisano, A., Kazuhiro, T., Wongjirad, T. (2018): *Machine learning at the energy and intensity frontiers of particle physics*, Nature 560(7716): 41-48.

Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A., Tyagi, S. (2008): *Imaging individual mRNA molecules using multiple singly labeled probes*, Nature Methods 5(10): 877-879.

Raj, A., van Oudenaarden, A. (2008): *Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences*, Cell 135(2): 216-226.

Rauzi, M., Krzic, U., Saunders, T. E., Krajnc, M., Ziherl, P., Hufnagel, L., Leptin, M. (2015): *Embryo-scale tissue mechanics during Drosophila gastrulation movements*, Nature Communications 6(1).

Ray, J., Johnny, O., Trovati, M., Sotiriadis, S., Bessis, N. (2018): *The Rise of Big Data Science: A Survey of Techniques, Methods and Approaches in the Field of Natural Language Processing and Network Theory*, Big Data and Cognitive Computing 2(3): 22.

Raybaut, P., Cordoba, C., and the Spyder contributors (2018): *Spyder-IDE GitHub Repository*, github.com/spyder-ide/spyder [online, accessed: 2018-10-29].

Redchuk, T. A., Omelina, E. S., Chernov, K. G., Verkhusha, V. V. (2017): *Near-infrared optogenetic pair for protein regulation and spectral multiplexing*, Nature Chemical Biology 13(6): 633-639.

Reiter, E., Lefkowitz, R. J. (2006): *GRKs and β-arrestins: roles in receptor silencing, trafficking and signaling*, Trends in Endocrinology & Metabolism 17(4): 159-165.

Repina, N. A., Rosenbloom, A., Mukherjee, A., Schaffer, D. V., Kane, R. S. (2017): *At Light Speed: Advances in Optogenetic Systems for Regulating Cell Signaling and Behavior*, Annual Review of Chemical and Biomolecular Engineering 8(1): 13-39.

Revenu, C., Streichan, S., Dona, E., Lecaudey, V., Hufnagel, L., Gilmour, D. (2014): *Quantitative cell polarity imaging defines leader-to-follower transitions during collective migration and the key role of microtubule-dependent adherens junction formation*, Development 141(6): 1282-1291.

Rid, R., Schiefermeier, N., Grigoriev, I., Small, J. V., Kaverina, I. (2005): *The last but not the least: The origin and significance of trailing adhesions in fibroblastic cells*, Cell Motility and the Cytoskeleton 61(3): 161-171.

Ridley, A. J. (2011): *Life at the Leading Edge*, Cell 145(7): 1012-1022.

Riedl, J., Crevenna, A. H., Kessenbrock, K., Yu, J. H., Neukirchen, D., Bista, M., Bradke, F., Jenne, D., Holak, T. A., Werb, Z., Sixt, M., Wedlich-Soldner, R. (2008): *Lifeact: a versatile marker to visualize F-actin*, Nature Methods 5(7): 605-607.

Rogers, K. W., Schier, A. F. (2011): *Morphogen Gradients: From Generation to Interpretation*, Annual Review of Cell and Developmental Biology 27(1): 377-407.

Röper, K. (2013): *Supracellular actomyosin assemblies during development*, BioArchitecture 3(2): 45-49.

Rørth, P. (2012): *Fellow travellers: emergent properties of collective cell migration*, EMBO Reports 13(11): 984-991.

Ross, J., Arkin, A. P. (2009): *Complex systems: From chemistry to systems biology*, PNAS 106/16: 6433-6434.

Rottner, K., Hall, A., Small, J. V. (1999): *Interplay between Rac and Rho in the control of substrate contact dynamics*, Current Biology 9(12): 640-648.

Rozario, T., DeSimone, D. W. (2010): *The extracellular matrix in development and morphogenesis: A dynamic view*, Developmental Biology 341(1): 126-140.

Saini, D. K., Kalyanaraman, V., Chisari, M., Gautam, N. (2007): *A Family of G Protein βγ Subunits Translocate Reversibly from the Plasma Membrane to Endomembranes on Receptor Activation*, Journal of Biological Chemistry 282(33): 24099-24108.

Salakhutdinov, R. (2015): *Learning Deep Generative Models*, Annual Review of Statistics and Its Application 2(1): 361-385.

Salbreux, G., Charras, G., Paluch, E. (2012): *Actin cortex mechanics and cellular morphogenesis*, Trends in Cell Biology 22(10): 536-545.

Samek, W., Wiegand, T., Müller K.-R. (2017): *Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models*, arXiv: 1708.08296v1 [preprint].

Sanghvi-Shah, R., Weber, G. F. (2017): *Intermediate Filaments at the Junction of Mechanotransduction, Migration, and Development*, Frontiers in Cell and Developmental Biology 5.

Sarrazin, A. F., Villablanca, E. J., Nuñez, V. A., Sandoval, P. C., Ghysen, A., Allende, M. L. (2006): *Proneural gene requirement for hair cell differentiation in the zebrafish lateral line*, Developmental Biology 295(2): 534-545.

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., Regev, A. (2015): *Spatial reconstruction of single-cell gene expression data*, Nature Biotechnology 33(5): 495-502.

Scarcelli, G., Yun, S. H. (2007): *Confocal Brillouin microscopy for three-dimensional mechanical imaging*, Nature Photonics 2(1): 39-43.

Scarpa, E., Szabó, A., Bibonne, A., Theveneau, E., Parsons, M., Mayor, R. (2015): *Cadherin Switch during EMT in Neural Crest Cells Leads to Contact Inhibition of Locomotion via Repolarization of Forces*, Developmental Cell 34(4): 421-434.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.-Y., White, D. J., Hartenstein, V., Eliceiri, K., Tomancak, P., Cardona, A. (2012): *Fiji: an open-source platform for biological-image analysis*, Nature Methods 9(7): 676-682.

Schlosser, G. (2006): *Induction and specification of cranial placodes*, Developmental Biology 294(2): 303-351.

Schneider, C. A., Rasband, W. S., Eliceiri, K. W. (2012): *NIH Image to ImageJ: 25 years of image analysis*, Nature Methods 9(7): 671-675.

Schwartz, M. (2004): *Rho signalling at a glance*, Journal of Cell Science 117(23): 5457-5458.

Scott, D. W. (1992): *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley & Sons.

Segler, M. H. S., Preuss, M., Waller, M. P. (2018): *Planning chemical syntheses with deep neural networks and symbolic AI*, Nature 555(7698): 604-610.

Shaner, N. C., Lin, M. Z., McKeown, M. R., Steinbach, P. A., Hazelwood, K. L., Davidson, M. W., Tsien, R. Y. (2008): *Improving the photostability of bright monomeric orange and red fluorescent proteins*, Nature Methods 5(6): 545-551.

Shaya, O., Binshtok, U., Hersch, M., Rivkin, D., Weinreb, S., Amir-Zilberstein, L., Khamaisi, B., Oppenheim, O., Desai, R. A., Goodyear, R. J., Richardson, G. P., Chen, C. S., Sprinzak, D. (2017): *Cell-Cell Contact Area Affects Notch Signaling and Notch-Dependent Patterning*, Developmental Cell 40(5): 505-511.

Shih, J., Keller, R. (1992): *Cell motility driving mediolateral intercalation in explants of Xenopus laevis*, Development 116(4): 901-914.

Shyer, A. E., Huycke, T. R., Lee, C., Mahadevan, L., Tabin, C. J. (2015): *Bending Gradients: How the Intestinal Stem Cell Gets Its Home*, Cell 161(3): 569-580.

Singhvi, R., Kumar, A., Lopez, G. P., Stephanopoulos, G. N., Wang, D. I. C., Whitesides, G. M., Ingber, D. E. (1994): *Engineering Cell Shape and Function*, Science 264: 696-699.

Sjöqvist, M., Andersson, E. R. (2017): *Do as I say, Not(ch) as I do: Lateral control of cell fate*, Developmental Biology.

Smutny M., Behrndt M., Campinho P., Ruprecht V., Heisenberg CP. (2015): *UV Laser Ablation to Measure Cell and Tissue-Generated Forces in the Zebrafish Embryo In Vivo and Ex Vivo*. in: *Tissue Morphogenesis*, Methods in Molecular Biology 1189.

Solon, J., Kaya-Çopur, A., Colombelli, J., Brunner, D. (2009): *Pulsed Forces Timed by a Ratchet-like Mechanism Drive Directed Tissue Movement during Dorsal Closure*, Cell 137(7): 1331-1342.

Somogyi, K., Rørth, P. (2004): *Evidence for Tension-Based Regulation of Drosophila MAL and SRF during Invasive Cell Migration*, Developmental Cell 7(1): 85-93.

Stent, G. S. (1969): *The coming of the Golden Age; a view of the end of progress*, The Natural History Press.

Sternberg, P. W., Horvitz, H. R. (1989): *The combined action of two intercellular signaling pathways specifies three cell fates during vulval induction in C. elegans*, Cell 58(4): 679-693.

Steinberg, M. S. (1970): *Does differential adhesion govern self-assembly processes in histogenesis? Equilibrium configurations and the emergence of a hierarchy among populations of embryonic cells*, Journal of Experimental Zoology 173(4): 395-433.

Steinberg, M. S. (2007): *Differential adhesion in morphogenesis: a modern view*, Current Opinion in Genetics & Development 17(4): 281-286.

Stephenson, R. E., Miller, A. L. (2017): *Tools for live imaging of active Rho GTPases in Xenopus*, Genesis 55(1-2): e22998.

Stewart, A. Huang, J., Fisher, R. A. (2012): *RGS proteins in heart: brakes on the vagus*, Frontiers in Physiology 3.

Sullivan, G. M., Feinn, R. (2012): *Using Effect Size—or Why thePValue Is Not Enough*, Journal of Graduate Medical Education 4(3): 279-282.

Swat, M. H., Thomas, G. L., Belmonte, J. M., Shirinifard, A., Hmeljak, D., Glazier, J. A. (2012): *Multi-Scale Modeling of Tissues Using CompuCell3D*, Computational Methods in Cell Biology 110: 325-366.

Sweeton, D., Parks, S., Costa, M., Wieschaus, E. (1991): *Gastrulation in Drosophila: the formation of the ventral furrow and posterior midgut invaginations*, Development 112(3): 775-789.

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N. T., Roth, A., Bork, P., Jensen, L. J., von Mering, C. (2016): *The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible*, Nucleic Acids Research 45(D1): D362-D368.

Tada, M., Heisenberg, C.-P. (2012): *Convergent extension: using collective cell migration and cell intercalation to shape embryos*, Development 139(21): 3897-3904.

Tambe, D. T., Corey Hardin, C., Angelini, T. E., Rajendran, K., Park, C. Y., Serra-Picamal, X., Zhou, E. H., Zaman, M. H., Butler, J. P., Weitz, D. A., Fredberg, J. J., Trepat, X. (2011): *Collective cell guidance by cooperative intercellular forces*, Nature Materials 10(6): 469-475.

Teague, B. P., Guye, P., Weiss, R. (2016): *Synthetic Morphogenesis*, Cold Spring Harbor Perspectives in Biology, 8/9: a023929.

Tepass, U., Godt, D., Winklbauer, R. (2002): *Cell sorting in animal development: signalling and adhesive mechanisms in the formation of tissue boundaries*, Current Opinion in Genetics & Development 12(5): 572-582.

Teruel, M. N., Blanpied, T. A., Shen, K., Augustine, G. J., Meyer, T. (1999): *A versatile microporation technique for the transfection of cultured CNS neurons*, Journal of Neuroscience Methods 93(1): 37-48.

Theveneau, E., Marchant, L., Kuriyama, S., Gull, M., Moepps, B., Parsons, M., Mayor, R. (2010): *Collective Chemotaxis Requires Contact-Dependent Cell Polarity*, Developmental Cell 19(1): 39-53.

Tischer, C., Hilsenstein, V., Hanson, K., Pepperkok, R. (2014): *Adaptive fluorescence microscopy by online feedback image analysis*, Quantitative Imaging in Cell Biology 489-503.

Tischer, D., Weiner, O. D. (2014): *Illuminating cell signalling with optogenetic tools*, Nature Reviews Molecular Cell Biology 15(8): 551-558.

Toni, T., Stumpf, M. P. H. (2010): *Simulation-based model selection for dynamical systems in systems and population biology*, Bioinformatics 26/1: 104-110.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., Stumpf, M. P. H. (2009): *Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems*, J. R. Soc. Interface 6: 187-202.

Torvalds, L., and the Git contributors (2018): *Git GitHub Repository*, github.com/git/git [online; access 2018-10-29].

Townes, P. L., Holtfreter, J. (1955): *Directed movements and selective adhesion of embryonic amphibian cells*, Journal of Experimental Zoology 128(1): 53-120.

Travis, E., Oliphant, E. (2006): *A guide to NumPy*, Trelgol Publishing.

Trepat, X., Wasserman, M. R., Angelini, T. E., Millet, E., Weitz, D. A., Butler, J. P., Fredberg, J. J. (2009): *Physical forces during collective cell migration*, Nature Physics 5(6): 426-430.

Turing, A. M. (1952): *The Chemical Basis of Morphogenesis*, Philosophical Transactions of the Royal Society B 237: 37-72.

Valentin, G., Haas, P., Gilmour, D. (2007): *The Chemokine SDF1a Coordinates Tissue Migration through the Spatially Restricted Activation of Cxcr7 and Cxcr4b*, Current Biology 17(12): 1026-1031.

Valon, L., Marín-Llauradó, A., Wyatt, T., Charras, G., Trepat, X. (2017): *Optogenetic control of cellular forces and mechanotransduction*, Nature Communications 8: 14396.

Van Aelst, L., D'Souza-Schorey, C. (1997): *Rho GTPases and signaling networks*, Genes & Development 11(18): 2295-2322.

Van Bergeijk, P., Adrian, M., Hoogenraad, C. C., Kapitein, L. C. (2015): *Optogenetic control of organelle transport and positioning*, Nature 518(7537): 111-114.

Van der Maaten, L. J. P., Hinton, G. E. (2008): *Visualizing High-Dimensional Data Using t-SNE*, Journal of Machine Learning Research 9: 2579-2605.

Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., Yu, T., and the scikit-image contributors (2014): *scikit-image: Image processing in Python*, PeerJ 2:e453.

Van Dyke, M. (1982): *An Album on Fluid Motion*, THE PARABOLIC PRESS.

Van Rossum, G. (1995): *Python Tutorial*, CWI Report: CS-R9526.

Van Unen, J., Reinhard, N. R., Yin, T., Wu, Y. I., Postma, M., Gadella, T. W. J., Goedhart, J. (2015): *Plasma membrane restricted RhoGEF activity is sufficient for RhoA-mediated actin polymerization*, Scientific Reports 5(1).

Varennes, J., Han, B., Mugler, A. (2016): *Collective Chemotaxis through Noisy Multicellular Gradient Sensing*, Biophysical Journal 111(3): 640-649.

Vergara, H. M., Bertucci, P. Y., Hantz, P., Tosches, M. A., Achim, K., Vopalensky, P., Arendt, D. (2017): *Whole-organism cellular gene-expression atlas reveals conserved cell types in the ventral nerve cord of Platynereis dumerilii*, Proc. Natl. Acad. Sci. USA 114(23): 5878-5885.

Vilela, M., Danuser, G. (2011): *What's wrong with correlative experiments?*, Nature Cell Biology 13(9): 1011-1011.

Villablanca, E. J., Renucci, A., Sapède, D., Lec, V., Soubiran, F., Sandoval, P. C., Dambly-Chaudière, C., Ghysen, A., Allende, M. L. (2006): *Control of cell migration in the zebrafish lateral line: Implication of the gene "Tumour-Associated Calcium Signal Transducer,"tacstd*, Developmental Dynamics 235(6): 1578-1588.

Villasenor, A., Chong, D. C., Henkemeyer, M., Cleaver, O. (2010): *Epithelial dynamics of pancreatic branching morphogenesis*, Development 137(24): 4295-4305.

Vitorino, P., Meyer, T. (2008): *Modular control of endothelial sheet migration*, Genes & Development 22(23): 3268-3281.

Wang, X., Ha, T. (2013): *Defining Single Molecular Forces Required to Activate Integrin and Notch Signaling*, Science 340(6135): 991-994.

Wang, X., He, L., Wu, Y. I., Hahn, K. M., Montell, D. J. (2010): *Light-mediated activation reveals a key role for Rac in collective guidance of cell movement in vivo*, Nature Cell Biology 12(6): 591-597.

Wartlick, O., Kicheva, A., González-Gaitán, M. (2009): *Morphogen gradient formation*, Cold Spring Harbor perspectives in biology: a001255.

Waskom, M., Botvinnik, O., drewokane, Hobson, P., David, Halchenko, Y., Lukauskas, S., Cole, J. B., Warmenhoven, J., de Ruiter, J., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Martin, M., Miles, A., Meyer, K., Augspurger, T., Yarkoni, T., Bachant, P., Williams, M., Evans, C., Fitzgerald, C., Brian, Wehner, D., Hitz, G., Ziegler, E., Qalieh, A, Lee, A. (2016): *seaborn: v0.7.1*, zenodo.

Way, G. P., Greene, C. S. (2017): *Extracting a Biologically Relevant Latent Space from Cancer Transcriptomes with Variational Autoencoders*, bioRxiv: 10.1101/174474 [preprint].

Weber, G. F., Bjerke, M. A., DeSimone, D. W. (2012): *A Mechanoresponsive Cadherin-Keratin Complex Directs Polarized Protrusive Behavior and Collective Cell Migration*, Developmental Cell 22(1): 104-115.

Weber, M., Hauschild, R., Schwarz, J., Moussion, C., de Vries, I., Legler, D. F., Luther, S. A., Bollenbach, T., Sixt, M. (2013): *Interstitial Dendritic Cell Guidance by Haptotactic Chemokine Gradients*, Science 339(6117): 328-332.

Weight, C., Parnham, D., Waites, R. (2008): *LeafAnalyser: a computational method for rapid and large-scale analyses of leaf shape variation*, The Plant Journal 53(3): 578-586.

Welf, E. S., Danuser, G. (2014): *Using Fluctuation Analysis to Establish Causal Relations between Cellular Events without Experimental Perturbation*, Biophysical Journal 107: 2492-2498.

Westerfield, M. (2000): *The zebrafish book. A guide for the laboratory use of zebrafish (Danio rerio)*, 4th ed., University of Oregon Press.

Williams, M. J., Habayeb, M. S., Hultmark, D. (2007): *Reciprocal regulation of Rac1 and Rho1 in Drosophila circulating immune surveillance cells*, Journal of Cell Science 120(3): 502-511.

Woese, C. R. (2004): *A New Biology for a New Century*, Microbiology and Molecular Biology Reviews 68(2): 173-186.

Wu, A. R., Neff, N. F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M. E., Mburu, F. M., Mantalas, G. L., Sim, S., Clarke, M. F., Quake, S. R. (2013): *Quantitative assessment of single-cell RNA-sequencing methods*, Nature Methods 11(1): 41-46.

Wu, Y. I., Frey, D., Lungu, O. I., Jaehrig, A., Schlichting, I., Kuhlman, B., Hahn, K. M. (2009): *A genetically encoded photoactivatable Rac controls the motility of living cells*, Nature 461(7260): 104-108.

Xu, Y., Hyun, Y.-M., Lim, K., Lee, H., Cummings, R. J., Gerber, S. A., Bae, S., Cho, T. Y., Lord, E. M., Kim, M. (2014): *Optogenetic control of chemokine receptor signal and T-cell migration*, Proceedings of the National Academy of Sciences 111(17): 6371-6376.

Yamaguchi, M., Yoshimoto, E., Kondo, S. (2007): *Pattern regulation in the stripe of zebrafish suggests an underlying dynamic and autonomous mechanism*, Proceedings of the National Academy of Sciences 104(12): 4790-4793.

Yan, D., Lin, X. (2009): *Shaping Morphogen Gradients by Proteoglycans*, Cold Spring Harbor Perspectives in Biology 1(3): a002493-a002493.

Yoo, S. K., Deng, Q., Cavnar, P. J., Wu, Y. I., Hahn, K. M., Huttenlocher, A. (2010): *Differential Regulation of Protrusion and Polarity by PI(3)K during Neutrophil Motility in Live Zebrafish*, Developmental Cell 18(2): 226-236.

Zaidel-Bar, R., Zhenhuan, G., Luxenburg, C. (2015): *The contractome - a systems view of actomyosin contractility in non-muscle cells*, Journal of Cell Science 128(12): 2209-2217.

Zaritsky, A., Welf, E. S., Tseng, Y.-Y., Angeles Rabadán, M., Serra-Picamal, X., Trepat, X., Danuser, G. (2015): *Seeds of Locally Aligned Motion and Stress Coordinate a Collective Cell Migration*, Biophysical Journal 109(12): 2492-2500.

Zayed, J. M., Nouvel, N., Rauwald, U., Scherman, O. A. (2009): *Chemical complexity—supramolecular self-assembly of synthetic and biological building blocks in water*, Chem. Soc. Rev. 39: 2806-2816.

Zhabotinsky, A. M. (1964): *Periodical oxidation of malonic acid in solution (a study of the Belousov reaction kinetics)*, Biofizika 9: 306-311.

Zhabotinsky, A. M., Zaikin, A. N. (1971): *Spatial effects in a self-oscillating chemical system*, in *Oscillatory processes in biological and chemical systems II*, Sci. Publ.

Zhabotinsky, A. M., Zaikin, A. N. (1973): *Autowave Processes in a Distributed Chemical System*, J. theor. Biol. 40: 45-61.

Zhabotinsky, A. M. (1991): *A history of chemical oscillations and waves*, Chaos 1: 379-386.

Zhabotinsky, A. M. (2007): *Belousov-Zhabotinsky reaction*, Scholarpedia 2(9): 1435.

Zhang, K., Cui, B. (2015): *Optogenetic control of intracellular signaling pathways*, Trends in Biotechnology 33(2): 92-100.

Zhang, K., Duan, L., Ong, Q., Lin, Z., Varman, P. M., Sung, K., Cui, B. (2014): *Light-Mediated Kinetic Control Reveals the Temporal Effect of the Raf/MEK/ERK Pathway in PC12 Cell Neurite Outgrowth*, PLoS ONE 9(3): e92917.

Acknowledgements

None of the work described in this thesis and – more importantly – none of what I've learned and none of the fun I've had throughout my PhD adventure would have been possible without countless contributions, tiny and large, from a wide array of people. My heartfelt thanks go out to all of you, whether you've become a close friend or whether you just barely tolerated my ramblings for a few minutes during a seminar.

First and foremost, many thanks Darren Gilmour for having me in his lab and continuing to put up with the constant barrage of ideas, opinions and criticisms I've subjected him to. It's been nothing if not interesting to work together for the past four years.

Thanks to Francesca Peri and Stefano De Renzis for kindly hosting me when the Gilmour lab started to osmose from EMBL to the University of Zürich.

My utmost thanks to all the current and former members of the Gilmour lab; you have been at the very core of my PhD experience. Thanks Sevi Durdu, Erika Donà, Lionel Newton, Sarfaraz Farooqui, Mie Wong, Florian de Molière, and most importantly Elisa Gallo, with whom I got to have so many inspiring discussions over microwave dinners and with whom I got to work on so many fun and mostly harmless side projects. Special thanks to Sruthi Raja, who worked with me on the CRY2 project for almost a year and with almost unreasonable dedication. Thanks also to the new team at the University in Zürich, especially its early pioneers Chelsea Kidwell and Greta Ebnicher.

Many thanks to the current and former members of the Peri group, who welcomed us with open arms when we invaded their lab. In particular, thanks to Ambra Villani, Jørgen Benjaminsen and Nils Norlin for our collaboration on 'bubble brain' and thanks to Marvin Albert for our 'Tischtennis' sessions and all the help with computers and maths. May the vectors be with you. Last but not least, thanks to Katrin Möller for sharing your enthusiasm and crazy stories from Iceland. Poor sheep!

Tremendous thanks to all those who are less visible but without whom our work would have been immeasurably more difficult: Andreas Kunze, Officer Sabine Görgens, Kerstin Richter, Antti Lempiäinen and Rick Kluiver, as well as all the support staff in IT, lab kitchen, workshop, graduate office and administration.

Thanks to the EMBL core facilities, both for your work to keep the tech running and for your support in using it. As almost all of the data in this thesis was generated using a light microscope, special thanks go to the ALMF, in particular Yury Belyaev, Christian Tischer, Kota Miura, Stefan Terjung, Aliaksandr Halavatyi, Volker Hilsenstein und Tobias Rasse.

Another great aspect of my time here has been the extracurricular work. Teaching people how to code was particularly fun, even more so because of the smooth collaboration with Bio-IT and EMBL's computational centres, in particular Toby Hodges, Karin Sasaki and Malvika Sharan, but also many others.

Thanks to my thesis advisory committee for their time and patience: Stefano De Renzis, Ursula Klingmüller, Lars Hufnagel and – joining for the defense commission – Steffen Lemke.

Huge thanks also to my partner, Julia Suter, for bearing with me when I babble on and on about science and for supporting me in every imaginable way. Importantly, my foray into machine learning in this thesis was partially inspired and supported by her fascinating work on natural language processing.

Last but not least, big thanks to the many good friends and colleagues who helped make EMBL an incredibly stimulating environment for me but didn't quite fit into any of the sections above, including Nade Abazova, Allyson Ryan, Jin Wang, Renato Alves, and many others.

I'm sure there are a few important names I've missed and many more I can't feasibly list here; it's late evening now and I want to print this thing tomorrow. If you have been unjustly omitted, please let me know and I'll make it up to you by paying for drinks the next time we meet.

Cheers!

Appendix

A1 Sequences

PA-Cxcr4b

5'_ATGAACGGGACCGAGGGCCCAAACTTCTACGTGCCTTTCTCCAACAAGACGGGCGTGGTGCGCAGCCCCTT
CGAGGCCCCGCAGTACTACCTGGCGGAGCCATGGCAGTTCTCCATGCTGGCCGCCTACATGTTCCTGCTGATCA
TGCTTGGCTTCCCCATCAACTTCCTCACGCTGTACGTCATGGGCTTCCAGAAGAAGTCAAAGAACATGCTCAAC
TACATCCTGCTCAACCTGGCCGTGGCCGACCTCTTCATGGTCTTCGGGGGCTTCACCACCACCCTCTACACCTCT
CTGCACGGGTACTTCGTCTTTGGGCCCACGGGCTGCAACCTGGAGGGCTTCTTTGCCACCTTGGGCGGTGAAA
TTGCACTGTGGTCCTTGGTGGTCCTGGCCATCGAGCGGTACGTGGTGGTGGTACGTGCCACAAACAGCCAAAA
CTTAAGGAAATTGCTTGCCATCATGGGCGTCGCCTTCACCTGGGTCATGGCTCTGGCCTGTGCCGCGCCCCCCC
TCGTCGGCTGGTCCAGGTACATCCCGGAGGGCATGCAGTGCTCGTGCGGGATTGACTACTACACGCCCCACGA
GGAGACCAACAATGAGTCGTTCGTCATCTACATGTTCGTGGTCCACTTCATCATCCCCCTGATTGTCATATTCTT
CTGCTACGGGATCATCATCTCAAAACTGTCCAAGAACTCCAAGGGTCAGACTCTAAAGAGGAAGGCACTGCGC
ATGGTGATCATCATGGTCATCGCTTTCCTAATCTGCTGGCTGCCCTACGCTGGGGTGGCGTTCTACATCTTCACC
CATCAGGGCTCTGACTTTGGCCCCATCTTCATGACCATCCCGGCTTTCTTTGCCAAGACTTCTGCCGTCTACAAC
CCCGTCATCTACATCATGATGAACAAGCAGTTCCGGAAATCTGCCCGTAACGCTCTGAGCATCAGCAGTAGATC
CAGTCACAAGATGCTGA_3'

zfRGS4Δ

5'_ATGCAAGAGCAAAAGACTCTGAAGGAAAAAGAGAAGGAGAAAGACAAGGAGAAGGTGAAGGACACTGT
GGTCAACAGAATCACTCCTGCCGAAACTGAGAAATGGAAAACATCATTTACCAACCTGATCAAAAACGACGAC
GGTCGCAAGGCTTTCGCATCCTTCCTACAGTCCGAATACAGTCAAGAGAACATTGAATTCTGGGTAGCCTGTGA
GGATTTCAAGCAGACGCCAGCAGACAAGATGAACCTGAAAGCCAGAAATATATTTGAGCGATACATTGAGGC
CGATTCCCCCCGTGAGGTCAATCTGGATTCAGTCACCAGGGAGCAGACCAGAAAGAACCTGGAAATGTGTGAT
GTTTCATGTTTTGACGAAGCTCAGAGTAAAATCTTCACTCTTATGGAAAAAGACTCGTATCGGCGATTCCTGAG
ATCCAGATTGTTCCTGGAACTGTCTCAACCTGCGATGGACAACAAACCCTGTGGTTTAGAGAAGAAAGTAAAG
CGACAGATTTCTGACTACAGTCAGTGTTTGCCTAGTTATGCCTAA_3'

Tiam1a-Gd

5'_GGCTCACAAGAGTACAGAGCAGGTGACCGCCTTCTGCCGCAACCTTCATGATATGAATTCCAACGAGGGTC
CCGTCTCCTGTTCCTCATCTTCCTCCACTTCTTCCTCATCTTCATCCTGCATGCCCAGCCCAATCTCTCCACTGCCT
GGTCTCTTCACCCCTCGCCAGCTCTCTGACGCAGATAAACTCCGCAAGGTCATCAGTGAGCTGGTGGACACTGA
GAGGACCTACGTTAAGGACCTGAACATTTTAATAGAGCGCTACCTGAACCCACTGCAGAAGGAGAGCTTCCTC
ACCCAGGATGAGCTGGACGTGCTGTTTGGAAACTTGGCGGAGATGGTGGAATTCCAAGTGGAGTTTCTGAAA

ACTCTAGAAGATGGAACCAGATTAGTTCCAGATTTAGACAAACTGGAGAGAGTCGATCAGTTTAAGAAAGTTC
TGTTTTCTCTTGGTGGATCTTTCCTCTACTATGCGGACCGTTTTAAGATCTACAGTGCATTTTGTGCCAGTCACAC
AAAAGTCCCCAAGGTCCTTACCAAAGCTAAAACAGACCCAGAGTTTAAGGCATTTCTGGCTGAGAGGAACCCC
AGACAGCAGCATTCCTCCACACTGGAGTCCTACCTGATCAAACCCATTCAGAGAGTCCTAAAATACCCACTGCT
TCTGAGGGAGCTTTACTCGCTCACCGACCCCGACAGCGAGGAGCACTACCACCTGGATGTGGCGATGAAAGCC
ATGAACAAAGTGGCCAGTCACATAAACGAGATGCAGAAGATTCACGAGGAATACGGAGCCGTGTTTGACCAA
CTCATCAGTGAACAGAGCTCAGAAAAGAAAGAGGTTGCTGATCTGTCAATGGGCGACCTCTTGTTGTATGACA
CAGTGGTCTGGATTAACCCACCGTCCTCTTTGATGAAGGGGAAGAGAGACCCAGAGCTGGCTGCCTTTGTTTT
CAGAACAGCGGTTGTTTTTGTGTGTAAGGACTGCTCCAAGCAAAAGAAAAAAATCGGTGGACCTCACAGAGTG
TCAACCCTTGATGAGAGAGACCCATTTCGTT_3'

ARHGEF25b-Gd

5'_GTGCTGACAGAGCTGGTAGAAACGGAGAAGCTGTATGTGGAAGATCTGGGGCTTGTTGTTGAGGGTTATA
TGCTTACAATGAGAAGTTTTGGAGTGCCCGAGTATTTGGAAGGAAAGGACAAAATAGTCTTTGGGAACATTCA
TCAGATCTACGACTGGCATAAAGACTATTTTCTTGGAGAGTTGGAGAAATGTGTGTCTGAACCCGACCTGCTA
GCACAGCTGTTTATTAAACATGAGAGACGGCTCAACATGTATGTTGTCTACTGTCAGAACATGCCAAAGTCAG
AGCACATCGTCTCAGAGTACATCGAAACTTATTTTGAGGATCTGAGGCAACAGCTGGGTCACAGACTGCAGTT
AAATGACCTGCTCATCAAACCCGTTCAGAGAATCATGAAATATCAGCTGCTGCTGAAGGACTTCCTGAAGTACT
ACACTAAAGCGGGCAGACAAACAGAAGATCTTGAGAGGGCCGTTGAGGTCATGTGTTTTGTTCCAAAGCGAT
GCAATGAC_3'

A2 Engineered Features

Sample Level

Features characterizing the entire primordium (sample). These are the same for every cell of a given primordium.

- covars.img.sample.**cellnum**
  - Number of cells in a given primordium

- covars.img.sample.**volume**
  - Segmented volume of entire primordium

- covars.pcl.sample.**extents**
  - Maximum extents of entire tissue along Z, Y and X axis
  - Corresponds to maximum height, width and length of the primordium

- covars.pcl.sample.**aspects**
  - Aspect ratios of `covars.pcl.sample.extents`, specifically Z/Y, Z/X and Y/X

Tissue Level

Features relating each cell to the tissue context.

- covars.img.tissue.**centroids**
  - ZYX position of cell centroids in the image frame of reference

- covars.img.tissue.**bboxes**
  - ZYX slice objects to slice the cells' bounding box from the image

- covars.img.tissue.**neighbor_ids**
  - Labels of the cells' direct-contact neighbors

- covars.img.tissue.**neighbor_num**
  - Total number of the cells' direct-contact neighbors

- covars.img.tissue.**neighbor_contact_areas**
  - Contact area (in number of voxels) shared with cell's direct contact neighbors
  - Ordered the same way as `covars.img.tissue.neighbor_ids`
  - The 'outer' surface area of the cell is used for this measurement

- covars.img.tissue.**outside_contact_area**
  - Contact area (in number of voxels) shared with the outside of the primordium
  - The 'outer' surface area of the cell is used for this measurement

- covars.pcl.tissue.**centroids**
  - ZYX position of cell centroids in TFOR

Cell Level (Segmentation-Based)

Cellular features based on the segmentation mask.

- covars.img.cell.**volume**
  - Total volume of cells (in number of voxels)

- covars.img.cell.**surface_area**
  - Total surface area of cells (in number of voxels)
  - The 'inner' surface area of the cell is used for this measurement

Cell Level (Intensity-Based)

Cellular features based on the intensity distribution of a given channel.

- covars.img.cell.{channel_name}.**sum_total**
  - Sum intensity within the entire cell

- covars.img.cell.{channel_name}.**mean_total**
  - Mean intensity within the entire cell

- covars.img.cell.{channel_name}.**sum_membrane**
  - Sum intensity in the membrane (shell) region of the cell

    - The thickness of the shell region is user-determined (here 3pxl)

- covars.img.cell.{channel_name}.**mean_membrane**
    - Mean intensity in the membrane (shell) region of the cell
    - The thickness of the shell region is user-determined (here 3pxl)

- covars.img.cell.{channel_name}.**sum_inside**
    - Sum intensity of the inside (core) region of the cells
    - This is equal to `sum_total - sum_membrane`
    - The core region is determined by the thickness of the shell region (here 3pxl)

- covars.img.cell.{channel_name}.**mean_inside**
    - Mean intensity of the inside (core) region of the cells
    - The core region is determined by the thickness of the shell region (here 3pxl)

- covars.img.cell.{channel_name}.**sum_apical**
    - Sum intensity on the "apical half" region of the cells
    - The apical half is determined as the section above the midslice of the bounding box

- covars.img.cell.{channel_name}.**mean_apical**
    - Mean intensity on the "apical half" region of the cells
    - The apical half is determined as the section above the midslice of the bounding box

- covars.img.cell.{channel_name}.**sum_basal**
    - Sum intensity on the "basal half" region of the cells
    - The basal half is determined as the section below the midslice of the bounding box

- covars.img.cell.{channel_name}.**mean_basal**
    - Mean intensity on the "basal half" region of the cells
    - The basal half is determined as the section below the midslice of the bounding box

- covars.pcl.cell.{channel_name}.**extents**
    - Maximum extents of the cell tissue along Z, Y and X axis (in TFOR)
    - Corresponds to maximum height, width and length of the cell

- covars.pcl.cell.{channel_name}.**aspects**
    - Aspect ratios of `covars.pcl.cell.{channel_name}.extents`, specifically Z/Y, Z/X and Y/X

- covars.pcl.cell.{channel_name}.**extents_pca**
    - Maximum extents of the cell along the first, second and third principal component axis

- covars.pcl.cell.{channel_name}.**aspects_pca**
    - Aspect ratios of `covars.pcl.cell.{channel_name}.extents_pca`, specifically PC 1/2, 1/3 and 2/3

- covars.pcl.cell.{channel_name}.**sphericity**
    - Measures how much spherical the points of the points are distributed
    - It is the mean deviation from mean sphere around the centroid, normalized and inverted
    - Possible values are `0 < sphericity <= 1`
    - A sphericity of `1` means perfectly spherical, smaller sphericity means less spherical

    - *Note:* This currently relates to the *surface* of the sphere.
      A filled sphere may have a low sphericity!

- covars.pcl.cell.{channel_name}.**symmetry**
    - *Deprecated* in favor of `eccentricity`!
    - An indication of how balanced/symmetrical the distribution is around the centroid
    - It is the distance of center of mass from geometrical center, normalized and inverted
    - Possible values are `0 < sphericity <= 1`
    - A symmetry of `1` means perfectly symmetrical, smaller symmetry means less symmetrical

- covars.pcl.cell.{channel_name}.**distp_coords**
    - Coordinates ZYX of the "distal-most point", the point furthest from the centroid
    - In local cloud frame of reference (usually corresponds to TFOR in terms of rotation & scaling)

- covars.pcl.cell.{channel_name}.**distp_dist**
    - Distance from the centroid to the "distal-most point", the point furthest from the centroid

- covars.pcl.cell.{channel_name}.**distp_angles**
    - Angles from the centroid to the "distal-most point", the point furthest from the centroid
    - The angles are given as the angles of a 2D vector relative to a (0,1) reference vector
    - Three angles are given for three planes: ZY, ZX and YX

- covars.pcl.cell.{channel_name}.**nn_dists_mean**
    - Mean of nearest-neighbor distances among points in cells' point clouds

- covars.pcl.cell.{channel_name}.**nn_dists_std**
    - Standard deviation of nearest-neighbor distances among points in cells' point clouds

- covars.pcl.cell.{channel_name}.**all_dists_mean**
    - Mean of all pairwise distances among points in cells' point clouds

- covars.pcl.cell.{channel_name}.**all_dists_std**
    - Standard deviation of all pairwise distances among points in cells' point clouds

- covars.pcl.cell.{channel_name}.**cen_dists_mean**
    - Mean of all distances from the centroid (magnitudes) among points in the cells' point clouds

- covars.pcl.cell.{channel_name}.**cen_dists_std**
    - Standard deviation of all distances from the centroid (magnitudes) among points in the cells' point clouds

- covars.pcl.cell.{channel_name}.**eccentricity**
    - Eccentricities of the ellipsoid of the cells' point clouds
    - Sorted according to the extent of the principal semi-axes, from largest to smallest