Data-driven approaches to modelling collective cell migration



Simon Martina-Perez

Hertford College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2023

"Sicut cervus desiderat ad fontes aquarum,
ita desiderat anima mea ad te, Deus."

(Psalm 42:2 Psalterium Romanum)

Acknowledgements

First and foremost, I would like to thank Ruth for her generosity, wisdom, and patience as my supervisor. These years I have worked in total freedom on many different topics. This is all due to Ruth's immense faith in my ideas and judgement. My thanks are infinite.

I would like to thank papá, mamá, Emilio, and all the members of my family for their endless love and support, now and always. I have missed you dearly, but have always felt your presence from across the Channel.

During my time in Oxford, I have made truly wonderful friends in expected and unexpected places. Each of you has challenged me intellectually, supported me emotionally, and given me a place in the incredible communities that make up this university. You know who you are.

I have had an awesome time doing my DPhil at the Wolfson Centre for Mathematical Biology. Thank you all for being great. A huge thanks also go out to Isaac for sharing a big part of our joint experimental-theoretical PhD journey in Princeton (and talking about rowing too much).

My DPhil, and indeed my life in Oxford, is made possible by the tireless work of a great number of people who often go unacknowledged. It has has been a pleasure to interact with the people working in college, the department, and this city. Thank you.

Finally, I would like to thank the UK Engineering and Physical Sciences Research Council for supporting me with a Doctoral Training Award.

Statement of Originality

This thesis is submitted to the University of Oxford in partial fulfillment of the requirements of the degree of Doctor of Philosophy. This thesis contains no material which has been previously submitted for a degree or diploma at this University or any other institution.

This thesis contains experimental work from collaborations, which I have used in applying, calibrating, and validating my own theoretical modelling work. The experimental data used in Chapter 1 were provided by Dr Heba Z. Sailem (University of Oxford). The experimental data used in Chapters 2 and 3 were obtained by Dr Daniel J. Cohen and Isaac B. Breinyn (Princeton Univeristy). The experimental data used in Chapter 4 have previously been published by Wolf *et al.* [1]. The data used in Chapter 5 were obtained by Dr Jen Kulesa-Kasemeier (Stowers Institute for Medical Research) and Dr Paul Kulesa (Notre Dame University). Otherwise, this thesis represents my own original work in developing theory, modelling, and computational implementation.

Simon Martina-Perez
University of Oxford
November 2023

Abstract

Collective cell migration is the defining characteristic of many biological events involved in morphogenesis, regeneration, and pathology. This abundance, together with the immense clinical and therapeutic advances that would arise from the ability to control collective migration in specific applications in regenerative medicine and oncology, has led to an enormous increase in the number of studies on collective cell migration. Yet, the integration of large biological data sets with modelling has proven challenging. In this thesis, we address three main challenges: connecting mathematical models with high-dimensional data, leveraging biological data to learn new mathematical models, and controlling biological systems using mathematical models. In our first case study, we combine individual-based models with computational Bayesian statistics to identify links between genetic perturbations and cellular phenotypes in a large siRNA screen. We propose a new approach to approximate Bayesian computation (ABC) and show that it is possible to identify the functional impact of a range of different genetic perturbations. Second, in the context of Madin-Darby canine kidney (MDCK) epithelial monolayers, we use a continuum mechanics model to characterise the role of mechanical power expenditure in regulating active cellular forces. We then utilise this knowledge in a continuum model for collective electrotaxis which takes into account energy expenditure during migration, and develop an optimal control framework to achieve experimentally desirable outcomes. By deriving a Keller-Segel type model of collective electrotaxis, we propose design of optimal electric fields that vary in space and in time. Finally, by combining single-cell tracking with deep attention networks, we discover a set of genes critical for the invasion of c8161 melanoma cells: failure to express these genes leads to abnormal leader-follower dynamics. Together, the contributions of this thesis are towards developing new approaches to data-driven mathematical modelling, using the derived models to control biological systems, and understanding the underlying mechanisms.

Contents

1 Introduction **1**

1.1 Grand challenges . . . 2

1.1.1 Controlling biological systems . . . 2

1.1.2 Connecting mathematical models with high-dimensional data . . . 3

1.1.3 Learning new models from data . . . 5

1.2 Aims and outline . . . 6

2 Efficient Bayesian inference for mechanistic modelling of collective cell invasion with high-throughput data **9**

2.1 Introduction . . . 9

2.2 Methods and model . . . 13

2.2.1 Data from a gene knockdown screen . . . 13

2.2.2 An individual-based mechanistic model of density-dependent cell movement, proliferation, and death . . . 13

2.2.3 Bayesian statistics: connecting models and data . . . 15

2.2.3.1 Approximate Bayesian computation . . . 16

2.2.3.2 Minibatch-ABC-SMC . . . 17

2.2.3.3 Summary statistics . . . 21

2.3 Results . . . 22

2.3.1 Small batch sizes provide reliable inference . . . 22

2.3.1.1 Variability of the posterior distributions . . . 23

2.3.1.2 Reliability of minibatch ABC-SMC using experimental data 23

2.3.2 Detecting differences between different experimental conditions . . . 24

2.3.3 High-throughput identification of mechanistic effects of gene knockdowns . . . 28

2.3.3.1 Clustering analysis reveals functional subgroups . . . 28

2.3.3.2 Density-dependent patterns of proliferation reproduce experimental patterns of cell count changes . . . 31

2.3.3.3 Finding mechanistic contributions to wound closure . . . 33

2.4 Discussion and outlook . . . 35

3 Cellular cruise control: mechanical power acts as a regulator of collective migration in epithelia 38
3.1 Introduction . . . 38
3.2 Data and methods . . . 41
3.2.1 Experimental set-up . . . 41
3.2.2 Image processing and velocity computations . . . 42
3.3 A model for cellular mechanics in a migrating epithelial monolayer . . . 43
3.3.1 Computing cellular tractions from stress balance . . . 43
3.3.2 Constitutive assumptions for MDCK epithelia . . . 44
3.3.3 Mechanical parameters . . . 45
3.3.4 Computing mechanical work in a migrating epithelium . . . 45
3.3.5 Computing average mechanical power in a migrating monolayer . . . 47
3.4 Mechanical homeostasis is maintained by regulating tissue power expenditure 48
3.4.1 Quantifying mechanical trends in migrating epithelia . . . 48
3.4.2 External drivers of collective migration are regulated by tissue power generation . . . 52
3.5 The energetic history of monolayers predicts future tissue velocity . . . 56
3.5.1 Energetic history predicts future tissue migratory behaviour . . . 56
3.6 Power regulation dynamics of single cells . . . 57
3.6.1 High and low power spending residence times follow a power law distribution . . . 58
3.6.2 High and low spenders: velocities of single cells are regulated by average mechanical power during migration . . . 60
3.7 Regulation is dynamic in time and adapts to external stimuli . . . 62
3.8 Discussion . . . 64
3.8.1 Understanding the biological origin of the discovered power regulatory mechanism . . . 64
3.8.2 Incorporating mechanical power in mathematical models of collective cell migration . . . 64

4 Optimal temporal control of collective electrotaxis in epithelial monolayers 66
4.1 Introduction . . . 66
4.2 Methods and data . . . 69
4.2.1 Different stimulation traces . . . 69
4.2.2 Experimental methods . . . 70
4.2.3 Data acquision at the tissue bulk . . . 70
4.3 An adaptation-excitation model of collective electrotaxis . . . 71

4.3.1 Model derivation . . . 71
4.3.2 Linear stability analysis . . . 73
4.3.3 Bayesian inference of model parameters . . . 74
4.3.4 Optimal control . . . 76
4.3.5 Optimal control for the terminal velocity . . . 80
4.3.6 Optimal control for constant velocity . . . 81
4.3.7 Constant velocity during a small time interval . . . 85
4.3.8 Comparison to bang-bang control . . . 89
4.3.9 Discussion . . . 89
4.4 Incorporating cell metabolism in electrotaxis . . . 90
4.4.1 Adaptation alone cannot explain temporal dynamics in a double-stimulation experiment . . . 91
4.4.2 A model for active regulation mediated by mechanical work . . . 92
4.4.3 Linear stability analysis . . . 93
4.4.4 Bayesian inference of the mechanically regulated model . . . 95
4.4.5 Active regulation working in tandem with passive adaption . . . 95
4.4.6 Bayesian inference for the adaption-regulation model: regulation is needed to explain long time scale behaviour . . . 97
4.4.7 Separating the contributions of regulation and adaptation . . . 98
4.4.8 Discussion . . . 100
4.5 Optimal control of a double stimulation experiment . . . 101
4.5.1 Optimal control problem . . . 101
4.5.2 Role of different stimulation interval lengths . . . 103
4.6 Discussion and outlook . . . 104

5 Optimal spatial control of collective electrotaxis in epithelial monolayers 106
5.1 Introduction . . . 106
5.2 Experimental data and methods . . . 108
5.3 Modelling spatial heterogeneity during electrotaxis . . . 109
5.3.1 An ODE model for vector superposition of competing migratory cues . . . 109
5.3.2 Bayesian inference of model parameters in the superposition model 112
5.4 Construction and validation of a continuum model for collective electrotaxis 114
5.4.1 Model development . . . 114
5.4.2 Bayesian inference of model parameters . . . 117
5.4.3 Tissue size affects migratory speed . . . 118
5.5 Optimal design of electric fields to achieve spatially uniform migration velocity 120
5.5.1 Physical insights from a one-dimensional model . . . 121

5.5.2 Design of electric fields for arbitrary shapes . . . 123
5.6 Discussion and a starting point for future modelling . . . 125
5.6.1 Spatially varying electric fields: consequences for modelling . . . 125
5.6.2 Understanding recoil in the top edge post-stimulation . . . 125
5.6.2.1 Model and constitutive equations . . . 126
5.6.2.2 Bayesian parameter identification . . . 127

6 Distinguishing the impact of gene knockdowns on collective invasion using deep attention networks 131
6.1 Introduction . . . 131
6.2 Data and methods . . . 132
6.2.1 A set of 45 candidate genes for cell invasion . . . 132
6.2.2 A rubber stopper high-content screening assay . . . 133
6.3 Statistical assessment of cell invasion dynamics . . . 134
6.3.1 Invasion is affected primarily by cell motility rather than proliferation 135
6.3.2 Leader-follower dynamics are a hallmark of c8161 melanoma collective migration . . . 137
6.3.3 Loss of leader cell behaviours upon genetic perturbation . . . 138
6.4 Quantifying cell-cell interactions using deep attention networks . . . 140
6.4.1 Distinct patterns in collective cell migration after alteration after gene knockdowns . . . 141
6.4.2 Alterations in attention maps are not related to the biological function of individual siRNA knockdowns . . . 144
6.5 Discussion and outlook . . . 145
6.5.1 Experimental follow-up to determine the biological origin of anisotropic attention between invading cells . . . 146
6.5.2 Incorporating anisotropic attention in mathematical models of interacting particles . . . 147

7 Conclusions and outlook 149
7.1 Controlling biological systems . . . 149
7.2 Connecting mathematical models with high-dimensional data . . . 151
7.3 Learning new models from data . . . 152
7.4 Concluding remarks . . . 154

A Uncertainty quantification for Chapter 2 **155**
A.1 Data processing: removing data artifacts . . . 155
A.1.1 Presence of both wound edges and final condition . . . 155
A.1.2 Presence of extruded cells in the wound area . . . 156
A.2 ABC-SMC parameter sampling . . . 156
A.2.1 Minibatch-ABC . . . 156
A.2.2 Choice of knockdown kernel . . . 158
A.2.3 Model calibration: number of ABC-SMC generations . . . 159
A.3 Justification of model reduction . . . 159
A.4 Reliable inference using small batch sizes . . . 161
A.4.1 Variability of the posterior mean . . . 162
A.4.2 Variability of the posterior variance . . . 163
A.4.3 Performance of different batch sizes on experimental data . . . 165
A.4.4 Capacity of inference with very small batch sizes on experimental data 166
A.5 Robustness of model predictions to cell size esimates . . . 168
A.6 Changing the number of clusters for K means clustering . . . 169
A.7 Confidence intervals for model parameters . . . 170
A.7.1 Confidence intervals for posterior means of Mock, CDC42, CDH5 knockdowns . . . 171
A.7.2 Confidence intervals for posterior means of different initial wound sizes 172

B Statistics for Chapter 6 **173**
B.1 Full free space and cell fold counts . . . 173
B.1.1 C8161 cell line . . . 173
B.1.2 HT1080 cell line . . . 174
B.2 Full edge and bulk cell displacement data . . . 175
B.2.1 C8161 cell line . . . 175
B.2.2 HT1080 cell line . . . 176

Bibliography **177**

List of Figures

2.1 Schematic representation of the process of combining mechanistic modelling and scratch assay data. . . . 12
2.2 Two initial conditions from the assays conducted on Mock knockdown cells. 18
2.3 Posterior predictive check for the individual based model calibrated to wound healing data. . . . 25
2.4 Posterior distributions for three genetic perturbations with large numbers of observations. . . . 25
2.5 Posterior distributions for model parameters when inference is carried out on subsets of the data based on wound size. . . . 27
2.6 Pairwise distributions of the posterior means for each of the knockdown experiments in the high-throughput screen. . . . 30
2.7 Identifying functional subgroups among perturbations. . . . 31
2.8 Observed cell fold change as a function of intrinsic net proliferation. . . . 33
2.9 Observed wound area fold change as a function of cell fold change. Across the range of cell fold changes, starkly different rates of wound closure are observed. . . . 34
2.10 Observed wound area fold change plotted against the different model parameters. . . . 35
2.11 Observed wound area fold change plotted as a function of combinations of posterior means. . . . 35

3.1 Spatial organisation of velocities and forces in a freely migrating epithelial monolayer . . . 49
3.2 Spatial and temporal distribution of average mechanical power in a freely expanding epithelial monolayer. . . . 50
3.3 Interplay between power, velocities, and forces in a freely migrating epithelial monolayer. . . . 51
3.4 Energetics of externally driven epithelial migration. . . . 53
3.5 Dynamic behaviour of velocity and power within a collectively moving epithelium. . . . 57

3.6 Empirical probability density function of single-cell residence times in high- (low-) energetic spending states. 59
3.7 Log-log plot of the empirical probability density function of single-cell residence times in high- (low-) energetic spending states. 59
3.8 Speeds of high and low power spending cells during collective migration. . 61
3.9 Dynamics of regulation during an external stimulus. 63

4.1 Experimental data of MDCK monolayer electrotaxis with different electric field stimulation protocols. 69
4.2 Experimental data of bulk velocity decay post-stimulation together with least-squares fit of exponential decay. 75
4.3 Bayesian inference on a simple model of adaptation and excitation using experimental data from an electrotaxis experiment. 76
4.4 Stimulation patterns for optimal control of collective electrotaxis. 82
4.5 Solution to optimal control problem for constant bulk speed during stimulation with an electric field. 83
4.6 Stimulation profiles and electric field strengths for varying terminal velocities. 84
4.7 Optimal stimulation to keep the velocity constant during stimulation. . . . 85
4.8 Optimal control to keep velocity constant in a small time frame. 88
4.9 Bayesian inference for the model using using only passive adaptation to the field. 92
4.10 Graphical summary and phase plane of an electrotaxis model incorporating cellular energetics. 94
4.11 Bayesian inference on a model of active regulation using experimental data from an electrotaxis experiment. 96
4.12 Bayesian inference for the model using including active regulation and passive adaptation. 98
4.13 Plots of the activity history, effective signal, and responsiveness function to the electric field, $\alpha(p)$, showing that the responsiveness wanes and as a result the tissue is pre-conditioned to be less responsive in the second electric field pulse. 100
4.14 Optimal control protocol for a double-pulse electric field experiment. . . . 103

5.1 Graphical summary of collective electrotaxis data from Wolf *et al.* 109
5.2 Bayesian inference with a model using vector superposition of the electric field with endogenous cues. 114
5.3 Marginal posterior distributions for a one-dimensional continuum model of collective electrotaxis in Equation (5.24). 118
5.4 Comparison between model predictions and experimental data of one-dimensional cell density profiles during collective electrotaxis. 119

5.5 Effect of tissue size and geometry on maximum migratory velocity during electrotaxis. . . . 120
5.6 Spatial distribution of the electric field after one hour of electrotaxis to achieve uniform migration in the x-direction, for a circular tissue (top row) and a triangular tissue (bottom row). . . . 124
5.7 Bayesian inference for the top edge velocity model in Equation (5.38). . . . 128
5.8 Analysis of model solutions on a parameter grid. . . . 129

6.1 Schematic showing workflow of the experimental set up and the resulting experimental data. . . . 135
6.2 Differences in displacement lead to different invasion outcomes in c8161 cells. 139
6.3 Loss of leader-cell type behaviour upon genetic perturbation. . . . 140
6.4 Deep learning reveals altered patterns of attention between cells. . . . 143
6.5 Clustering of knockdowns by their free space invasiveness shows no relationship with quantitative metrics of their respective attention networks. . . . 143
6.6 Genes clustered by knockdown show no systematic relationship in metrics from attention maps compared to that of Neg. . . . 145

A.1 KL divergence between the posterior distributions at one generation and the next, plotted for a total of 10 generations. . . . 160
A.2 Boxplots for the means of the posterior distributions obtained by performing minibatch ABC-SMC on data generated in parameter regime I when the batch size is varied. . . . 163
A.3 Boxplots for the means of the posterior distributions obtained by performing minibatch-ABC-SMC on data generated in parameter regime II when the batch size is varied. . . . 163
A.4 Posterior distributions obtained using the two synthetic data sets corresponding to parameter regimes I and II. . . . 164
A.5 Boxplots for the standard deviations of the posterior distributions obtained by performing minibatch-ABC-SMC on parameter regime I when the batch size is varied. . . . 165
A.6 Boxplots for the standard deviations of the posterior distributions obtained by performing minibatch-ABC-SMC on parameter regime II when the batch size is varied. . . . 165
A.7 Posterior densities of model parameters for the Mock data set. . . . 166
A.8 Boxplots showing the means of the posterior distributions found by performing minibatch-ABC-SMC on *miniature data sets* created by taking subsets of the data. . . . 167

A.9 Boxplots of the means of the posterior distributions found by performing ABC-SMC on subsets of the data as well as using minibatch-ABC-SMC. Subsampling yields more variable posterior means than minibatching. . . . 167
A.10 Boxplots of the variances of the posterior distributions found by performing ABC-SMC on subsets of the data as well as using minibatch-ABC-SMC. Subsampling yields larger posterior variances than minibatching. 168
A.11 Distribution of cell sizes in the dataset. 168
A.12 Relative error plotted against deviations of the cell area as input parameters for minibatch-ABC-SMC. 169
A.13 Sum of squared distances for K-means clustering as K is increased. 170

Chapter 1

Introduction

Collective cell migration is the defining characteristic of many biological events implicated in morphogenesis [2], regeneration [3], and pathology [4]. From gastrulation in early embryonic development [5] to wound healing [3], intricate collective cell migration events require the coordination of several thousands of cells, often necessitating highly patterned proliferation, extreme mechanical deformation, and interaction with the physical environment. The abundance of collective migration phenomena across all of life makes investigation of the precise mechanisms underlying such processes almost synonymous with understanding life itself. Moreover, the ability to control collective migration in specific applications in regenerative medicine [6] and oncology [7] would herald immense clinical and therapeutic advances. For this reason, recent years have seen an enormous increase in the number of studies into collective cell migration in all of its contexts.

Simultaneously, advances in experimental techniques, microscopy, and data processing have enabled researchers to collect vast amounts of experimental data at unprecedented spatial and temporal resolution. Automated high-throughput experiments, for example, allow researchers to study thousands of experimental and genetic conditions at once, often at single-cell resolution or finer. Given such data, detailed hypotheses about cell-cell interactions can be formalised using mathematical and physical models, which can be tested, calibrated, and refined using the data. In fact, it has been argued that "the closer the complete systematic understanding of cell behavior is, the more important modeling becomes and the more experiment and theory merge" [8].

Nevertheless, almost two decades after this statement by Mogilner [8], the integration of large biological datasets with detailed models remains challenging. Calibrating complicated mathematical models to experimental data is computationally costly, sometimes to the point of it being infeasible, which has necessitated the development of approximate and heuristic methods [9–11]. On the other hand, the complexity of high-dimensional data sets can make analysis so intractable that it is practically difficult to formulate testable

hypotheses from the data. Finally, perhaps more contentiously, the fundamental *weakness* of traditional mathematical modelling is that it encodes prescriptive hypotheses that may lead researchers to view the problem of interest through the lens of a mathematical *idée fixe*. The abundance of data should, instead, empower researchers to transcend their preconceptions about the underlying biological mechanisms by discovering new mechanisms computationally.

The remainder of this introduction will propose three grand challenges in data-driven modelling of collective cell migration, the role their answers may play in providing biophysical insight into the mechanisms underlying collective cell migration in a multiplicity of settings, as well as their application in controlling collective cell migration.

1.1 Grand challenges

In this section, we will discuss three challenges that involve detailed mathematical modelling with high-quality data. The three challenges are developed in order of conceptual complexity, starting with how to utilise mathematical modelling to design and control biological systems, then discussing the challenges in combining modelling with data, before arriving at the final grand challenge of how to use data to learn models from the system when little is known about the problem at hand.

1.1.1 Controlling biological systems

The grand promise of mathematical modelling in the life sciences is that it provides key mechanistic insights into the biological system under investigation. This grand promise, though, comes with a question: given that a model can *fit* the experimental data well, can it also *predict* what would happen if the conditions of the experiment were altered? If the answer to this is in the affirmative, we arrive at a grand challenge: given a suitable model that can predict a biological system's response to some experimental or clinical intervention, can it also be used to derive a strategy that will lead to a desired outcome?

Practitioners, experimental and clinical, could then leverage that understanding to control the behavior of their system of study. In such a case, an external intervention would be used to achieve a change in the desired collective behaviour of the experimental system. Conceptually, the question of designing an external intervention that achieves a desired outcome can be addressed mathematically, for example through the lens of optimal control theory [12]. While the mathematical tools to design bespoke experiments exist, there are few examples still of practical implementation. In this thesis, we will offer a case study addressing the question of translating a desired experimental outcome into a

mathematical optimisation problem. All the while, we offer a quantification of how certain one may be of the model predictions in these new circumstances.

1.1.2 Connecting mathematical models with high-dimensional data

Assessing whether a mathematical model can fit a given data set often reduces to finding the set of parameters that produce model outcomes that replicate the observed data. Often, in mathematical biology, the values of these parameters are thought to be physically meaningful, representing the magnitude of various effects that the model is intended to capture. When working with data, then, the ability to make quantitative statements about which parameter values for a given model are appropriate to describe a biological system will play centre stage. Modern applied mathematics has seen a colossal body of work designed to address this question. In broad strokes, there are three main questions that current scholarship has concerned itself with. First, given a data set, can parameter values of a model be determined unambigiously? In the usual terminology: are the model parameters *identifiable*? Second, given that the model parameters are identifiable, how does one find the 'best set of parameters to recapitulate the experimental data? Finally, how can one express uncertainty in parameter estimates, given the noise that is inherent in biological systems?

Formally, a model can be considered structurally identifiable if distinct parameter values imply distinct distributions of observations, and hence when it is possible to uniquely determine the model parameters using an infinite amount of ideal data [13, 14]. Since data is in practice not infinite and contains observation noise, a related concept is that of *practical* identifiability: the ability to identify the parameters given practical constraints on the data [13–15]. Studying the practical identifiability of models in systems biology has become a field in its own right, since in this field there is an abundance of instances of competing models designed to describe the same phenomena [13, 16, 17]. In this thesis, we will use concepts from identifiability theory to ensure that data are adequately selected for each modelling problem.

This work will feature techniques to find an optimal set of model parameters, given experimental data. For an identifiable model, we suppose that the data can be described as a realisation of the model, which is corrupted with some stochastic noise [14]. Given a model and a statistical description of the noise, one can express the *likelihood*, ℓ, that the data, $\mathcal{D}$, are observed as a noisy realisation of the model parameterised by a parameter set, $\boldsymbol{\theta}$. That is,

$$\ell = \ell(\mathcal{D}; \boldsymbol{\theta}). \tag{1.1}$$

Frequentist approaches to parameter estimation consider the problem of finding the parameter vector that maximises the likelihood, given the data. This parameter is called the maximum likelihood estimator and is usually denoted $\hat{\boldsymbol{\theta}}$ [18–20]. Maximising the likelihood function is often computationally difficult if the parameter space is high-dimensional [21], or impossible if the likelihood is intractable [9]. However, the more fundamental problem with a single parameter value to describe data is that biological data is inherently noisy and so there is uncertainty associated with any experiment. Therefore, a single parameter value, such as the maximum likelihood estimator, will fail to capture the *range* of parameter values that might be likely to produce model outputs similar to the observed data.

To address this, Bayesian statistics expresses model parameters as a stochastic variable with its own probability distribution [9, 10]. In the absence of data, prior beliefs about the likely value of the parameters are encoded in a prior distribution $\pi(\boldsymbol{\theta})$, which defines a probability density function for parameters $\boldsymbol{\theta}$. Given a parameter vector $\boldsymbol{\theta}$, the model defines a density function for the observations, which, when normalised, is called the likelihood, as in the frequentist case in Equation (1.1). Bayes' rule allows one to combine the likelihood with the prior distribution to give the posterior distribution:

$$P(\boldsymbol{\theta} \,|\, \mathcal{D}_{\text{obs}}) \propto \ell(\mathcal{D}; \boldsymbol{\theta}) \pi(\boldsymbol{\theta}). \tag{1.2}$$

Simply put, the the goal of Bayesian parameter estimation is to update prior beliefs about the model parameters with real-world data [9, 10, 22, 23].

In practice, the likelihood for complicated stochastic models is mathematically intractable and typically only accessible through simulation. Approximate Bayesian computation (ABC) is a popular likelihood-free tool to estimate the left-hand side in Equation (1.2) [9, 10]. It approximates the likelihood, ℓ, using repeated simulation of the model, accepting simulated parameters, $\boldsymbol{\theta}$, only if the resulting model output, $\mathcal{D}_{\text{sim}}(\boldsymbol{\theta})$, is closer than some threshold, ϵ, to the data, $\mathcal{D}_{\text{obs}}$:

$$P_{\text{ABC}}(\boldsymbol{\theta} \,|\, \mathcal{D}_{\text{obs}}) \propto P\left(d(\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{sim}}) < \varepsilon \,|\, \boldsymbol{\theta}\right) \pi(\boldsymbol{\theta}), \tag{1.3}$$

where d is some distance function that quantifies the difference between the data, $\mathcal{D}_{\text{obs}}$, and model output, $\mathcal{D}_{\text{sim}}$. In many applications of ABC, the data are high-dimensional, meaning that, from a computational point of view, it is necessary to summarise the data in lower-dimensional summary statistics so that the distance $d(\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{sim}})$ in Equation (1.3) can be efficiently computed. We note that the choice of summary statistic is non-trivial and many works exist to address the problem of choosing appropriate summary statistics [24–26]. Computationally, many approaches to ABC are available, the most popular nowadays

being sequential Monte Carlo sampling (ABC-SMC) [9–11] and Markov Chain Monte Carlo (MCMC) sampling [27].

The grand challenge that we propose in this work is to develop novel algorithmic and summary statistic design for a practical application where the state-of-the-art ABC implementations are not able to perform computations within a feasible time. We remind the reader that the weakness of ABC is the cost of repeated simulation of the model, and comparison with data. In the context of a high-content scratch assay screen considered in Martina-Perez *et al.* [23], the amount of data made estimation of the posterior distribution using ABC computationally prohibitive even using state-of-the-art methods. Inspired by the method of stochastic gradient descent, we propose a *minibatch* approach to ABC, thus providing a solution to a computational bottleneck.

1.1.3 Learning new models from data

Testing specific hypotheses with mathematical modelling and experimental data lies at the heart of modern quantitative biology. However, prescribing a model *a priori* and calibrating it to data might be restrictive in cases with insufficient knowledge about the underlying mechanisms. Equation learning methods take a different approach and aim to directly *infer* a model that best describes the data. This is done by leveraging statistical and machine learning tools to learn, for example, the terms of a differential equation model directly from the data [22, 28–31]. There exist a large number of such approaches. One example is the PDE-FIND algorithm [28, 32], which aims to learn which terms out of a library of candidate terms for a partial differential equation (PDE) model to include in the final model, as well as their coefficients. Another example is biologically informed neural networks [29], an extension of physically informed neural networks [30], that have been developed to learn the different terms of a PDE model without the need to specify a library of possible terms. For an introduction to recent equation learning methods, we refer the reader to [31]. The fact that equation learning can discover previously undetected mechanisms, discriminate between competing models, or estimate biological quantities of interest that are difficult to measure experimentally, makes equation learning attractive to scientists working with real-world data.

In previous work, which is not included in this thesis, we have addressed the problem of uncertainty quantification when equation learning methods are used on biologically realistic data [22]. This was a timely contribution to the field, since practitioners wishing to develop models that they can use in real-world settings require, in addition, a thorough quantification of uncertainty [11, 33–36]. This need comes from the fact that the, often significant, noise in real-world data can impact the models predicted by equation learning methods, and hence the predictive capability of the models for unseen data or scenarios [10, 37]. In Martina-Perez *et al.* [22] we have demonstrated that noise can significantly impact

both the structure and the values of the parameters of learned differential equation models, and have developed a novel algorithmic strategy to combine equation learning using a library search with fully Bayesian uncertainty quantification.

Having understood that equation learning methods can be used on biological datasets to discover underlying mechanistic rules with practically acceptable fidelity, this thesis will propose a final grand challenge in data-driven approaches to collective cell migration. We will concern ourselves with understanding how equation learning approaches can be used in conjunction with high-dimensional data sets to explore how cell-cell interactions can change under different experimental and genetic conditions in a cell invasion assay. All the while, we will impose upon ourselves the additional constraint that such learnt equations should provide interpretable and experimentally tractable insights that can be relayed to experimental and clinical researchers working on these data sets.

1.2 Aims and outline

The overarching aim of this thesis is to use existing theoretical approaches to exploit the vast amounts of information contained in high-resolution biological data sets of collective cell migration, on the one hand, and to develop novel algorithmic and computational approaches when this abundance of information is problematic for existing approaches, on the other. In Chapter 2, we encounter the first application of this ethos. In this chapter, which is based on Martina-Perez *et al.* [23], we encounter a first problem with high-dimensional data, namely a high-throughput gene knockdown screen of human endothelial cells subject to genetic perturbations using RNAi in the context of a wound healing assay. While the dataset contains many different gene knockdowns, each experimental record only consists of two time points, which necessitates mathematical modelling to infer the interaction rules between the cells. We leverage a novel computational framework to combine mechanistic modelling with vast amounts of data and show that reliable inference can be performed at a fraction of the computational cost of a traditional Bayesian inference scheme. On the biological side, we find that density-dependent interactions play a crucial role in the process of wound healing.

Whereas the problem in Chapter 2 consists of developing a novel algorithmic approach to combine an existing model with high-dimensional model, we take a different approach in Chapter 3, where we leverage a wide range of computational tools to learn a new mathematical model from data. Using a continuum mechanics framework, we define what it means for an epithelium to migrate 'efficiently' and show that freely expanding epithelia regulate themselves to operate with maximal efficiency. By turning to the single-cell level, we understand the dynamics of single-cell energetic contributions to the collective, pointing to cell-level mechanisms of regulation through energetic expenditure.

By using straightforward mathematical modelling in conjunction with detailed microscopy measurements, this work complements existing biochemical metabolic studies by providing a novel physical mechanism for the regulation of active forces during epithelial migration.

Having established a novel mechanism that regulates the production of active forces in collective electrotaxis, we turn to using this insight in Chapter 4, where we aim to control the motion of the centre of mass of epithelial monolayers migrating during electrotaxis. Electrotaxis is the collective migration of epithelial monolayers that occurs when they are stimulated with a DC electric field (EF) [1,38–40]. In Chapter 4, we encode the mechanisms learned in Chapter 3 by defining an optimal control problem for the velocity of the bulk of a tissue undergoing collective electrotaxis. We identify various experimental challenges in designing electrotactic stimulation protocols and arrive at quantitative predictions for the optimal field design. In so doing, we discuss the optimality of existing experimental techniques and provide testable implementations to achieve various experimental goals.

Having established ways to understand and optimise collective electrotaxis by controlling for the velocity of the bulk of the monolayer, we turn to controlling the full spatial extent of the tissue. For this reason, in Chapter 5, we explore the spatial heterogeneity of epithelial electrotaxis. During electrotaxis, the EF serves as an exogenous cue that competes with and temporarily overrides endogenous cues within the tissue. These endogenous cues are present at the edges of the monolayer, which, in the absence of an EF, experience a characteristic outward motion; stimulation with an EF can temporarily slow down or even reverse this preferred direction of migration. By using detailed mathematical modelling, we explain the biological mechanism behind how cells integrate competing migratory cues at different locations. We then explore the continuous dependence of EF cue integration on the spatial location within the tissue. We encode the mechanisms underlying the spatial heterogeneity of collective electrotaxis in a reaction-diffusion model that incorporates random cell motility, directed movement in response to the EF, and proliferation. We use this model to generate spatially varying EFs that can be used to achieve a wide range of tissue engineering objectives. In this way, we contribute to the grand challenge of controlling biological systems.

In our final case study, we return to a problem of learning mathematical models directly from data. So far, the high-dimensionality of the data has provided us with opportunities to learn and calibrate models in Chapters 2-5. In particular, the data in Chapter 2 contained large numbers of experiments with a paucity of time points each, and mathematical modelling was used to find a mechanistic interpretation for the observed differences in the recorded wound healing assay data. A different methodological difficulty is posed by the data set we consider in Chapter 6. In this cell invasion assay, 45 candidate genes of interest were knocked down individually to study their effects in neural-crest derived carcinoma invasion. The availability of imaging at single-cell resolution in 10 minute intervals enables individual cell tracking. While the available temporal resolution could make this data an

ideal candidate to test and calibrate (existing) agent-based models, the question posed by our collaborators was to carefully characterise *how* cell behaviours and cell-cell interactions changed upon genetic perturbations. Therefore, we took an equation learning approach to learn the interaction rules between the constitutive cells. By constructing so-called deep attention networks [41,42], we detect changes in cell-cell interactions that are not otherwise observable from traditional single-cell tracking metrics, such as mean squared displacement and directionality. Our attention maps point to a select subset of four candidate genes that lead to severely altered leader-cell behaviour in neural-crest derived cancers, which was confirmed in subsequent *in vivo* testing. Biologically, this amounts to a successful application of equation learning methods to learn new rules of interaction. Mathematically, our interaction rules can be used, in the future, to build novel mathematical interacting particle models that encode our observed *in vitro* cell-cell interaction rules.

The research in Chapter 2 has arisen from a collaboration with Heba Z. Sailem (Kings College London). The research in Chapters 3 and 4 has arisen from a collaboration with Daniel J. Cohen and Isaac B. Breinyn (Princeton Univeristy). The experimental results shown in Chapter 5 have previously been published by Wolf *et al.* [1]. The research in Chapter 6 has arisen from a collaboration with Paul Kulesa (Stowers Institute for Medical Research). My contribution lies in data analysis, modelling, inference, and the development of methodology.

Chapter 2

Efficient Bayesian inference for mechanistic modelling of collective cell invasion with high-throughput data

The work in this chapter is published in Martina-Perez *et al.* [23]. The data used in the publications were originally published by Williams *et al.* [43]. My contributions consist of the main algorithmic design in this section, data curation, computational implementation and writing of the publication.

2.1 Introduction

High-throughput methods entail the acquisition and processing of vast amounts of experimental data with ever increasing detail. This wealth of data creates unique opportunities to derive insights about real-world phenomena. However, the quantitative metrics used for analysis of high-throughput data sets often do not exploit the full spatial and temporal information in the data. For example, in scratch assay experiments routinely used in cell biology studies, typically only cell count and wound area are used to interpret outcomes [44, 45]. Such straight-forward summary statistics can identify biologically relevant differences between different experimental conditions, but they are not generally suited to identify complex spatial and temporal phenomena that arise from, for example, interactions between individual cells. Further, while the ever increasing availability of

detailed experimental data has the potential to inform, calibrate and refine complicated mathematical models that incorporate such cell-cell interactions, the computational cost of this process quickly becomes a bottleneck when the amount of data to be analysed is increased. In this work, we seek to develop a method to combine mathematical modelling and high-throughput data to extract detailed, real-world insights from a large data set with a fraction of the computational cost of existing methods.

Mathematical modelling and experimental data are routinely combined using Bayesian inference, and of late there has been intense interest in both the opportunities offered by high-throughput data and models (see for instance the editorial overview by Hasenauer and Banga [46]) as well as the required protocols for combining mathematical modelling and high-throughput data successfully [47]. Given a mathematical model and observed data, Bayesian methods express the information gained from data in probability distributions for the parameters that constitute the model. Such *posterior distributions* provide information as to which model parameters reproduce observed data well while also expressing the associated uncertainty in these values given the data. One of the challenges in performing Bayesian analysis of complex mathematical models is that the likelihood function of simulated data given a parameter value is typically intractable, and so simulation-based approaches to likelihood estimation are needed. This means that large amounts of computationally costly simulations are required to compare model outputs with experimental observations, which is complicated further when the parameter set of the model is high-dimensional, as the cost of accurate likelihood estimation scales poorly with the dimension of the parameter space [48, 49]. The computational cost of likelihood-free Bayesian inference methods has prompted a huge volume of research into improving the computational efficiency of Bayesian methods [50], such as multifidelity methods [51], optimal perturbation kernels [52, 53], and delayed acceptance [54], to name but a few.

The central challenge for likelihood-free Bayesian inference with high-throughput data is that in practice one might need to repeatedly simulate a new instance of the model for each of the observed data points in order to accurately estimate the likelihood. This may be necessary, for instance, when the data consist of time series with highly variable initial conditions: a reasonable comparison between simulated and observed data can then only be made when each observed and simulated data point share the same initial condition. When the number of data points is large, such as with high-throughput data, the resulting computational cost becomes prohibitive, as the simulation cost grows linearly with the amount of data, in addition to growing exponentially with the dimension of the parameter space. Even when other state-of-the-art approaches are used, the fundamental bottleneck of repeated simulation remains. Inspired by the method of stochastic gradient descent (SGD) [55–57] in machine learning, we propose a *minibatch* approach to tackle this issue: for each comparison between simulated and observed data, we use a stochastically sampled subset (minibatch) of the data. A similar minibatch method has been employed

very recently by Stapor *et al.* [58] to successfully calibrate ordinary differential equation (ODE) models with a significant improvement in computational performance, and by Seita *et al.* [59] within the context of Markov Chain Monte Carlo (MCMC), likewise with a significant computational speed-up. We demonstrate that choosing a large enough minibatch ensures that the relevant signatures in the observed data can be accurately estimated, while avoiding unnecessary comparisons that slow down inference. We apply our minibatch approach to perform Approximate Bayesian computation (ABC) with a well-established stochastic individual-based model (IBM) of density-dependent cell migration, proliferation, and death [60, 61] on the high-throughput data set from Williams *et al.*'s scratch assay screen [43]. Scratch assays involve growing a cell monolayer to confluence and then mechanically removing (scratching away) a portion of the cell population to leave an "in vitro wound". Typically, a small region around the scratched area is imaged to provide information on the dynamics of wound closure [62]. Such assays are a simple, fast, and inexpensive method to investigate collective cell invasion under varying environmental or genetic conditions.

In the screen of Williams *et al.* [43], human dermal lymphatic endothelial cells (HDLECs) were used, both in a control setting and after knockdown of some 500 genes using RNA interference. Williams *et al.* [43] identified the effect of different gene knockdowns on wound healing by comparing the extent of wound closure after a period of 24 hours. While this metric is informative of general, large-scale, features of collective cell invasion, it does not intrinsically reflect the mechanisms that allow the different knockdowns to display different wound healing behaviours. In this study, we use computational Bayesian inference with an interpretable, mechanistic model to obtain detailed descriptions of cell-cell interactions. By calculating the posterior parameter distribution of the model for each gene knockdown, we characterise the functional behaviour of each perturbation according to the cell-cell interactions and cell-intrinsic mechanisms of the model. By analysing the extent of similarity between different gene knockdowns, we identify three major functional subgroups, which differ mainly according to their intrinsic and density-dependent movement rates. To our knowledge, this is the first time a mechanistic model has been used to interpret and classify high-throughput data to elucidate functional subgroups.

The structure of this work is as follows. In Section 2.2, we describe the high-throughput data, the IBM and the proposed minibatch Bayesian inference scheme. In Section 2.3, we characterise the performance of the method using different batch sizes and apply it to characterise functional mechanistic differences between different genetic perturbations. We conclude in Section 2.4 with a discussion of our results. All code and data to reproduce the study are publicly available at `https://github.com/simonmape/minibatch-ABC` and our Zenodo repository: DOI 10.5281/zenodo.5898532.

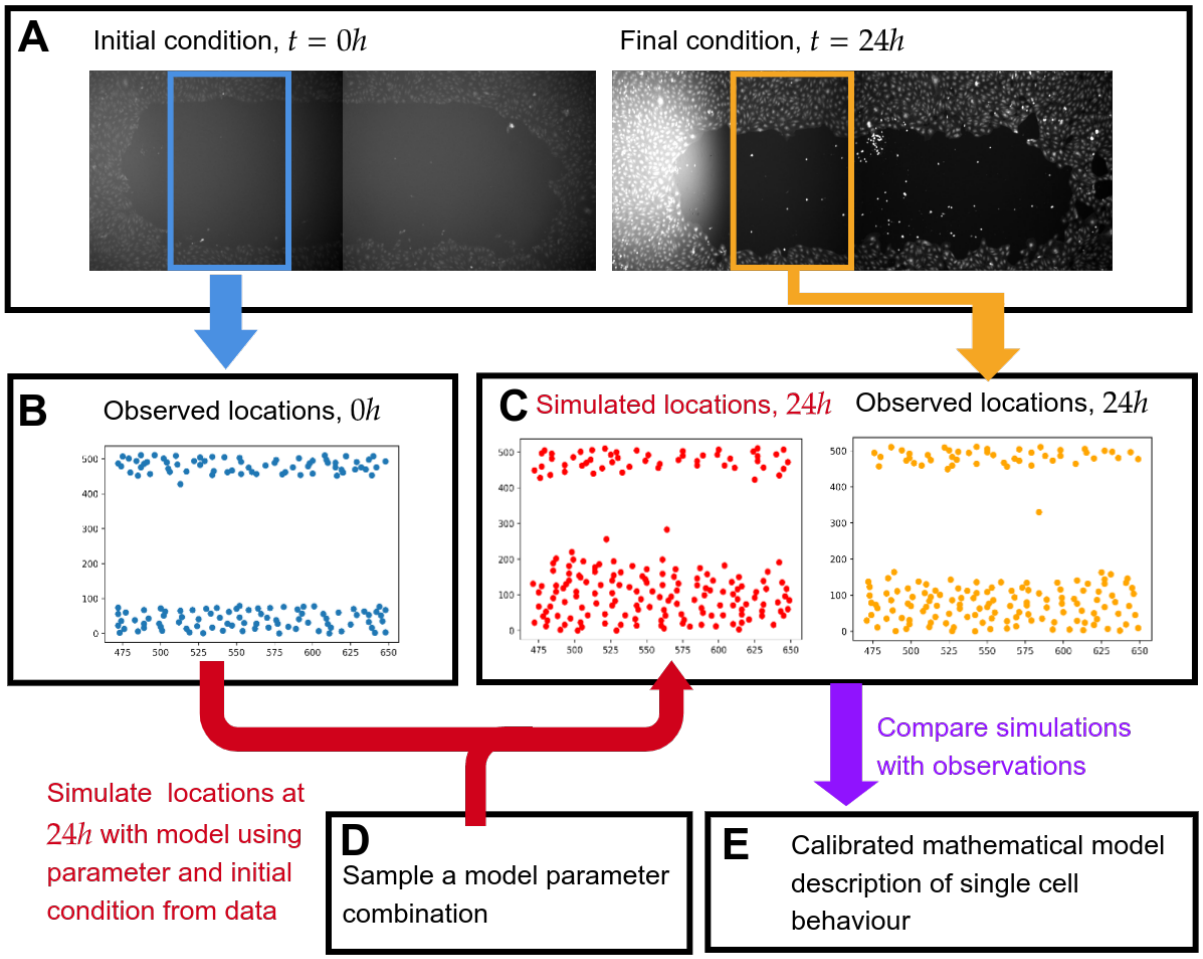


Figure 2.1: Schematic representation of the process of combining mechanistic modelling and scratch assay data. A: each observation in the data set consists of an initial condition at $t = 0$h and a final condition at $t = 24h$. B: the field of view of the initial condition is cropped and cell locations are extracted. C: the field of view of the final condition is cropped and cell locations are extracted, as for the initial condition. A mechanistic mathematical model is used to simulate cell locations at $t = 24$h using a sampled parameter from box D. E: by simulating many times with different parameter values we can compare which parameters generate simulated data that is similar to observations.

2.2 Methods and model

In this section, we detail the experimental data that will be used in this work (Section 2.2.1), outline the computational model (Section 2.2.2), and the Bayesian method used to combine model and data (Section 2.2.3).

2.2.1 Data from a gene knockdown screen

We used a previously published high-throughput siRNA screen of wound healing in HDLECs [43] where more than 500 gene knockdowns were identified as significant for altering wound healing [43]. During the wound healing experiment, images of the wound were captured at $t = 0$h and $t = 24$h after the scratch was made. Images of cells were acquired at 4x objective, which allowed the entire experimental well to be captured in two images that were stitched together, resulting in 512×1392-pixel images, where each pixel corresponds to $2.97\mu m$ [43].

We used DeepScratch [63] to analyse image data from various gene knockdowns. DeepScratch is a neural network with U-NET architecture. It detects cell centres from raw cell images stained with markers of the nucleus or membrane, and in addition it uses the locations of detected cells to segment the wound as the area that is void of cells. Following image analysis using DeepScratch, we computed various features including the distance of each of the cells to the wound edge and cell density. As the image data lack a cytoplasmic stain, we approximated cell area by generating a Voronoi tessellation where a polygon is defined for each cell centre. Cell locations are then given in a rectangular grid with dimensions 512×1392 pixels. Only assays where both an initial ($t = 0h$) and final condition ($t = 24h$) are present in the data set are included, and we also only include assays where both wound edges are contained in the initial condition. For each image, a wound mask was computed to isolate the extent of the wound and exclude dead and extruded cells. We describe the filtering process in detail in Appendix A.1. To exclude noise from measurements at the location where images are stitched together, and to ensure that the entire field of view contains a wound edge, we further reduce the field of view to a 512×180-pixel subset of the image, 10 pixels left of where the images were stitched together. Applying both filters to the data set yields 118 candidate gene knockdowns for investigation, with between two and 117 biological replicates per knockdown experiment.

2.2.2 An individual-based mechanistic model of density-dependent cell movement, proliferation, and death

In this work, we use a well-established stochastic, lattice-free IBM developed by Binny *et al.* [61] to model density-dependent cell migration, proliferation, and death. This model

has previously been shown by Browning *et al.* [60] to accurately represent collective cell migration in a wound healing assay experiment. Density-dependent behaviour in the model is incorporated through cell-to-cell interactions, so that cell behaviours depend on local crowding [60]. Following [61], we let $\mathbf{x}_n = (x_n, y_n)$ be the location of cell n, where $n = 1, \ldots, N(t)$ and $N(t)$ is the population size at time t. Interactions between cells are governed by a measure of crowding, where each cell senses the cells in its close proximity. Mathematically, the crowding function at location $\mathbf{x}$ is given by

$$B(\mathbf{x}) = \sum_{i=1}^{N(t)} \gamma_b w(\|\mathbf{x} - \mathbf{x}_n\|^2), \tag{2.1}$$

where the function w is an exponential kernel that quantifies the impact of local crowding upon cell behaviours. The specific choice of $w(r) = \exp(-r^2/2\sigma^2)$ has been found to give rise to outputs that match experimental data well, but other choices for w can also be made [61]. To simplify computations, we assume that $w(r) = 0$ whenever $r \geq 3\sigma$. The parameter $\gamma_b > 0$ in Equation (2.1) describes the magnitude of the crowding effect, and $\sigma > 0$ governs the length scale over which agents can interact with other agents. Browning *et al.* [60] suggest that a suitable choice for the interaction parameter σ is $\sigma = \varphi/2$, where φ is the cell diameter. In this work, we will estimate the cell diameter, φ, from experimental data by using the measured cell areas in the images acquired at $t = 0h$ and $t = 24h$. In Appendix A.5, we report this estimation procedure in detail.

The crowding function, B, governs the direction in which agents prefer to move and proliferate: agents prefer to move and proliferate along the steepest descent of B. The direction of steepest descent is also called the *bias vector*. For the n-th agent, we define the bias vector as

$$\mathbf{B}_n = -\nabla B(\mathbf{x}_n). \tag{2.2}$$

When a cell attempts to move in the model, the direction of movement, ϕ, is sampled from a von Mises distribution as

$$\phi \sim \text{von Mises}(\arg(\mathbf{B}_n), \|\mathbf{B}_n\|). \tag{2.3}$$

The rationale for this is that $\mathbb{E}[\phi] = \arg(\mathbf{B}_n)$, meaning that the direction of movement and proliferation is on average in the direction of the bias vector, and $\phi \sim \mathcal{U}(0, 2\pi)$ in the limit as $\|\mathbf{B}_n\| \to 0$, meaning that cells in uniform density regions exhibit unbiased movement.

Cell movement, proliferation, and death occur according to Poisson processes, which we simulate using the Gillespie algorithm [64]. Each cell's movement and proliferation rates depend on the extent of local crowding, whereas we let the death rate be constant, *i.e.*, independent of crowding. In Appendix A.3 we explore the effect of including a

density-dependent death term in the model. We compute Bayes factors for both models (with and without density-dependent death) using the resulting posteriors [65–67] to find that the data contains stronger evidence in favour of density-independent death.

The movement rate for the n-th cell is denoted by M_n, and the proliferation rate for the n-th cell is denoted by P_n. Each individual rate consists of a base, intrinsic rate, which is then modulated by the cell's interactions with its neighbours. The movement and proliferation rates for cell n are given by

$$M_n = \max\left\{0, m - \gamma_m \sum_{i=1, i\neq n}^{N(t)} \exp\left(-\frac{\|\mathbf{x}_n - \mathbf{x}_i\|^2}{2\sigma^2}\right)\right\}, \tag{2.4}$$

$$P_n = \max\left\{0, p - \gamma_p \sum_{i=1, i\neq n}^{N(t)} \exp\left(-\frac{\|\mathbf{x}_n - \mathbf{x}_i\|^2}{2\sigma^2}\right)\right\}, \tag{2.5}$$

and the death rate is a constant $d > 0$. The parameters γ_m and γ_p modulate the strength of a cell's response to crowding. For example, when $\gamma_m < 0$, local crowding increases motility, when $\gamma_m = 0$ motility is independent of local density, and when $\gamma_m > 0$ motility decreases with local crowding.

When a movement event occurs for any given cell, that cell moves one cell diameter, φ, in the direction prescribed by the distribution of the angle ϕ. When a proliferation event occurs instead, the cell places a daughter cell one cell diameter away in the direction of ϕ. We assume periodic boundary conditions to model the fact that the field of view is a small subset of a much larger tissue. In our implementation, we choose the domain to match the field of view of the cropped high-throughput imaging data. Hence, we simulate on a domain of size 180×512 pixels.

2.2.3 Bayesian statistics: connecting models and data

In this work, we are interested in how observed collective behaviours within the scratch assays might be described in terms of the mechanistic properties of individual cells. That is, given observed data of wound closure in the scratch assay, we wish to understand which parameters $\theta = (m, p, d, \gamma_m, \gamma_p, \gamma_b)$ might make such observations probable. The goal of Bayesian parameter estimation is to update prior beliefs about the model parameters of the IBM from Section 2.2.2 with real-world data. Prior beliefs are encoded in a prior distribution $\pi(\theta)$, which defines a probability density function for parameters θ. Given a parameter vector θ, the model defines a density function for the observations, which, when normalised, is called the likelihood, $P(\mathcal{D}_{\text{obs}} \,|\, \theta)$. Bayes' rule allows one to combine

the likelihood with the prior distribution to give the posterior distribution:

$$P(\theta \,|\, \mathcal{D}_{\text{obs}}) \propto P(\mathcal{D}_{\text{obs}} \,|\, \theta)\pi(\theta). \tag{2.6}$$

By using Bayes' rule, the prior is updated with data, which in this work consists of the repeated measurements of cell centre locations at $t = 0$h and $t = 24$h. In this work, we choose independent uniform priors for each of the parameters, and we let the prior ranges be informed by the findings of Browning *et al.* [60] so that $m \sim \mathcal{U}(0, 10)\text{h}^{-1}$, $p, d \sim \mathcal{U}(0, 0.05)\text{h}^{-1}$, $\gamma_b \sim \mathcal{U}(-2.5, 2.5)\text{h}^{-1}$, $\gamma_p \sim \mathcal{U}(0, 0.05)\text{h}^{-1}$, and $\gamma_b \sim \mathcal{U}(0, 50)\mu\text{m}$. We note that we estimate the cell area directly using DeepScratch, rather than assigning a prior distribution and performing Bayesian inference, as has been the standard approach in previous works, such as Browning *et al.* [60]. To understand the reliability of our results, in Appendix A we investigated the sensitivity of the posterior distributions to the input cell area parameter and find that for values of the cell area within three standard deviations of the median cell estimate, relative differences between the resulting posterior means are within 10% of the parameter value found using the estimated cell size, suggesting that the model predictions are robust. At the same time, we observe that a large variation exists in the measured cell areas across the high-throughput screen, such that it becomes unwise to assign a single value for cell area to all the knockdowns in the screen.

2.2.3.1 Approximate Bayesian computation

In practice, the likelihood for complex stochastic models is mathematically intractable and typically only accessible through simulation. Approximate Bayesian computation is a popular likelihood-free tool to estimate the left-hand side in Equation (2.6) [9, 66]. It approximates the likelihood, $P(\mathcal{D}_{\text{obs}} \,|\, \theta)$, using repeated simulation of the model, accepting simulated parameters θ only if the resulting model output, $\mathcal{D}_{\text{sim}}(\theta)$, is closer than some threshold, ϵ, to the data, $\mathcal{D}_{\text{obs}}$:

$$P_{\text{ABC}}(\theta \,|\, \mathcal{D}_{\text{obs}}) \propto P\left(d(\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{sim}}) < \varepsilon \,|\, \theta\right) \pi(\theta), \tag{2.7}$$

where d is some distance function that quantifies the difference between the data, $\mathcal{D}_{\text{obs}}$, and model output, $\mathcal{D}_{\text{sim}}$. In many applications of ABC, the data are high-dimensional, meaning that, from a computational point of view, it is necessary to summarise the data in lower-dimensional summary statistics so that the distance $d(\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{sim}})$ in Equation (2.7) can be efficiently computed. In Section 2.2.3.3, we describe our choice of summary statistic for this problem.

We use a version of sequential Monte Carlo sampling (ABC-SMC) [9, 66] as our implementation of ABC. In brief, ABC-SMC propagates a sample from the prior dis-

tribution, $\pi(\theta)$, through a sequence of intermediate distributions to approximate the target distribution, $P(\theta|\mathcal{D}_{\text{obs}})$. The intermediate distributions are a sequence of ABC approximations P_t for $t \in \{1, \ldots, T\}$. We use a sequential importance sampling approach to ABC-SMC [66, 68, 69], where each Monte Carlo sample $\{\theta_i^{(t)}, w_i^{(t)}\}$ built at generation t is used to construct an importance distribution at generation $t+1$. We define N_{gen} as the total number of parameters that are drawn from the importance distribution at each generation, and N_{acc} as the total number of parameters that are accepted. Given a Monte Carlo sample at generation t, the sample at generation $t+1$ is obtained by using a *perturbation kernel* $K(\cdot|t)$ to propose N_{gen} perturbed parameters from the Monte Carlo sample at generation t, from which the N_{acc} proposed parameters with smallest distance to the real-world data are accepted into the Monte Carlo sample for generation $t+1$. In this work, we use the Normal perturbation kernel proposed by Filippi *et al.* [52] and let $N_{\text{gen}} = 2 \cdot 10^4$ and $N_{\text{acc}} = 500$, meaning that each generation has an acceptance rate of 2.5%. The final sample, $\{\theta_i^{(T)}, w_i^{(T)}\}$, forms the output of the ABC-SMC algorithm. All implementation details are outlined in Appendix A.2.

2.2.3.2 Minibatch-ABC-SMC

In some problems, one model simulation can be compared with arbitrarily many observations, using the distance function d. This is particularly the case when the model is deterministic and one realisation of the model suffices to capture the relevant behaviour of the model. When the model is stochastic, however, a single model realisation will likely not provide sufficient information on the distribution of model outputs to sensibly compare to observed data. An additional, and significant, problem occurs with time series data because the initial conditions across different experiments can be vastly different. This means that for each comparison with a different datapoint, a new model simulation with a specific initial condition is needed to ensure that the comparison between the summary statistics of the simulated and observed data is fair and informative. For high-throughput data, the implications of these additional computational costs are significant. As an illustration of this point, the data set in this work consists of many hundreds of observations, each with different initial wound location, initial cell densities and initial topology. Figure 2.2 illustrates the extent to which two wells aimed at studying the same genetic profile can differ. The wells in Figure 2.2 are 'Mock' knockdowns, which is the use of a not-target protein siRNA that does not affect overall cell function. To apply Bayesian inference approaches to calibrate models to such data requires individual simulations, which mirror the initial conditions of each experiment separately.

When observed data, $\mathcal{D}_{\text{obs}} = \{y_{\text{obs}}^i\}_{i=1}^{N_{\text{obs}}}$, can only be compared individually to model simulations that share the same initial condition, the simulated data will equally consist of N_{obs} datapoints, such that $\mathcal{D}_{\text{sim}} = \{y_{\text{sim}}^i\}_{i=1}^{N_{\text{obs}}}$, and y_{sim}^i shares its initial condition with

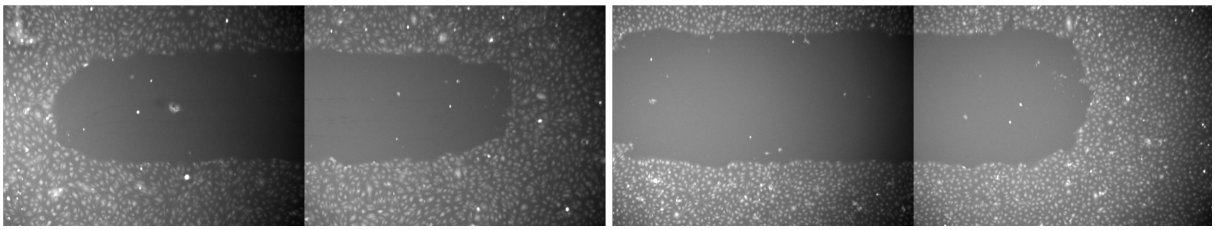


Figure 2.2: Two initial conditions from the assays conducted on Mock knockdown cells. Note the variation in both the location of the wound as well as the size of the initial wound, highlighting the need to carry out separate simulations, one for each initial condition.

y^i_{obs}. A scalar measure of the discrepancy between the observed and simulated data can then be obtained by taking the mean of distances between the observed data points and their corresponding model simulations:

$$d(\mathcal{D}_{\text{sim}}, \mathcal{D}_{\text{obs}}) = \frac{1}{N_{\text{obs}}} \sum_{i=1}^{N_{\text{obs}}} d(y^i_{\text{sim}}, y^i_{\text{obs}}). \tag{2.8}$$

When the size of observed data, N_{obs}, is large, *i.e.* $N_{\text{obs}} \gg 1$, and the model needs to be simulated separately for each data point, ABC-SMC approaches fail to scale well with the number of observations: simulation time – which is normally a bottleneck – increases linearly with the number of observations.

Drawing inspiration from minibatch methods in SGD in machine learning [55–57], we propose a minibatch approach to computing the distance in Equation (2.8): rather than using all N_{obs} observations to compute the distance whenever a parameter θ is proposed in ABC-SMC, we suggest computing the distance between observed and simulated data based on randomly chosen *batches* of the data. This is done by choosing a *batch size*, N_{bs}, such that $N_{\text{bs}} \ll N_{\text{obs}}$, that controls the number of datapoints considered in each comparison between model simulations and data. Formally, this amounts to drawing N_{bs} random numbers i_j independently with replacement, from $\{1, \ldots, N_{\text{obs}}\}$, obtaining the corresponding data point $y^{i_j}_{\text{obs}}$ and simulating a model output, $y^{i_j}_{\text{sim}}$, that shares the same initial condition. Then, the discrepancy between model and data can be approximated by

$$d(\mathcal{D}_{\text{sim}}, \mathcal{D}_{\text{obs}}) \approx \frac{1}{N_{\text{bs}}} \sum_{j=1}^{N_{\text{bs}}} d(y^{i_j}_{\text{sim}}, y^{i_j}_{\text{obs}}). \tag{2.9}$$

This approach immediately reduces simulation time by a factor $N_{\text{obs}}/N_{\text{bs}}$. The minibatch approach is simple to implement within the importance sampling part of the ABC-SMC framework described in Section 2.2.3.1. At each generation, t, an importance distribution is given (for $t = 1$ this is the prior distribution). Then, N_{gen} parameters are sampled according to this distribution. For each parameter a minibatch of size N_{bs} is sampled

from the data (with replacement) and for each observation within that minibatch a corresponding model simulation is generated given the parameter and observed initial condition. Mathematically, if $y^{\xi}_{\text{obs}}(0)$ is the initial condition corresponding to observed data point y^{ξ}_{obs}, the model simulates a corresponding output $y_{\text{sim}} \sim f\left(\bullet|\theta_n, y^{\xi}_{\text{obs}}(0)\right)$. The discrepancy between simulated and observed data is computed for each parameter using Equation (2.9). The method is summarised in Algorithm 1.

Algorithm 1: Minibatch-importance sampling

Input: Data, $\mathcal{D}_{\text{obs}} = \{y^i_{\text{obs}}\}_{i=1}^{N_{\text{obs}}}$; number of parameters to sample, N_{gen}; number of parameters to accept, N_{acc}; batch size, N_{bs}; prior distribution, $\pi(\theta)$; importance distribution, $\hat{q}(\theta)$; model, $f(\bullet \,|\, \theta)$; distance function, d.

Output: N_{acc} weighted samples $\{\theta_n, w_n\}_{n=1}^{N_{\text{acc}}}$.

for $i = 0, \ldots, N_{gen}$ **do**
  Generate parameter vector θ_i according to given importance distribution $\hat{q}(\theta)$;
  for $k = 1, \ldots, N_{bs}$ **do**
    Draw data index ξ uniformly at random from $\{1, \ldots, N_{\text{obs}}\}$;
    Simulate $y_{\text{sim}} \sim f\left(\bullet \,|\, \theta_n, y^{\xi}_{\text{obs}}(0)\right)$;
  end
  Calculate ϵ_i using Equation (2.9).
end
From $\{\theta_i, w_i\}_{n=1}^{N_{\text{gen}}}$, select the N_{acc} parameters with lowest value of ϵ_i;
Set $w_i = \pi(\theta_i)/q(\theta_i)$ for all accepted parameters θ_i.

At this point, we remark that an alternative to using minibatch ABC-SMC might be to *a priori* select a subset of the data (*e.g.*, using stratified sampling) and use ABC-SMC on this subset. While reducing the amount of data to be processed would certainly reduce the computational burden of ABC-SMC, this approach also vastly reduces the amount of available information. Moreover, subsampling *a priori* from the data might introduce bias into the problem, as it is up to the practitioner to define a sensible subset of the data. In Appendix A.4, we carefully examine the quality of the posteriors generated by subsampling from the data in comparison to those generated using minibatch ABC-SMC. We find that *a priori* selected subsets give rise to more variability in the estimated posterior mean, and higher posterior variance across all parameters, than minibatch ABC-SMC

with minibatches of the same size. We conclude that our approach is to be preferred over simply reducing the amount of available information by subsampling the data.

Our method contains the minibatch size, N_{bs}, and the number of generations, T, as tunable hyperparameters. In machine learning applications, it has been noted that inference with smaller batch sizes when optimizing an objective function leads to less overfitting to the data, as the algorithm is better able to observe the full range of parameter values that yield results that resemble the full distribution of the data, rather than the parameters that only produce summary statistics that are close to the average observed in the data set [56, 57]. At the same time, using a very small batch size might not provide a good estimate of the model distribution, meaning that careful consideration of the batch size is necessary. This will be the focus of Section 2.3.1. For the number of generations, we choose $T = 4$. This hyperparameter choice comes from performing test runs of minibatch ABC-SMC with various batch sizes and numbers of generations and computing the Kullback-Leibler (KL) divergence [70] between the successive posteriors as a measure of change of the distribution from one generation to another, as is done by Filippi *et al.* [52]. As a general stopping criterion, we suggest to choose the smallest T so that the KL divergences for the subsequent generations converge to a fixed value. In this specific problem, given the model and the data, this amounts to $T = 4$, indicating that the distributions do not change significantly after this generation. In other problems, a different number of generations will typically be needed. For details on how the KL divergence between successive generations was calculated, we refer the reader to Appendix A.2.3.

We note that the random sampling of minibatches ensures that with very high probability every datapoint in the dataset will be used in each generation. Since for each of the N_{gen} parameters in each generation there are N_{bs} datapoints sampled uniformly at random, the probability that one generation of minibatch ABC-SMC fails to contain a given datapoint, say $i_\star \in \{1, \ldots, N_{\text{obs}}\}$, can be computed. Denoting this event $E_{i_\star}$, a simple computation shows that this probability can be bounded by

$$\mathbb{P}(E_{i_\star}) = \left(1 - \frac{1}{N_{\text{obs}}}\right)^{N_{\text{gen}} \cdot N_{\text{bs}}} \leq e^{-(N_{\text{gen}} \cdot N_{\text{bs}})/N_{\text{obs}}}, \tag{2.10}$$

since $(1 + x/n)^n < e^x$ for any integer n. For $N_{\text{gen}} = 2 \cdot 10^4$,

$$\mathbb{P}(E_{i_\star}) \leq \exp\left(-\frac{2 \cdot N_{\text{bs}}}{N_{\text{obs}}} \cdot 10^4\right), \tag{2.11}$$

meaning that the probability that in one generation not all datapoints are used is essentially negligble for sensible choices of N_{bs} compared to N_{obs}.

All simulations were carried out using a development branch of the parallel open source ABC-SMC software Pakman [71]. Each ABC-SMC experiment was performed on a high

performance compute (HPC) cluster using five 48-core Cascade Lake (Intel Xeon Platinum 8268 CPU @ 2.90GHz) nodes.

2.2.3.3 Summary statistics

To compute the distance d between observed and simulated data, we employ the summary statistics proposed by Browning *et al.* [60], with the aim to capture three key features: population size increase, spatial structure and density profile [60]. To capture population size increase, we use the population size, $N(t)$, at $t = 24h$. To capture spatial structure, we use the pair correlation function, $P(r, t)$, which describes the density of pairs of cells separated by a distance r at time t, relative to the expected density of pairs if the population were uniformly distributed [60]. For our discrete data, the pair correlation function describes the relative number of pairs separated by distances in the range $(j-1)\Delta r < r < j\Delta r$. Following [60], we compute

$$P(j,t) = \frac{LH}{N(t)^2 \pi \Delta r(2j + \Delta r)} \sum_{n=1}^{N(t)} \sum_{i=1, i\neq n}^{N(t)} \mathbb{I}((j-1)\Delta r < \|x_i - x_n\| < j\Delta r). \tag{2.12}$$

In Equation (2.12), L and H are length and height, respectively, of the field of view. Following the recommendation in [60], we choose $\Delta r = 5\mu$m, and calculate the pair correlation function up to a distance of 100μm. To capture the density profile, we compute the density across cross-sections of the wound by dividing the field of view into 64 different bins in the direction parallel to the scratch and computing the cell density in each bin. Using 64 bins results in a bin width of approximately $24\mu m$, or, one cell diameter. Assigning a centroid for each cell, and using this centroid to define the location of cell i as y_i, results in a one-dimensional cell count $\rho(j, t)$ at time t, where

$$\rho(j,t) = \sum_{i=0}^{N(t)} \mathbb{I}(j\Delta y \leq y_i < (j+1)\Delta j). \tag{2.13}$$

These three distance functions allow us to define the distance between an observed data point y^i_{obs} and a simulated data point y^i_{sim} as

$$\begin{aligned} d(y^i_{\text{sim}}, y^i_{\text{obs}}) = \frac{[N_{\text{sim}}(24) - N_{\text{obs}}(24)]^2}{N_{\text{obs}}(24)^2} &+ \frac{\sum_{j=1}^{20}[P_{\text{sim}}(j,24) - P_{\text{obs}}(j,24)]^2}{\sum_{j=1}^{20} P_{\text{obs}}(j,24)^2} \\ &+ \frac{\sum_{j=1}^{64}[\rho_{\text{sim}}(j,24) - \rho_{\text{obs}}(j,24)]^2}{\sum_{j=1}^{64} \rho_{\text{obs}}(j,24)^2}. \end{aligned} \tag{2.14}$$

2.3 Results

In this section, we demonstrate that minibatch ABC-SMC is accurate and efficient at obtaining posterior distributions over model parameters given the high-throughput data from [43], and that it can be used to shed light on the mechanistic impact of a wide range of genetic perturbations. In Subsection 2.3.1, we explore how the choice of hyperparameters in the algorithm, as well as the quantity of available information, affects the reliability of the inference process so that we can confidently apply the minibatch ABC-SMC algorithm to data from the high-throughput screen. In Subsection 2.3.2, we show how the IBM can be used in tandem with minibatch ACB-SMC to detect functional differences between different gene knockdowns. In Subsection 2.3.3, we use minibatch ABC-SMC on the data from 118 gene knockdowns to locate each genetic perturbation in the five-dimensional parameter space of the IBM, according to its estimated posterior mean. Analysing the spatial structure of the locations of each gene knockdown in this space allows us to distinguish a range of functionally different subgroups within the data that represent genetic perturbations with functionally similar behaviours. Given the biological interpretation of the different model parameters, this allows us to reach conclusions about the primary drivers of wound closure in the different genetic perturbations. Importantly, our analysis shows the importance of density-dependent effects on observed changes in cell count number during the experiment.

2.3.1 Small batch sizes provide reliable inference

The introduction of minibatches in ABC-SMC comes with the task of selecting the batch size as an algorithm hyperparameter. A good choice for the batch size, N_{bs}, needs to balance computational complexity on the one hand and the generation of a reliable estimate of the discrepancy between model and data on the other hand. To understand the capacity of the minibatch ABC-SMC algorithm to distinguish different model parameter values from data in the context of the IBM, we generate two synthetic datasets, each corresponding to a different parameter regime. Parameter regime I corresponds to a high movement rate with a positive impact of density on movement and a low proliferation rate, such that $m = 1.5\text{h}^{-1}$, $p - d = 0.01\text{h}^{-1}$, $\gamma_m = -1\text{h}^{-1}$, $\gamma_p = 0.01\text{h}^{-1}$ and $\gamma_b = 20\mu\text{m}$. Parameter regime II corresponds to a low movement rate with contact inhibition for movement and a high proliferation rate, such that $m = 0.5\text{h}^{-1}$, $p - d = 0.025\text{h}^{-1}$, $\gamma_m = 1.5\text{h}^{-1}$, $\gamma_p = 0.01\text{h}^{-1}$ and $\gamma_b = 20\mu\text{m}$. For each parameter regime, an *in silico* data set is created by simulating one model output at $t = 24\text{h}$ for every initial condition contained in the Mock knockdown dataset. That is, each of the two *in silico* datasets contains 117 data points, each corresponding to one initial datum from the high-throughput experiment. Then, we perform minibatch-ABC-SMC on the synthetic datasets, varying the batch size,

such that $N_{\text{bs}} = 1, 5, 10, 50, 117$. For each batch size, we repeat the inference process five times.

2.3.1.1 Variability of the posterior distributions

In Appendix A.4, we report the posterior means of the resulting posterior distributions, and their variance, at each batch size, as well as representative plots of the posterior distributions. At any value of the batch size, minibatch ABC-SMC correctly identifies the features of each parameter regime. For instance, for the dataset for parameter regime I, the estimated intrinsic motility rate, m, is consistently estimated as higher than that of the dataset for parameter regime II. Conversely, the net intrinsic proliferation rate, $p - d$, is correctly estimated as higher in parameter regime II than in parameter regime I. For the density-dependent interaction term, γ_m, the minibatch ABC-SMC implementation discerned at any batch size that the effect of crowding on motility is positive in regime I, and negative in regime II. Increasing the batch size improved the quality of the posterior distributions, although the improvements diminish rapidly as the batch size, N_{bs}, is increased. There is a critical value – Figure A.2 shows this is typically around $N_{\text{bs}} = 10$ – after which increasing the batch size, N_{bs}, does not futher reduce the variability of the estimated posterior means. Similarly, after this critical value, the capacity of minibatch ABC-SMC to identify the true posterior mean does not improve, even when the batch size is increased to the full data size. Similarly, the standard deviation of the posterior distributions diminishes rapidly when the batch size is increased initially, but then stays relatively constant across larger batch sizes. In both parameter regimes, the standard deviation does not decrease noticeably with increasing batch size past a batch size of $N_{\text{bs}} = 10$. Taken together, analysis of the posterior means and posterior standard deviations suggests that a minibatch implementation of ABC-SMC provides a robust approximation of the posterior distribution at a fraction of the computational cost. In the synthetic data set, the computational speedup of using a batch size of 10 samples compared to the full data set amounts to a factor $117/10 = 11.7$.

2.3.1.2 Reliability of minibatch ABC-SMC using experimental data

Finally, to compare performance on real-life data from the scratch assay, we perform inference on the Mock type data set in Appendix A.4.3, where we vary the batch size from 1 to 117 and find no meaningful differences between the different batch sizes in terms of their ability to identify the posterior means consistently. With growing batch size, just as with synthetic data, the variance of the posterior distribution decreases and shows a similar behaviour to the synthetic control data around a batch size of $N_{\text{bs}} = 10$. We conclude that choosing a minibatch size of $N_{\text{bs}} = 10$ allows for minibatch ABC-SMC to reliably identify the posterior mean.

While some genetic perturbations have a large number of observations and the choice of $N_{\text{bs}} = 10$ offers a good balance between computational complexity, robust inference of the posterior mean and low variance of the posterior distribution, some perturbations have very few (as few as two) observations in the high-throughput screen used in this work. As such we address the suitability of ABC-SMC in identifying mechanistically relevant differences between the different gene knockdowns by selecting 40 datapoints from each of the Mock and CDH5 knockdown experiments at random and without replacement, and dividing them into 20 pairs of two observations for each of the datasets. We perform ABC-SMC on each of the resulting *miniature data sets* and record their posterior distributions. We find that while the spread of the posteriors is indeed larger for each of the experimental conditions than in the case when more datasets are included, the posterior means can still identify meaningful differences between the two different genetic perturbations. This suggests that as few as just two observations can be used to confidently estimate model parameters.

Finally, we assess how well our IBM predicts experimentally observed wound healing behaviour when parametrised according to the posteriors generated using minibatch ABC-SMC. We sample 1000 parameters from the posterior distribution obtained using minibatch ABC-SMC with $N_{\text{bs}} = 10$ on the full Mock dataset and select one Mock initial condition uniformly at random. For this condition, we perform forward model simulations, one with each sampled parameter, and record the summary statistics defined in Section 2.3.3, which are the density profile, pair correlation function, and number of detected cells at $t = 24$h. From these forward simulations, we compute the mean and standard deviation for the density profile and pair correlation function. In Figure 3, we display the experimental data for density profile and pair correlation together with a range of one standard interval from the posterior mean. We also show a histogram of simulated cell counts at $t = 24$h, compared to the experimentally observed cell count at $t = 24$h. The posterior predictive ranges for the density profile and pair correlation are excellent fits to the data, and the simulated cell count histogram is centered around the experimentally observed cell count.

2.3.2 Detecting differences between different experimental conditions

Our data set contains three genetic perturbations with a large number of observations. These are: Mock, CDC42 knockdown and CDH5 knockdown. For each of the observations, we apply the minibatch ABC-SMC algorithm using the implementation details from Section 2.2.3. We use the findings from Section 2.3.1 to choose a batch size $N_{\text{bs}} = 10$ for all experiments. We note that the total number of observations for Mock, CDC42 and CDH5 knockdowns are 117, 52 and 31, respectively, meaning that in the inference process

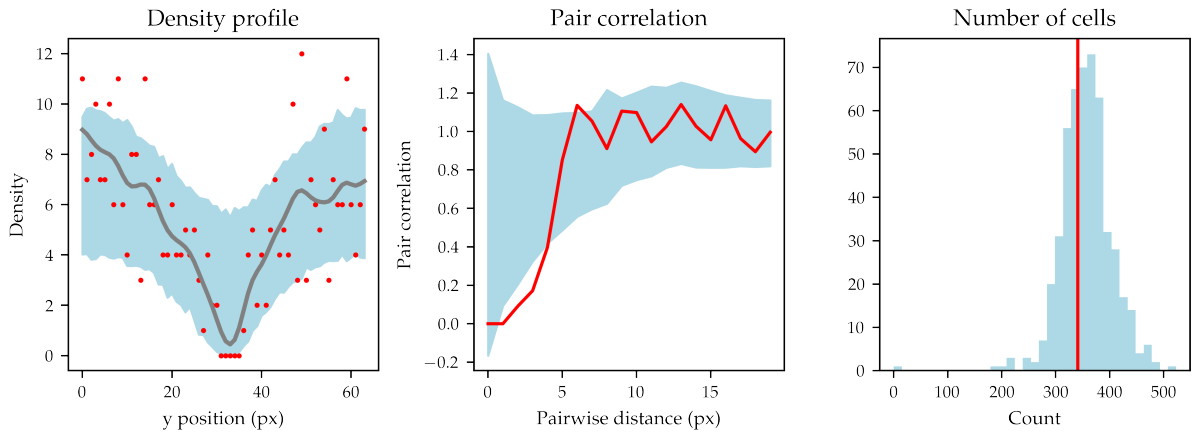


Figure 2.3: Posterior predictive check for the model. Left: experimentally observed density profile for a single experimental replicate sampled at random from the Mock dataset (red points) displayed against the posterior range of one standard deviation away from the posterior mean (blue shaded region). The gray line is the smoothed density profile generated using a Gaussian kernel. Middle: experimental pair correlation function (red line) displayed against the posterior range of one standard deviation away from the posterior mean (blue shaded region). Right: experimentally observed cell count at $t = 24$h (red vertical line) compared to the histogram of model simulations of cell count (blue). All summary statistics demonstrate an excellent fit to the experimental data.

for these experiments using $N_{\text{bs}} = 10$, minibatch ABC-SMC yields a speedup of roughly $12\times, 5\times, 3\times$, respectively.

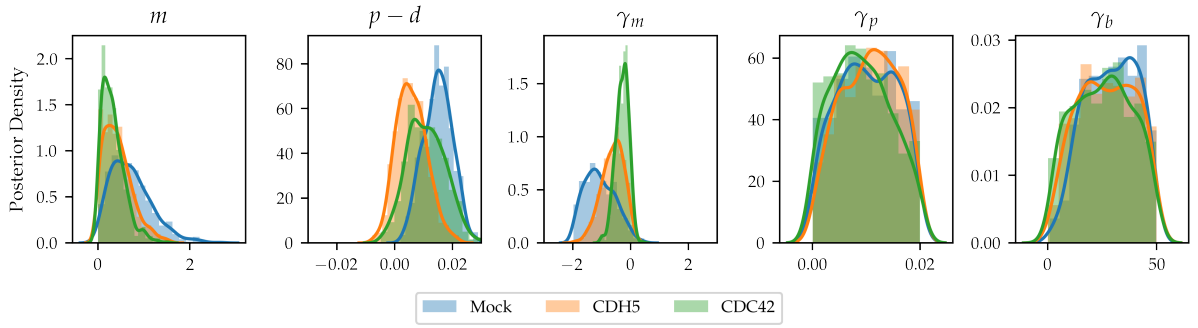


Figure 2.4: Posterior distributions for three genetic perturbations with large numbers of observations. Blue: Mock condition, orange: CDH5 knockdown, green: CDC42 knockdown.

Figure 2.4 reveals a striking difference between the three genetic profiles in terms of their intrinsic movement, intrinsic net proliferation and motility parameters. In Appendix A.7, we calculate confidence intervals for the posterior means of each parameter and find that the confidence intervals are disjoint for the intrinsic motility rate, m, net intrinsic proliferation rate, $p - d$, and density-dependent motility rate, γ_m, while only CDH5 has a confidence interval disjoint from CDC42 and Mock, respectively, in the density-dependent parameters γ_p and γ_b. Of these genetic profiles, Mock knockdown is shown to have behaviour best described by strong contact-mediated motility, while motility is much less impacted by

local cell density in the CDH5 knockdown case, and this effect is much smaller still in the CDC42 knockdown case. Further, Mock knockdown has the highest intrinsic net proliferation rate, whereas the CDH5 and CDC42 knockdowns are inferred to have lower net proliferation rates. In fact, the posterior distributions for the CDH5 and CDC42 knockdowns have support on negative values of $p-d$, which is consistent with the fact that some of the CDH5 and CDC42 knockdown assays show a decrease in the total number of cells over the course of the experiment. In line with previous inference performed on this mechanistic model of density-dependent cell movement and proliferation, the parameters controlling density-dependent proliferation, γ_p, and the strength of density-dependent interactions, γ_b, are not conclusively identified from the data [60].

Our findings are in agreement with previous experiments. For example, loss of CDC42 expression is associated with defective adhesion, wound healing, polarity establishment, and migration (see, for instance, the review by Mendelez *et al.* [72]), which is consistent with our finding that the CDC42 knockdown has weaker density-dependent migration than Mock: a loss of polarity establishment and adhesion is consistent with less directed movement into the wound and away from areas of high local tissue density. Likewise, CDH5 plays "a[n] important role in endothelial cell biology through control of the cohesion and organization of the intercellular junctions" [73], which is again consistent with a loss of local density-dependent mechanisms.

In addition to comparing *between* the different knockdowns in terms of their cellular migratory dynamics, the large number of observations per knockdown in a high-throughput screen also enables inquiry into the factors that might create variation *within* any given genetic perturbation. For instance, the data for Mock and CDH5 knockdown contain very large variations in initial wound size, meaning that an important question that can be explored is the functional effect of initial wound size on migratory dynamics. It is known from the literature that changing the initial density of cells in a scratch assay has far-reaching consequences for the resulting dynamics [60]. In the same way, Jin *et al.* found that the initial geometry in experimental models of wound closing plays an important role in the wound closing dynamics [74], raising the possibility that the size of the wound could also have a strong effect on the wound-closure dynamics. To address this question, we measured the initial wound size for both the Mock and CDH5 knockdown datasets.[1] We find that initial wound sizes in the Mock dataset range between $\sim 1.79\text{mm}^2$ and $\sim 3.06\text{mm}^2$, and between $\sim 1.89\text{mm}^2$ and $\sim 2.7\text{mm}^2$ in the CDH5 knockdown dataset. We choose to divide the Mock data set into three different wound size categories: small ($< 2.43\text{mm}^2$), medium ($2.43\text{mm}^2 - 2.70\text{mm}^2$), and large ($> 2.70\text{mm}^2$). Similarly, we divide the CDH5 data set in two categories: small ($< 2.25\text{mm}^2$) and large ($> 2.25\text{mm}^2$). For each of the data subsets – large, medium, and small Mock; large and small CDH5 knockdown,

[1]The CDC42 knockdown experiments had a much smaller variation in initial wound size, making it impossible to differentiate meaningfully between experiments based on initial wound size.

respectively – we perform minibatch ABC-SMC with a batch size $N_{\text{bs}} = 10$ to identify the migratory dynamics within these conditions.

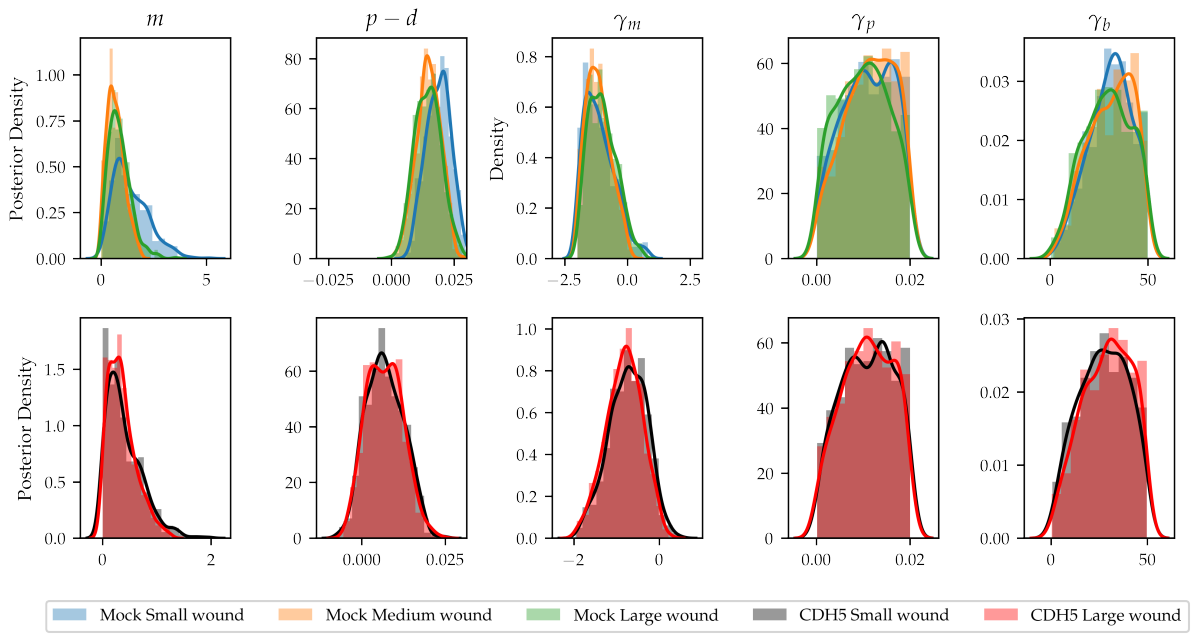


Figure 2.5: Posterior distributions for model parameters when inference is carried out on subsets of the data based on wound size. Top row: Mock experiments, three different wound size categories are identified – small (blue, $< 2.43\text{mm}^2$), middle (orange, $2.43\text{mm}^2 - 2.70\text{mm}^2$), and large (green, $> 2.70\text{mm}^2$). Bottom row: CDH5 knockdown, two different wound size categories are identified – small (black, $< 2.25\text{mm}^2$) and large (red, $> 2.25\text{mm}^2$).

Figure 2.5 shows the marginal posterior distributions for model parameters for both the Mock and CDH5 knockdown datasets. To understand the large variance seen in Figure 2.5 for the intrinsic motility parameter, m, in the small wound category, we further analysed the wound closure data and observed that the smallest wounds at time $t = 0$h in the Mock case were fully closed by $t = 24$h. Further, when this is the case, the intrinsic motility parameter becomes practically unidentifiable: by simulating new model outputs using the IBM, we found that for any fixed value of the net intrinsic proliferation rate, $p - d$, and density-dependent movement parameter, γ_m, higher intrinsic movement rates did not influence the density or pair correlation summary statistics detailed in Section 2.2.3.3. Therefore, the higher variance of the posterior distribution in the small wound size category for the Mock data set is likely attributable to parameter non-identifiability given the available data, rather than to biologically significant mechanistic signatures in the data. We quantify the (dissimilarity) between the resulting marginal posterior distributions by computing confidence intervals for the posterior means in A.7. We find that none of the confidence intervals, except for that of the intrinsic motility rate, m, in Mock are disjoint from those of the other initial wound size categories. This implies that migratory dynamics for the different initial wound sizes are very similar.

2.3.3 High-throughput identification of mechanistic effects of gene knockdowns

Having established that minibatch ABC-SMC can accurately identify different mechanistic cell behaviours, we employ ABC-SMC on a large number of gene knockdowns in the high-throughput screen in order to assess and identify the range of different behaviours observed in the set of knockdown experiments. For each gene knockdown, a posterior distribution is obtained and each genetic perturbation is identified with a point in five-dimensional parameter space, given by the value of its posterior mean. We reiterate that there is a large variability in the number of available data points for each knockdown. Here, we consider all gene knockdowns for which more than two data points exist in the data set, and for all gene knockdowns we choose $N_{\text{bs}} = 2$ accordingly.

2.3.3.1 Clustering analysis reveals functional subgroups

The set of posterior means for the different gene knockdowns is a set of points in five-dimensional parameter space, and it can be studied to reveal patterns of cell behaviours across the different gene knockdown experiments. One popular method of understanding the spatial structure of data, when it cannot be readily visualised due to the dimension of the data space, is K-means clustering. The idea behind K-means clustering is to divide a set of points in Euclidean space into K *clusters* such that the sum of the pairwise distances within each cluster is minimised. In other words, it is an unsupervised method to assign labels to the datapoints such that points with the same label are close to each other given some defined distance function. Mathematically, given K clusters, the classification problem is to subdivide the data into K distinct clusters, $\mathcal{S} = \{S_1, \ldots, S_K\}$, where the clusters are found by solving the optimisation problem

$$\mathcal{S} = \arg\min \sum_{i=1}^{k} \frac{1}{2|S_i|} \sum_{\mathbf{x},\mathbf{y} \in S_i} \|\mathbf{x} - \mathbf{y}\|^2, \tag{2.15}$$

where $|S_i|$ denotes the number of points in cluster i, and $\| \bullet \|$ denotes the Euclidian distance. The number of clusters, K, is a hyperparameter that needs to be carefully chosen. Increasing the number of clusters will always decrease the total sum of pairwise distances, however it also leads to a loss of interpretability of the cluster labels and potentially overfits the data. A common practice in data science is to use the *elbow method* [75], which was first employed by Thorndike [76]. The elbow method amounts to choosing K as the elbow of the curve of the minimum of the objective function in Equation (2.15) as K is varied. The interpretation of picking the elbow of the curve, in clustering, corresponds to choosing K such that adding futher clusters does not provide a significantly better fit to

the data. We perform K-means clustering for $K = 1, \ldots, 10$, and find that the marginal improvement of the sum of pairwise distances rapidly decreases after $K = 3$ (see Appendix A.13). Therefore, we perform clustering with three clusters. To further analyse robustness, we assess the difference between clusters when we set $K = 2, 4, 5$ to find that the addition of extra clusters yields further sub-clustering within the clusters found at $K = 2$. From this we conclude that $K = 3$ is an good choice to balance mechanistic insight with more detailed spatial information.

The pairwise marginal distributions of the posterior means for the different gene knockdowns is shown in Figure 2.6, where the colour key is defined by the cluster labels. The separation between the clusters is best seen in the pairwise martinal posterior distribution of the intrinsic motility parameter, m, and the density-dependent motility parameter, γ_m. The marginal posterior distributions of the two parameters show that the cluster labels correspond to changes in the sign of the density-dependent motility parameter, γ_m, and the magnitude of the intrinsic motility parameter, m. Interpreting this separation mechanistically allows us to identify three functionally distinct clusters in the data. We identify Cluster I as knockdowns with low intrinsic motility, with the motility enhanced by high local cell density (negative values for γ_m). In Cluster II, knockdowns have low intrinsic motility, and motility is inhibited by high local cell density (positive values for γ_m). In Cluster III, knockdowns have high intrinsic motility, and motility is enhanced by high local cell density (positive values for γ_m).

Furthermore, the pairwise plot of the intrinsic net proliferation parameter, $p - d$, and the density-dependent motility parameter, γ_p, shows that for knockdowns with a value of $p - d$ close to zero, there is a wide range of possible values for the density-dependent proliferation parameter, γ_p, whereas the variability of γ_p diminishes for $p - d$ further away from zero. As such, the mechanistic interpretation of this finding is that high intrinsic proliferation rates are associated with lower density-dependent inhibition of proliferation. Figure 2.6 also suggests that there is a correlation between a higher increased intrinsic motility, m, and the bias parameter, γ_b: knockdowns in Cluster III also show the highest values of γ_b. This correlation is much less pronounced when the intrinsic net proliferation rate is compared to the density-dependent crowding term, γ_b. In summary, analysis of the posterior means suggests that high intrinsic motility, but not high intrinsic net proliferation, is associated with stronger sensing of local cell density gradients through the bias parameter, γ_b. Finally, the pairwise plots show that most of the functional differences between the knockdowns can be described in terms of the intrinsic motility rate, m, the net proliferation rate, $p - d$, and the density-dependent motility rate, γ_m. In the remainder of this section, we explore the impact of these parameters on wound healing outcomes by analysing the observed cell count and wound area changes.

Figure 2.7 illustrates how the clusters can help tease apart the different mechanisms at play in wound healing for each of the gene knockouts. For each of the clusters, we

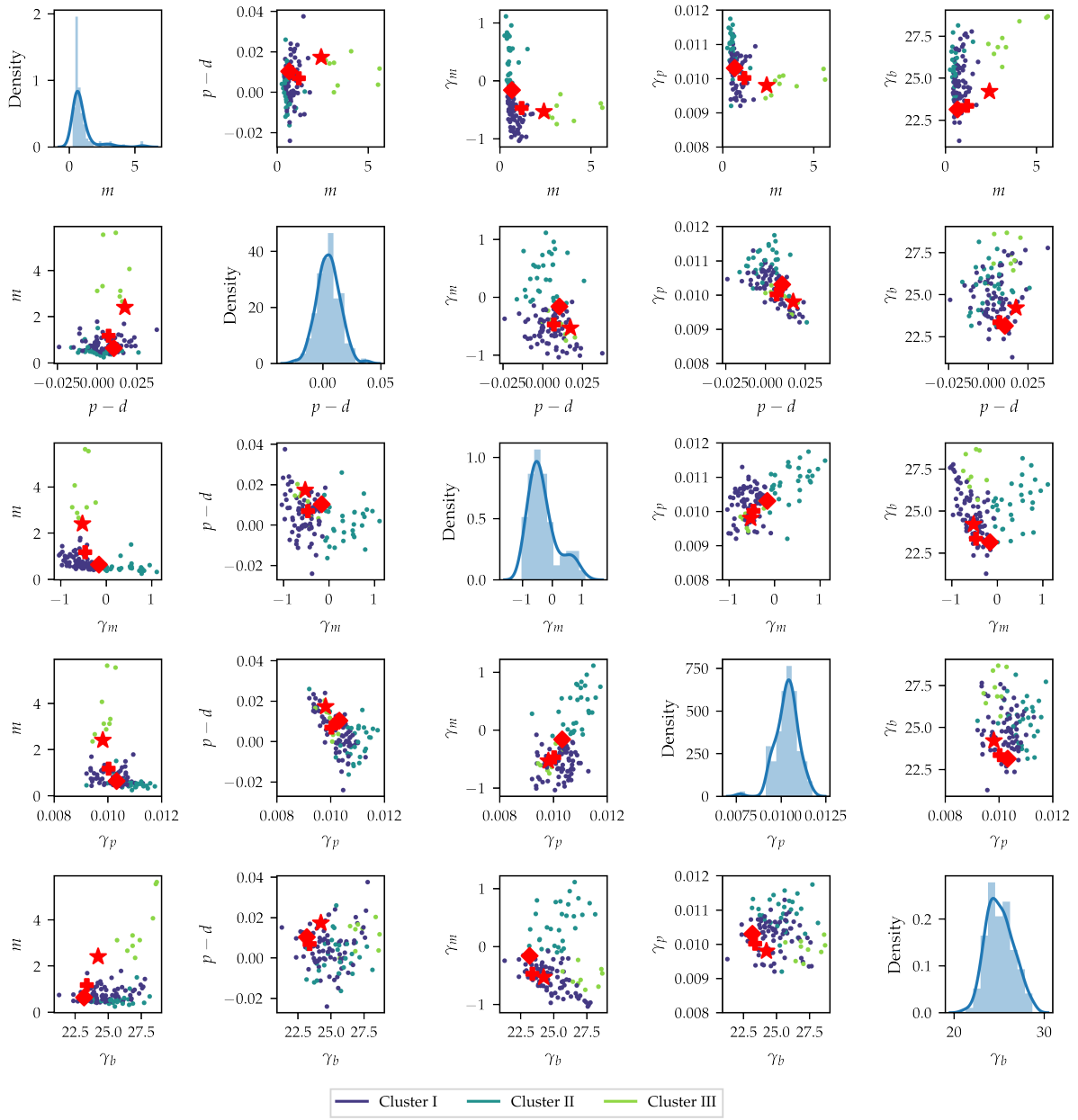


Figure 2.6: Pairwise distributions of the posterior means for each of the knockdown experiments in the high-throughput screen. Diagonal plots show the distribution of the posterior means of each of the parameters, while off-diagonal plots show the pairwise distributions of the different knockdown posterior means. The colour key denotes the different clusters found using K-means clustering with $K = 3$. Star denotes Mock, diamond CDC42 knockdown and plus CDH5 knockdown.

select the knockouts with greatest dissimilarity with the other clusters. These are EPHA7, USP18 and ITPR1 for Clusters I, II, and III, respectively. While the EPHA7 and USP18 knockdowns have similar degrees of wound closure, the posterior means for the EPHA7 knockdown suggest that incomplete wound closure is mainly due to net cell death during the experiment, which is consistent with previous studies [77]. The high value for γ_m for USP18 knockdown suggests that cells are unable to move into the wound area due to

defects in cell adhesion. This indicates that USP18 might be required for the turnover of cell adhesion through deubiquitination. Moreover, ITPR1 is known to regulate apoptosis and our results confirm its dual role on cell motility and proliferation [78].

In summary, Bayesian inference with a large screen of gene knockdowns can be used to characterise the functional role of the genes under investigation given a sufficiently detailed mathematical model. Given such a characterisation (in our work, this is the division into functional clusters), experimentally testable hypotheses can be generated to further refine and explore our understanding of the genes under consideration.

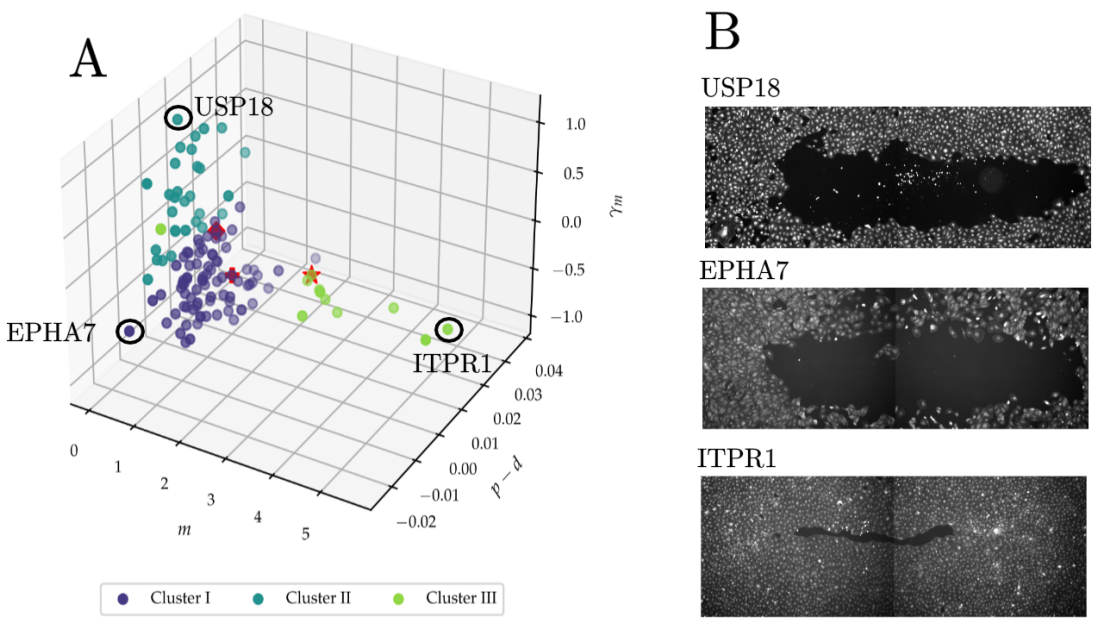


Figure 2.7: Identifying functional subgroups among perturbations. A: Three-dimensional scatter plot of the posterior means for the intrinsic motility parameter, m, intrinsic net proliferation parameter, $p - d$, and density-dependent motility parameter,γ_m, for each of the samples in the high-throughput screen. Star denotes Mock, diamond denotes CDC42 knockdown and plus denotes CDH5 knockdown. B, top to bottom: wound mask for one final condition of the USP18, EPHA7, and ITPR1 knockdowns, respectively.

2.3.3.2 Density-dependent patterns of proliferation reproduce experimental patterns of cell count changes

Changes in cell counts are a simple summary statistic to extract from scratch assay data [60, 62, 63] and provide an indication of the average net proliferation rate in a large population of cells, such as the cell monolayer used for scratch assays. However, computing an average net proliferation rate does not reflect more complex mechanisms, such as contact inhibition, that may influence the change in cell numbers throughout the experiment. In

this section, we turn to the question of the extent to which density-dependent effects are observed in the high-throughput measurements in cell count changes for each of the knockdowns. Let c_0, c_{24} denote the cell counts at $t = 0$h, and $t = 24$h for a given knockdown. The fold change in cell count, $\mathcal{C}$, is given by $\mathcal{C} = c_{24}/c_0$. By the definition of the IBM, cell count changes occur directly as a function of the proliferation parameters p, d, and γ_p, and indirectly through the parameters m, γ_m, and γ_b, since these will affect local cell density. In fact, Equation (2.4) relates the local proliferation rate to the intrinsic and density-dependent proliferation rates and hence relates the expected number of cells to the values of the different model parameters. In the case where density-dependent effects are ignored, *i.e.* $\gamma_p = 0$, one would expect that the expected value of the cell fold change, $\mathcal{C}$, is given by $\mathcal{C} = \exp(24(p - d))$, since $p - d$ is the net proliferation rate measured in h^{-1}. Figure 2.8 shows that this curve fails to describe the observed fold change in cell numbers across the screen, indicating that density-dependent effects affect fold changes in observed cell counts. In this work, density-dependent effects affect the proliferation rate through the sum in Equation (2.4). This sum depends on the cell locations at each simulation step, meaning that it cannot be determined from the available data. However, by assuming local cell density is constant across the assay, and by approximating local cell density as the average of the local cell densities at $t = 0h$ and $t = 24h$, this difficulty can be overcome. Letting F be the area of the field of view, and w_0, w_{24} the wound area at $t = 0$h, and $t = 24$h, respectively, one can approximate the sum in Equation (2.4) as follows. Assuming constant density, ρ_0, one can estimate that

$$\gamma_p \sum_{i=1, i\neq n}^{N(t)} \exp\left(-\frac{\|\mathbf{x}_n - \mathbf{x}_i\|^2}{2\sigma^2}\right) \approx 2\pi\rho_0 \int_0^\infty \exp(-r^2/2\sigma^2)\mathrm{d}r = \gamma_p \sigma\sqrt{2\pi}\rho_0. \tag{2.16}$$

By now estimating the average density, ρ_0, as the average of the density at $t = 0$h and $t = 24$h, *i.e.*,

$$\rho_0 = \frac{1}{2}\left(\frac{c_0}{F - w_0} + \frac{c_{24}}{F - w_{24}}\right), \tag{2.17}$$

we obtain the approximation

$$\gamma_p \sum_{i=1, i\neq n}^{N(t)} \exp\left(-\frac{\|\mathbf{x}_n - \mathbf{x}_i\|^2}{2\sigma^2}\right) \approx \sigma\gamma_p\sqrt{\pi/2}\left(\frac{c_0}{F - w_0} + \frac{c_{24}}{F - w_{24}}\right). \tag{2.18}$$

Under this approximation, a coarse-grained estimate for the fold change in cell numbers over the 24-hour experiment is given by

$$\mathcal{C} \approx \exp\left(24\left(p - d - \sigma\gamma_p\sqrt{\pi/2}\left(\frac{c_0}{F - w_0} + \frac{c_{24}}{F - w_{24}}\right)\right)\right). \tag{2.19}$$

For each sample in the high-throughput screen, we compute the value of $\mathcal{C}$ through Equation (2.19) and fit an exponential curve through the obtained fold estimates. In Figure 2.8 we show that the density-corrected cell fold change estimate accurately predicts the observed cell fold change as a function of the intrinsic motility parameter. We conclude that density-dependent effects contribute to the observed cell fold change by modulating the intrinsic net proliferation rate and that the IBM can shed light on the nature of density-dependent effects for each of the knockdowns.

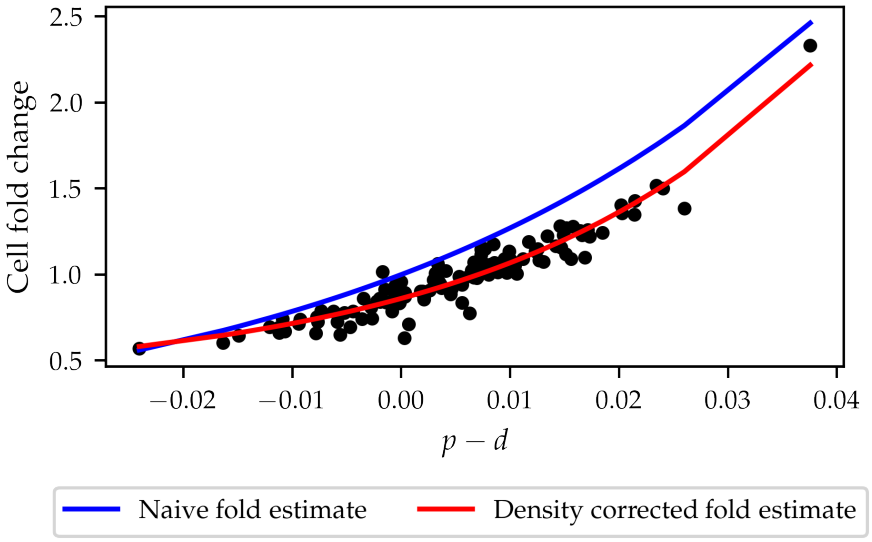


Figure 2.8: Observed cell fold change as a function of intrinsic net proliferation rate, $p - d$. When density-dependent effects are ignored (blue) line, the estimate of cell fold change as a function of net intrinsic proliferation overestimates cell fold change. When density-dependent effects are considered, model predictions accurately capture cell fold change.

2.3.3.3 Finding mechanistic contributions to wound closure

While the fold change in cell counts is associated with the proliferation parameters, the fold change in wound size depends delicately and nonlinearly on all model parameters. We define the wound area fold change, $\mathcal{W}$, as w_{24}/w_0. Note that $\mathcal{W} = 1$ means that the wounds at $t = 0$h, and $t = 24$h are equally large, and $\mathcal{W} = 0$ means the wound is fully closed. In this section we wish to understand the different roles of the various mechanistic terms in the model and explain the range of observed wound closure rates. In Figure 2.9, the observed wound area fold change is plotted against the observed cell fold change. While wound closure is not completely independent of increases in cell numbers, the data suggest that the dependence of wound closure on cell number increases is not as strong as might be expected if wound closure is thought to be primarily driven by proliferation: values of wound area fold change of approximately 0.75 are observed across knockdowns with 50%

cell death all the way to 25% increases in cell counts, *i.e.* for net positive and net negative cell number changes, there can be similar amounts of wound closure. To investigate which

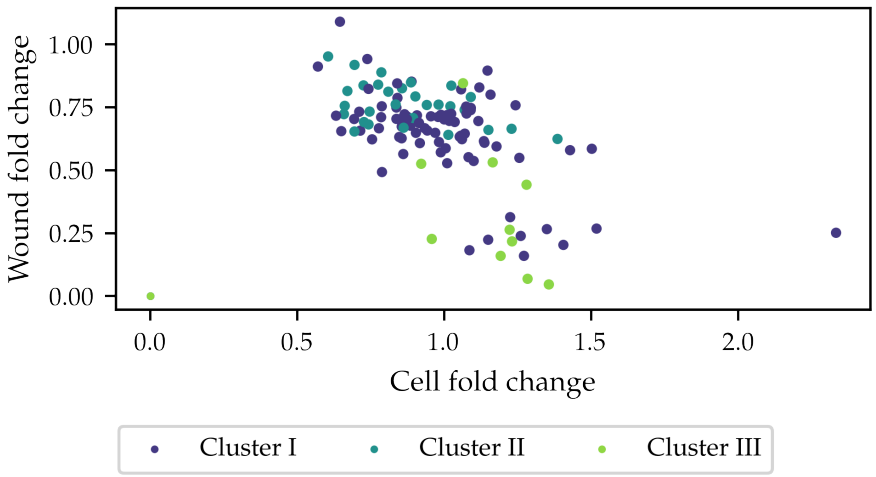


Figure 2.9: Observed wound area fold change as a function of cell fold change. Across the range of cell fold changes, starkly different rates of wound closure are observed.

mechanistic features of the model are key in wound closure, Figure 2.10 shows the observed wound area fold change, plotted against the different model parameters. Surprisingly, high intrinsic motility rates do not necessarily correlate with enhanced wound closure: the datapoints with labels corresponding to high intrinsic motility and density-enhanced motility all have much higher intrinsic motility posterior means than Mock, but their wound closure rates are not, in general, higher than Mock. At any given value of the intrinsic motility, m, a wide range of different wound closure extents are observed. In contrast, a stronger coupling is observed between the extent of wound closure and the density-dependent terms: when motility is strongly enhanced by local crowding, *i.e.* large, negative values of γ_m, very low values of wound area fold change (*i.e.*, significant wound closure) are observed, whereas the wound area fold change is generally high when γ_m is high. Likewise, when γ_p is increased, wound area fold changes increase as well. These findings suggest that, rather than the intrinsic rates of movement and proliferation, it is the density-dependent interactions between cells that determine the degree to which wound healing is accomplished.

While Figure 2.10 provides information as to how the different model parameters influence wound closure, it is to be expected that there is an interplay between the estimated intrinsic rates and their density-dependent counterparts. To investigate the extent to which different model parameters combine to give rise to wound closure, we investigate the dependence of wound closure on pairs of parameters. In Figure 2.11, wound area fold change is plotted as a function of the parameter pairs $(m, p-d)$, $(p-d, \gamma_p)$, (m, γ_m).

Figure 2.11 suggests that high intrinsic motility rates can counteract low intrinsic proliferation rates to close the wound rapidly, but not vice versa. Moreover, the density-

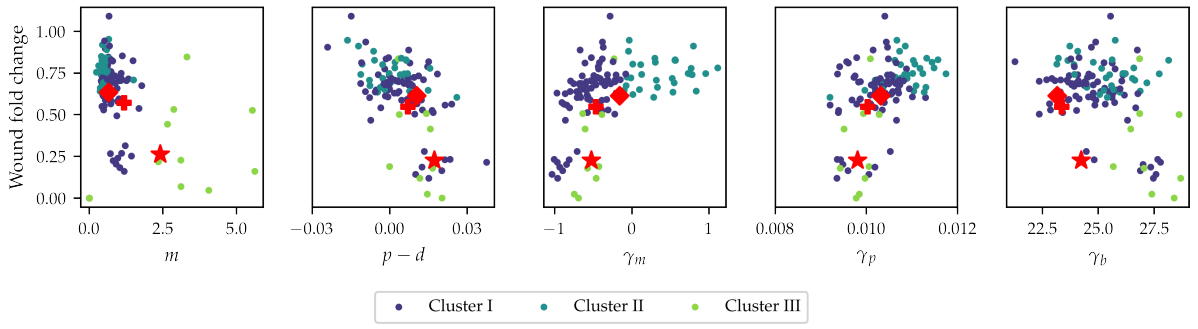


Figure 2.10: Observed wound area fold change plotted against the different model parameters. The colour key denotes the different clusters found using K-means clustering with $K = 3$. Star denotes Mock, diamond denotes CDC42 knockdown and plus denotes CDH5 knockdown.

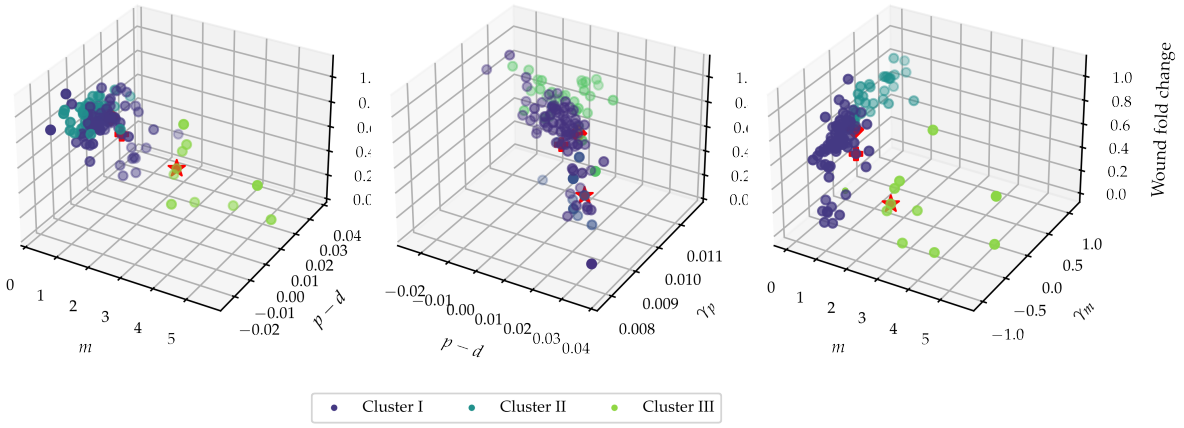


Figure 2.11: Observed wound area fold change plotted as a function of combinations of posterior means. Each three-dimensional plot shows the multivariate dependence of wound area fold change as a function of various parameters. The colour key denotes the different clusters found using K-means clustering with $K = 3$. Star denotes Mock, diamond denotes CDC42 knockdown and plus denotes CDH5 knockdown.

dependent mechanistic parameters have a profound impact upon wound area fold change. At the same time, across the range of values for the intrinsic net proliferation rate, $p - d$, the extent of density-dependent proliferation, γ_p, influences wound healing by increasing the extent of wound closure at similar levels of $p - d$.

2.4 Discussion and outlook

The aim of this work was to develop and showcase a method to address the task of performing Bayesian inference on a large dataset, where the likelihood is obtained through simulation of a detailed mathematical model and the nature of the data requires repeated simulation. In essence, our method entails estimating the likelihood by repeatedly drawing

subsamples from the data and estimating the likelihood function through this sampling scheme. The framework was developed in the context of a high-throughput scratch assay experiment, using a mathematical model with density-dependent interactions to describe the behaviours of the individual cells in the experiment, but the ethos of the approach is sufficiently general so as to apply to other large datasets and other computationally costly likelihood estimation methodologies.

The motivation for developing this framework stems from the fact that traditional high-throughput experiments often use simple summary statistics to understand and explain observed differences in the experiment. While these can lead to valuable information, a detailed mathematical model can provide further insights that are not captured by simple summary statistics. For example, this work focused on extracting detailed density-dependent interactions between cells from only considering spatial statistics in very sparse data. In this context, the identification of detailed density-dependent interactions allowed us to identify subgroups of genetic perturbations that have functionally different mechanistic behaviours. These differences help to explain and understand the different degrees of cell count changes and wound closure rates, which are summary statistics that are routinely collected from high-throughput experiments. Our findings identify different mechanistic behaviours that are consistent with some previous experimental measurements of wound healing after gene knockdowns. At the same time, the scale on which we are able to perform simulations opens up the possibility of characterising in an automated way the functional role played by a large amount of gene knockdowns in the future. By using different, possibly more detailed, mathematical models that are better tailored to a specific application, we conjecture that efficient Bayesian inference on a high-throughput experiment can be used to uncover previously undiscovered mechanisms – in the context of cell migration, or other applications.

We would like to highlight the importance of using a mathematical model that can faithfully identify and isolate mechanistic effects in the presence of observation noise or other factors. For instance, we noted in the process of carrying out the work that unless careful consideration was given to incorporating measured cell sizes in the assay, the model was unable to confidently identify cell-cell interactions, as the model components involved in estimating local cell density depend heavily on an accurate representation of cell size. Hence, by carefully extracting these detailed statistics, we can isolate the effect of cell-cell interactions and avoid the confounding factor that cell-cell interactions naturally occur over different spatial scales as a result of possibly different cell sizes. By probing how further mechanistic behaviours of the gene knockdowns affect these summary statistics, we conclude that density-dependent effects play a crucial role in wound healing.

There are a number of ways in which our method can be further improved going forward. Firstly, in this work we have used a simple method to choose the batch size, which amounts to analysing the systematic error and variability of the estimated parameter

posterior distributions when different batch sizes are used, and choosing a value of the batch size where the variability plateaus while not incurring systematic errors in estimating model parameters. We remark that this amounts to a grid search through the batch size hyperparameter. Performing a grid search through algorithm hyperparameters prior to using the method on experimental data is a common approach in computational science to choose hyperparameters. In this case, the method yields additionally a rigorous uncertainty quantification of model predictions. The potential shortcoming of performing the hyperparameter grid search is that the computational cost of the grid search using different batch sizes might undo the computational gains from performing minibatch ABC-SMC. However, in many applications there may exist a need to perform ABC-SBC a great number of times. For example, in this work, there are 118 datasets, each of which requires one ABC-SMC run. If, in each such ABC-SMC run, the computational time is substantially reduced (and we note that the speed-up reported for *e.g.* the Mock dataset was a factor 11), the computational savings can still greatly exceed the extra cost of the hyperparameter grid search. At present, our method offers a computational improvement in such a setting. In future work, a different, perhaps more computationally efficient, method for choosing the minibatch size might be proposed, which would greatly extend and improve the applicability of our work.

More generally, the capacity of any ABC implementation to infer parameter values from data will depend on the specific computational model and the data available (for instance, through its spatial and temporal resolution, or its signal to noise ratio). For this work, data availability for several of the gene knockdowns was limited, and our work demonstrated that increasing the number of data points for these knockdowns would provide less variable estimates of model parameters, without incurring extra computational cost (by choosing a small batch size, the computational cost of the likelihood estimation does not increase, while the total variance does decrease). At the same time, a different choice of summary statistics might aid in the identification of some of the model parameters for which the posterior density is uninformative. The ability to do so will depend again on data availability, as well as a further understanding of the specific computational model (practical non-identifiability was also observed by Browning *et al.* [60] for the same parameters even though temporal resolution of the data was finer). These points notwithstanding, uncertainty quantification demonstrated that, even with modest amounts of data, minibatch ABC-SMC can reliably identify functional differences between different gene knockdowns. This opens the way to using ABC-SMC in understanding and interpreting the outcomes of high-throughput experiments as well as using ABC-SMC in such contexts to generate hypotheses regarding potential biological mechanisms that can be experimentally tested.

Chapter 3

Cellular cruise control: mechanical power acts as a regulator of collective migration in epithelia

The work in this chapter has arisen from a collaboration with Isaac Breinyn and Dr Daniel J. Cohen (Princeton University). All experimental data were collected by Isaac Breinyn. My contributions to the work were the development of the continuum mechanics modelling, all computational implementations, and the writing of the results in this chapter.

3.1 Introduction

In his 1859 'Cellular-Pathologie', Rudolf Virchow coined the term *Zellenstaat*, or 'Cell State', to describe tissues as 'a society of cells, a tiny well-ordered state' [79]. Indeed, modern studies have shown that cellular collectives, both *in vivo* and *in vitro*, are highly structured, with domains of highly correlated collective motion, complicated leader-follower dynamics, and self-regulatory mechanisms that have both bio-molecular and physical components [80]. Of late, the role of mechanics during collective migration has received evermore attention in the scholarship [81–85], demonstrating that tightly coordinated patterns of force production in a wide range of tissues are both the cause and the consequence of morphogenetic and migratory events [5, 80, 86–88]. Although there have been massive strides towards developing robust, relevant and tractable laboratory setups for studying forces in epithelial migration, there is a distinct lack of interpretation of the roles of those forces in guiding and regulating collective cellular migration. A problem that is particularly poorly understood is how the constituent cells in a tissue

use the propagation of mechanical forces, and stresses, to inform their own contribution to force production and collective migration. In the parlance of the *Zellenstaat*, cells must contribute to collective migration by exerting energetically costly active forces, but the question that arises is: how is this energetic burden shared between cells? The key questions that we will explore in this chapter, and which have gone largely unstudied in the scholarship, are thus: do all cells use the same share of their energy budgets to produce active forces; are there some cells that contribute more to the collective motion than others; and how is the amount of energy used in collective migration regulated in cellular collectives?

To answer these questions, we used Madin-Darby Canine Kidney (MDCK) epithelial monolayers as an experimental system to study collective migration and its associated energetic cost. Epithelial monolayers are a convenient system to study collective cell migration due to their experimental tractability and their highly coordinated movement resulting from strong cell-cell adhesions [38,81,89,90]. It is for this reason that a large body of research exists to characterise the collective behaviour of epithelia in a tremendous range of different processes [81]. This body of research has yielded insightful quantifications of the spatial and temporal patterning of velocities within epithelial monolayers as well as the forces that give rise to these velocities [83, 84, 91, 92]. The use of traction force microscopy (TFM), an experimental technique used to measure the cell-substrate tractions exerted by cells during migration, has yielded powerful insights relating the roles of cell polarity and location to their patterns of force production [82,93], yet there is a lack of understanding of how these observed forces translate to the rate at which individual regions of the epithelial monolayer are using energy to propel migration, and how this rate of energy consumption is fed back to the overall patterns of migration.

Recently, the study of collective migration in epithelial monolayers has been expanded with works focussing on the concept of energy. On the one hand, there exist biochemical metabolic studies that study the rate of energy consumption in the monolayer using the NADH/NAD^+ ratio, ATP utilisation, or the rate of oxidation [94, 95]. These studies show that collective migration is *energetically costly* and that cell metabolic rates are not uniformly spread throughout the monolayer [94, 95]. On the other hand, computational physical metrics such as entropy, enthalpy, and vorticity, which are all based on automated computation of velocities in the monolayer, have been used to describe abstract concepts of energy in migrating monolayers through the lens of statistical mechanics and fluid dynamics [91, 96]. While insightful, there are several reasons as to why these studies must be complemented to address the main question of this chapter which is how the energetic burden of migration is shared throughout the monolayer. First, neither of these two classes of studies directly assess the energy expended by producing active cell forces during migration. In the case of biochemical metabolic studies, the markers for cellular metabolism reflect the total energetic usage of cells, which also depends on a host of other

processes, such as proliferation [97]. Likewise, the abstract continuum physics approaches of *e.g.* Lin *et al.* [91, 96] do not relate energy to forces in their models.

Therefore, in an attempt to close the gap identified in the understanding of energetics related to active forces in collective migration, this chapter proposes that continuum modelling of cellular monolayers can be used to characterise the collective mechanics of collective migration in conjunction with their associated mechanical energy. Previously, classical continuum mechanical metrics such as strain and stress have been used to address open problems in collective migration and characterise the onset of epithelial events such as migration, expansion, buckling, symmetry-breaking, and wound healing [87]. In this work, we demonstrate that a continuum mechanics model of epithelial migration can be used to compute the mechanical work done by the tissue in generating forces within a migrating epithelial monolayer. Mechanical work is an established physical quantity that can be computed using continuum mechanics, given experimental velocity data from a migrating monolayer, making this approach an experimentally simple method to assess several mechanical metrics concurrently, which are experimentally and computationally intractable in more complicated experimental protocols such as TFM.

We address the key question of this chapter by relating the average rate of mechanical work in the monolayer to the migratory dynamics, and show that there is a heretofore unknown migration regulation mechanism – a rate of 'optimal' energy expenditure set by the ensemble averaged behaviour of the constituent cells, above which 'high spending' regions of the tissue are decelerated, and 'low spending' regions are accelerated. We show that in freely expanding (homeostatic) tissues and those subject to external stimuli alike, this regulation occurs persistently, dynamically, and precisely. To explore how the self-regulation of migratory dynamics in an epithelium is impacted by external stimuli, we used an external electric field of biologically relevant strength, $\mathcal{O}(1\ \mathrm{V/cm})$, to induce electrotaxis, a process by which cells exhibit highly directed and accelerated migration in response to an external DC electric field [1,38–40]. To span the space of biologically relevant electric stimuli that vary in both time and strength, we used three unique waveforms to stimulate epithelial tissues *in vitro*: a 3 V/cm pulse; a 1.5 V/cm pulse; and a 3 V/cm linear ramp. This allowed us to characterise the spatiotemporal dynamics of energy expenditure in epithelia migrating under the influence of an external stimulus. Interestingly, we find that over-expenditure may be the cause of commonly reported 'slow-downs', wherein electrotaxing cells will slow their migration during exposure to the field. Finally, a careful quantification of the dynamics of forces, velocities, and energies allows us to explore what it means for a tissue to migrate 'efficiently', both in the absence and presence of an external stimulus, and show that freely expanding epithelia actively regulate themselves to operate within a maximally efficient regime.

This chapter is organised as follows. In Section 3.2, we detail the experimental protocol used and the resulting data. In Section 3.3 we present the continuum mechanics model

used to infer traction forces in the monolayer from the experimental data, and derive the associated metrics that relate mechanics to measures of energetic expenditure. We use the method derived in Section 3.3 to elucidate a regulatory mechanism in Section 3.4. By carefully considering the dynamics of forces, velocities, and energies, we propose a simple regulatory mechanism, which we show is unaffected by the presence of an external stimulus in the form of an EF. In Section 3.5, we analyse this regulation mechanism in detail, to find that a simple mechanical model of force regulation dependent on previous energetic expenditure alone is capable of recapitulating the observed joint dynamics of velocity and average mechanical power. This mechanism is then demonstrated at the single-cell level in Section 3.6, where we show that single cells have a characteristic temporal distribution of velocities and power expenditures. Putting these findings together, we take a closer look at regulation during stimulation with an external stimulus in Section 3.7, and propose that passive adaptation to the EF signal also occurs. Finally, we discuss the consequences for the mathematical modelling of epithelial dynamics, which we will take into account in the next chapters, in Section 3.8.

3.2 Data and methods

3.2.1 Experimental set-up

Wild type Madin-Darby Canine Kidney II cells (MDCK-II, courtesy of Nelson Lab, Stanford) were cultured in media consisting of low-glucose (1 g/L) DMEM with phenol red (Gibco; powder stock), 1 g/L sodium bicarbonate, 1% streptomycin and penicillin (Gibco), and 10% fetal bovine serum (Atlanta Biological). Cells were maintained at 37°C and 5% CO_2 in humidified air. All experiments were conducted between P12 and P18 and cells were mycoplasma tested twice annually. For all experiments, time-lapsed montage images were captured every 10 minutes for 8 hours with an automated inverted microscope (Zeiss Axio Observer Z1) equipped with a 5x/0.16 phase contrast Zeiss objective, a Photometrics Iris 9 (Photometrics, Inc.) sCMOS camera, and an XY motorised stage, all controlled using Slidebook (3i Intelligent Imaging Innovations). Additionally, the stage-top was enclosed in a cage incubator kept at 37°C. Captured images were stitched using (Fiji Is Just) ImageJ's `Grid/Collection stitching` plug-in (three 3×3 montages with 15% spatial overlap) [98], template matched using the `Template Matching` plug-in, masked, and aligned before further analyses.

The bioreactor used for electrotaxis was modified from the SCHEEPDOG platform for uniaxial stimulation of large-scale tissues developed by the Cohen Laboratory at Princeton University [40]. Briefly, a laser-cut and glued acrylic body houses an electrode (Ag—AgCl) - electrolyte (PBS) - salt bridge (4% w/v agarose/PBS) assembly and is clamped to a

10cm tissue culture dish to create a microfluidic region supported by a 250μm thick stencil of PDMS. Electrodes protrude from the top of the device, allowing for direct integration with external controllers, and the body contains both an inlet and outlet for fresh media perfusion orthogonal to the direction of electrical stimulation; this aids in temperature regulation (see [40] for full characterisation) and perfusion of electrochemical byproducts such as reactive oxygen species. Titanium probes are inserted into the agarose salt bridges to monitor the voltage and current across the microfluidic channel, measured by a USB oscilloscope (Analog Discovery 2, Digilent Inc.) and proportionally controlled by a Keithley 2450 SourceMeter (Tektronix) and custom MATLAB script. For a complete summary of the approach and guidelines, CAD files of the full device, and/or custom closed-loop control scripts, see previous works [1, 39, 40, 90, 99, 100] and the corresponding repositories (`github.com/cohenlabprinceton`). Per experiment, three 5×5mm^2 square tissues were seeded onto 10cm tissue-culture plastic dishes (CELLTREAT) using stencils of 250μm-thick silicone elastomer (Bisco HT6240) cut with a hobbyist razor-writer (Silhouette Cricut). Suspensions of MDCK cells were seeded into the stencil patterns at a volume of 10μL and density of 2.5 ($\pm$0.15) x 10^6 cells/mL. Cells were given one hour to adhere before the dish was flooded with media and incubated for 18 hours prior to stencil pull and device integration, as described below.

In all experiments (with and without electrotaxis), tissue culture stencils were pulled approximately one hour before imaging began. Immediately post-stencil pull, dishes were refilled with fresh culture media followed by direct integration with the bioreactor. The assembled bioreactor is then placed into the microscope, situated within a custom-built cage incubator maintained at 37°C and perfused with fresh media continuously bubbled with 5% CO_2 using a peristaltic pump (Instech Laboratories) at a rate of 2.5 mL/h. When integrating electrical stimulation, experiments begin with a one hour 'control' period before stimulation is turned on. Relevant stimulation is then delivered for three hours before a four hour 'relaxation' period during which there is no stimulation but imaging persists.

3.2.2 Image processing and velocity computations

All nuclear data used for computational analyses were obtained from transmitted light phase contrast microscopy using an in-house Fluorescence Reconstruction Microscopy tool (a convolution neural network trained to produce images of various biomarkers from phase contrast images) [101]. After initial image preparation, velocity vector fields were calculated using MATLAB's `PIVlab` [102]. A two-pass fast Fourier transform was performed with interrogation window sizes of 128 x 128 pixels and 64 x 64 pixels with 50% overlap. Vectors outside of five standard deviations were treated as outliers and replaced with interpolated data. After conversion of phase contrast data to nuclear data, we used ImageJ's TrackMate [103] to track all cells throughout the experimental timecourse.

3.3 A model for cellular mechanics in a migrating epithelial monolayer

3.3.1 Computing cellular tractions from stress balance

In this chapter, we are interested in quantifying mechanical metrics from experimental data of collective MDCK migration. MDCK epithelial monolayers make up a flat sheet, which means that the ratio of the height, h, to the out-of-plane curvature, R, is small, *i.e.*,

$$\frac{h}{R} \ll 1. \tag{3.1}$$

This scaling justifies the common assumption in the literature that cell forces upon the substrate and within the monolayer are planar and do not have any out of plane contributions [82]. Since cell movements within a monolayer, which we can quantify using PIV (see Figure 3.1a), are tightly coupled through intercellular adhesions, active forces within the monolayer create local stresses, which are transmitted throughout the tissue across a typical length scale [81, 83, 84, 104]. In a continuum mechanics formulation, forces in a two-dimensional material, like an epithelial monolayer, are described by a force vector field, $\mathbf{F}$, and monolayer stresses are described by the (Cauchy) stress tensor, $\boldsymbol{\sigma}$, which is a tensor field that describes the state of stress of a deformed material. The cell forces within the monolayer are typically interpreted as the active force transmitted by cells onto other cells, as well as on the substrate, to balance the interstitial stresses inside the monolayer, as well to generate their active movement. It is important to note that a continuum force field in an epithelial monolayer describes a variety of forces that cells exert, including active and passive forces exerted on their neighbours, as well as active traction forces.

Forces and stresses are physically related through force balance, which is captured by the law of conservation of linear momentum. In the context of epithelia, inertial effects can be ignored [105], and so conservation of linear momentum reduces to a force balance condition between monolayer stresses and cell (active) tractions, $\mathbf{T}$ [82], which then reads

$$\frac{\partial \boldsymbol{\sigma}_{\boldsymbol{xx}}}{\partial x} + \frac{\partial \boldsymbol{\sigma}_{xy}}{\partial y} = -\mathbf{T}^{\star}_{x}, \tag{3.2}$$

$$\frac{\partial \boldsymbol{\sigma}_{xy}}{\partial x} + \frac{\partial \boldsymbol{\sigma}_{yy}}{\partial y} = -\mathbf{T}^{\star}_{y}, \tag{3.3}$$

where

$$\mathbf{T}^{\star}_{x} = \mathbf{T}_{x}/h, \quad \mathbf{T}^{\star}_{y} = \mathbf{T}_{y}/h, \tag{3.4}$$

and $\mathbf{T}_x, \mathbf{T}_y$ are the x and y-components of the tractions, respectively. Equations (3.2)

and (3.3) represent force balance per unit height in the tissue. By additionally making the assumption that the monolayer is homogeneous and isotropic, Equations (3.2) and (3.3) allow for the direct computation of monolayer tractions when the state of stress in the tissue is known. Likewise, when the forces are known, for example through traction force microscopy (TFM), the components of the stress tensor, σ, can be found by solving Equations (3.2) and (3.3) as PDEs subject to appropriate boundary conditions – for example, no-slip or free boundary conditions [82].

3.3.2 Constitutive assumptions for MDCK epithelia

With a known stress field, forces in the tissue can be directly computed through Equations (3.2) and (3.3). Physically, the state of stress in the tissue can be directly determined by the state of deformation by introducing a *constitutive assumption* that prescribes the relation between stresses and strains. The continuum mechanics of collectively migrating epithelia are characterised by several well-established canonical models that account for a wide range of physical contexts, like varying levels of proliferation, substrate stiffness and tissue geometries [105–109]. The use of such models is supported by a large body of research establishing the validity and applicability of modelling epithelial monolayers as active viscoelastic materials. Viscoelastic models capture key characteristics of the migrating epithelial monolayer, like stress relaxation, which follow from the physical properties of the constituent cells [81, 108].

In general, such models of viscoelasticity provide constitutive relations for the monolayer stresses and activity that are described entirely by the directly observable physical quantities of strain,

$$\boldsymbol{\epsilon} = \frac{1}{2}(\nabla \mathbf{u} + \nabla^T \mathbf{u}), \tag{3.5}$$

and strain rate,

$$\dot{\boldsymbol{\epsilon}} = \frac{1}{2}(\nabla \mathbf{v} + \nabla^T \mathbf{v}), \tag{3.6}$$

respectively, where $\mathbf{u}$ is the displacement field, and $\mathbf{v} = \dot{\mathbf{u}}$ is the velocity. Recall that in this work, the displacement and velocity fields are found experimentally using PIV from microscopy images. Appropriate constitutive relations may be given by, for example, the Maxwell, Kelvin-Voigt, or simple viscous models. For a review, see [81, 108]. While these models all give rise to different mechanical behaviours in response to external forces and internal activity [81, 83, 108, 110], their applicability as suitable descriptions of monolayer mechanics in different contexts is well established.

In sum, having constitutive equations that link the directly observable velocity field, $\mathbf{v}$, which is measured using PIV, to the stress tensor, $\boldsymbol{\sigma}$, allows for the direct computation of a continuum traction field through Equations (3.2) and (3.3). In this view, inferring monolayer stresses and forces from velocities reduces to choosing an appropriate constitutive

assumption that most faithfully describes the epithelial system of interest. An MDCK epithelial monolayer, with its typical proliferation and strong cell-cell adhesions, can be effectively described as a freely migrating viscous liquid, such that a constitutive assumption of an active viscous material at low Reynolds number is the most appropriate [105]. Such models take into account the stresses within the monolayer and their direct transmission to the substrate by cell-substrate adhesions. By choosing a simple viscous constitutive law the Cauchy stress tensor is readily expressed in terms of the rate of strain tensor,

$$\boldsymbol{\sigma} = \eta\dot{\boldsymbol{\epsilon}}, \tag{3.7}$$

where η is the effective viscosity of the material. By then making the canonical assumption that tissue flow is incompressible, *i.e.*,

$$\nabla \cdot \mathbf{v} = 0, \tag{3.8}$$

and imposing conservation of linear momentum to find the tractions, through Equations (3.2) and (3.3), the velocity field is described as a generalised Stokes flow given by

$$\eta\nabla^2\mathbf{v} = \mathbf{T} + \gamma\mathbf{v}. \tag{3.9}$$

The left-hand side of Equation (3.9) can be interpreted as a viscous force per unit area that opposes spatial variations in the velocity. The right-hand side expresses cell tractions as an external driving force per unit area, $\mathbf{T}$, and also contains a friction force per unit area with the substrate of the form $\gamma\mathbf{v}$, where $\gamma > 0$. This term, which is linear in the cell velocities, is a canonical approximation for the cell-substrate friction and is analogous to the viscous drag experienced by a low Reynolds number fluid flow over a surface [108, 110], which is likewise represented by linear friction.

3.3.3 Mechanical parameters

The mechanical properties of MDCK epithelial monolayers on soft substrates are well characterised, and theoretical estimates of key mechanical parameters are supplied in the literature. Here, we report the mechanical parameters used in all force computations.

3.3.4 Computing mechanical work in a migrating epithelium

To assess the energetic cost associated with the active cell forces within the monolayer, a simple model for the energy associated with migration in the monolayer is needed. Previous investigations into continuum descriptions of energetics in cellular collectives

Parameter	Reference value	Source
Friction coefficient, γ	100 Pas/μm^2	[104, 111]
Monolayer viscosity, η	25 MPa·s	[104, 111]
Hydrodynamic length, $\sqrt{\eta/\gamma}$	0.5 mm	Computed

Table 3.1: Mechanical parameters used for the computation of forces and stresses throughout all experiments.

have used various energy metrics to relate the velocity field to cellular energetics [91, 96]. For example, Lin *et al.* [91, 96] study the distribution of kinetic energy,

$$E = (1/2)|\mathbf{v}|^2, \tag{3.10}$$

and enstrophy,

$$\Omega = (1/2)|\boldsymbol{\omega}|^2, \tag{3.11}$$

respectively, where $\boldsymbol{\omega}$ is the vorticity tensor. Their investigations find that the spatiotemporal distribution of energy and velocity in the tissue is stochastic and follows a q-Gaussian distribution, which is similar to the empirical distribution of monolayer velocities [91]. Although both kinetic energy and enstrophy provide a good mathematical abstraction for the energy stored in a tissue, these two metrics do not reflect the physical work done by the tissue as a consequence of the forces associated with migration. Since Equation (3.9) provides a theoretically justified continuum framework to describe forces from velocity data, the question becomes how to use this framework to derive a metric directly informative of the mechanical energy spent during collective migration in the cell monolayer. Perhaps the simplest metric for the mechanical energy associated with force production is that of mechanical work, W. Given a force field, $\mathbf{F}$, and a velocity field, $\mathbf{v}$, the mechanical work is readily expressed in the Eulerian frame of reference using the standard expression

$$\frac{\partial W(\mathbf{x}, t)}{\partial t} = \mathbf{F}(\mathbf{x}, t) \cdot \mathbf{v}(\mathbf{x}, t). \tag{3.12}$$

As we are interested in the regulation of active forces jointly with mechanical work, we would like to track the work done by a particular tissue element, as opposed to that at a particular fixed location in two-dimensional space. These two concepts are different, since monolayer migration is characterised by a flow that creates advection in the tissue, *i.e.*, the cells that occupy one spatial location at some time point will occupy another spatial location at a future time point. Put differently, the fact that the tissue is migrating means that the work, W, if interpreted as a scalar quantity associated with a fixed tissue element, is convected along with the monolayer velocity. Therefore, to track the mechanical work done by the monolayer during the course of its migration, it is important to find an

expression for the mechanical work done by the part of the tissue occupying any fixed location x at time t during its history. To take into account the convection with the monolayer velocity, and quantify the work that corresponds to a fixed region of tissue that occupies given coordinates in space and time, we use the material time derivative, $\mathrm{D}/\mathrm{D}t$. For a scalar function, φ, the material derivative is defined as

$$\frac{\mathrm{D}\varphi}{\mathrm{D}t} = \frac{\partial\varphi}{\partial t} + \mathbf{v}\cdot\nabla\varphi. \tag{3.13}$$

Combining Equations (3.12) and (3.13), yields the material derivative of the mechanical work, W,

$$\frac{\mathrm{D}W(\mathbf{x},t)}{\mathrm{D}t} = \mathbf{F}(\mathbf{x},t)\cdot\mathbf{v}(\mathbf{x},t) + \mathbf{v}\cdot\nabla W. \tag{3.14}$$

3.3.5 Computing average mechanical power in a migrating monolayer

By integrating in time, Equation (3.14) yields the total mechanical energy used by the region of tissue that is at location x at time t, *i.e.*,

$$W(\mathbf{x},t) = \int_0^t \frac{\mathrm{D}W(\mathbf{x},s)}{\mathrm{D}s}\mathrm{d}s. \tag{3.15}$$

This is a useful metric to understand how much energy has been expended by different regions of tissue at the same time point. In the context of epithelial migration, cells have processes that allow them to absorb and metabolise energy from their environment continuously. Therefore, a metric that relates to the *rate* of energy expenditure in the monolayer can be used more meaningfully to assess energetic cost for the epithelial monolayer during its migration. To achieve this, we consider the quantity of *average mechanical power*. The average power expended by a tissue element allows us to assess the average rate at which energy has been used by a tissue element during epithelial migration, as a function of force production against the substrate and against local deformations.

Defining average mechanical work would be straightforward if one were to ignore tissue expansion, and assume that the monolayer occupies the same spatial extent during the experiment. In that case, the average power expended could be readily computed from the work computed in Equation (3.14) by taking the temporal average of the integrand, *i.e.*,

$$P(\mathbf{x},t) = \frac{W(\mathbf{x},t)}{t}. \tag{3.16}$$

During epithelial expansion, however, the edges of the tissue move to invade new space,

meaning that some spatial locations will only contain tissue for a subset of the experimental time. This implies that dividing through by the total time, as in Equation (3.16), will bias the average power expended by the tissue at the tissue edges. We address this issue by correcting for the total time that any given spatial location has been occupied by epithelium to write

$$P(\mathbf{x},t)=\frac{\int_0^t \mathbb{I}_{\text{tissue}}(\mathbf{x},s)W(\mathbf{x},s)\mathrm{d}s}{\int_0^t \mathbb{I}_{\text{tissue}}(\mathbf{x},s)\mathrm{d}s}. \tag{3.17}$$

Here, $\mathbb{I}_{\text{tissue}}(\mathbf{x},s)$ is an indicator function that is equal to one if there is tissue present at location $\mathbf{x}$ at time s and zero otherwise. Note that Equation (3.17) reduces to Equation (3.16) when $\mathbb{I}_{\text{tissue}}(\mathbf{x},s)\equiv 1$, which is particularly true in the bulk, *i.e.*, the centre of the tissue, which is always populated. The general formulation in Equation (3.17) provides an unbiased quantity that compares the time-averaged rate of mechanical work at any spatial location in the tissue and at any time. Finally, we remark that this average power metric takes into account the average rate of work from the beginning of the experiment, and so all of past time contributes equally to this metric.

3.4 Mechanical homeostasis is maintained by regulating tissue power expenditure

3.4.1 Quantifying mechanical trends in migrating epithelia

The continuum framework to quantify mechanical trends of force, mechanical work, and average mechanical power developed in Section 3.3 can be applied to experimental data of freely expanding epithelial monolayers. During free epithelial expansion, it is well-known that the bulk of the tissue exhibits velocities that strongly fluctuate in time and in space, with these fluctuations being organised in small and heavily correlated regions [83, 91]. In contrast, the edges show highly directed motion towards the free space surrounding the tissue. These trends are well captured by the PIV data of the freely expanding monolayer in Figure 3.1a. Previous studies using TFM have related the stereotypical outward migratory patterns of MDCK monolayers to typical patterns of localised traction across the monolayer [81,82,84]. We begin by using our continuum method to estimate and quantify cellular forces in a freely expanding epithelial monolayer to make a comparison between the patterns of total cell forces computed using a continuum framework, and the mechanical trends that have been previously described in the literature [81, 82, 84].

First, we computed monolayer forces by combining PIV data from an experiment involving a freely expanding monolayer and the constituent equation for monolayer forces, Equation (3.9). Figure 3.1b shows a heatmap of the magnitude of the resulting computed

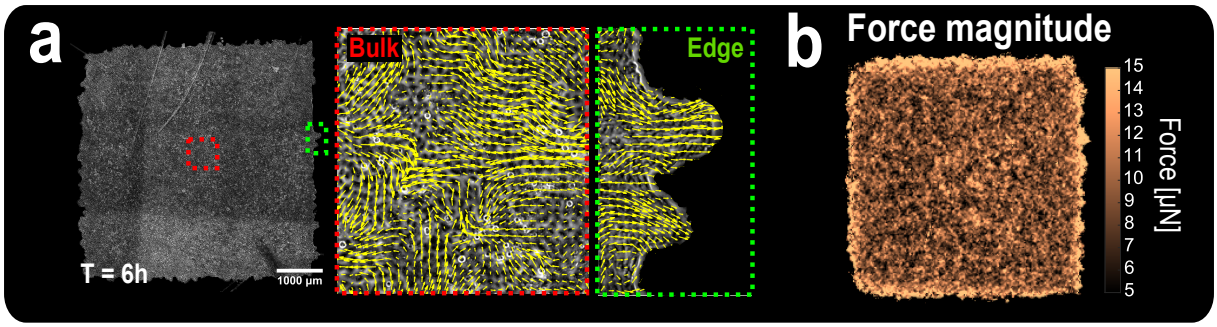


Figure 3.1: Spatial organisation of velocities and forces in a freely migrating epithelial monolayer. **(a)** Phase image of a representative $5\times5\text{mm}^2$ MDCK-II monolayer six hours after stencil pull. (Red) A $500\times500\mu\text{m}^2$ region in the bulk, with overlayed PIV vectors. (Green) A $250\times500\mu\text{m}^2$ region at the edge of the tissue. **(b)** Heatmap of the computed magnitude of computed traction forces, $\mathbf{T}$, at $t = 1$ hour using Equation (3.9).

x-component of the traction field at $t = 1$ hour. The bulk of the tissue is characterised by intercalated regions of low and high force. These regions of low and high force are known to propagate through the tissue as standing waves [84, 112]. In contrast, the edges of the freely expanding monolayer are defined by an uninterrupted front of high force magnitude, as shown in Figure 3.1b. This is consistent with the observation that in free epithelial expansion, the speed of migration and the directional order of migration are highest at the edges [39, 113]. Given this intuitive spatial patterning in force generation, we then extended our analysis to address the question of how much energetic cost is borne at different regions within the monolayer.

To understand the spatial and temporal distribution of time-averaged mechanical power in the freely migrating monolayer, we computed the mechanical power by using the computed forces from Equation (3.9) in the model of Equation (3.17). The distribution of time-averaged mechanical work is visualised using a heatmap at $t = 6$ hours in Figure 3.2a. The heatmap in Figure 3.2a shows that the average mechanical power is highest and constant at the edges, lowest in the region between the edges and the bulk, and increases towards the bulk of the tissue. This characteristic spatial distribution of power is conserved throughout the experiment, as shown by several representative temporal cross-sections in Figure 3.2b. The distribution of mechanical work supports the view that collective cell migration creates energetic costs that are not reflected equally in all tissue locations. At the same time, the spatial distribution of the time-averaged mechanical power shown in Figure 3.2 is physically intuitive. The fact that the highest average mechanical power is located at the edges is consistent with the rapid outward migration at the edges of MDCK monolayers [114]. Likewise, the highly turbulent and random movement of cells in the bulk of the tissue creates large viscous forces, which are reflected by the Laplacian term in Equation (3.9). The transition region between edges and bulk is characterised by slower outward movement than the edges, but also more coordinated movement than the bulk, explaining why the energetic contributions from these regions are lower.

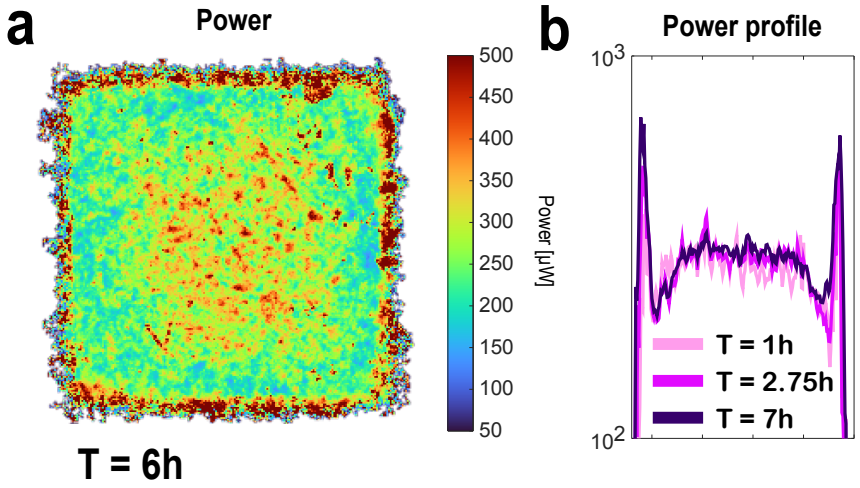


Figure 3.2: Spatial and temporal distribution of average mechanical power in a freely expanding epithelial monolayer. **(a)** Heatmap of computed power in one representative tissue at $t = 6$ hours. **(b)** Cross-section through the middle 2mm strip of power heatmap at different time points during the experiment.

Having characterised the spatiotemporal patterns of average mechanical power, we investigate how average mechanical work impacts collective migration. Given that each tissue region has its characteristic distribution of average mechanical power, and given that the edges are exposed to a persistent outward migratory cue, an optimal region to study the effect of mechanical power on collective migration in isolation is in the bulk of the tissue. For this reason, we limit all future analyses to the $2\times2\text{mm}^2$ bulk region of the tissues introduced in the previous section. The first thing to note is that, although epithelia are inertia-free, there are large fluctuations in the velocity field that can span up to 10 cell lengths [112]. This divides the tissue into regions that effectively accelerate, thus fuelling collective migration by incurring a high energetic cost, while other regions of the monolayer expend far less energy and so incur a low energetic cost. A mode of migration in which the energetic cost is not borne equally across the tissue seems inefficient *a priori*, so we explore whether there is a mechanism that regulates force production in high-spending regions.

As a starting point to investigate whether the average power expended has an effect on the amount of force that is produced in the tissue in the future, we related the average mechanical power to temporal changes in the force, *i.e.*, the jerk. The rationale for doing so is that we hypothesised that tissue parcels that have expended greater amounts of mechanical power would moderate their future force production, and tissue parcels that have expended more moderate amounts of mechanical power would be able to increase their force production in the future. To investigate whether such a regulation mechanism might be present, we used the computed averaged power for each spatial location at each given time point, across the different biological replicates using Equation (3.17). For each such data point, we computed the corresponding jerk at that time and spatial location

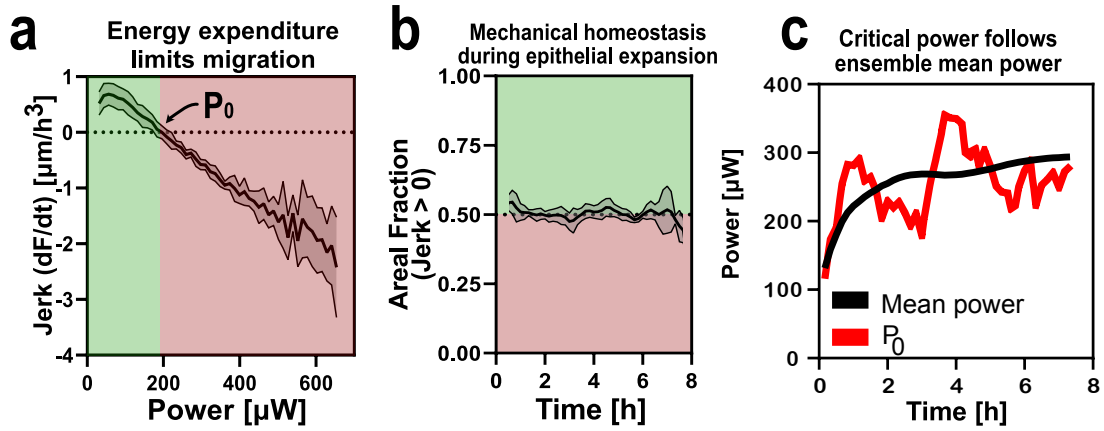


Figure 3.3: Interplay between power, velocities, and forces. **(a)** Change in force, *i.e.* jerk, changes sign at a critical power, p_0. **(b)** Temporal evolution of the fraction of tissue area that is accelerated, averaged over all tissues. **(c)** Temporal dynamics of average mechanical power and the critical power, p_0.

using a first-order time derivative of the forces predicted by Equation (3.9). We compute a piecewise linear spline to fit through the datapoints that lie within the 5% and 95% percentiles, along with its standard deviation, which is shown in Figure 3.3a. The plot in Figure 3.3a shows the average mechanical power against the jerk. Strikingly, the jerk is indeed directly related to average power: future force production increases (decreases) when average power is below (above) a *critical power*, which we denote here by p_0.

Given that the time-averaged mechanical power seems to play a regulatory role in future force production across different locations in the monolayer, we explore how this critical power itself evolves in time. The distribution of average mechanical power around its critical value, p_0, we hypothesised, might be driving a type of mechanical homeostasis: regions of tissue that have expended more energy than p_0 reduce their active force and so decelerate (*idem* for regions below p_0). To understand first whether the distribution of velocities in the epithelium is always such that there is no net acceleration, we calculated the fraction of the tissue area that had a positive jerk, averaged across all tissues. Throughout the experiment, the fraction of tissue area that is accelerating remains at 0.5, as shown in Figure 3.3b. A constant value of the accelerating area, together with the fact that the rate of force production is impacted by the time-averaged mechanical power, suggests that the critical power, p_0, might fluctuate in time to respond to the energetic output of the migrating monolayer. To relate the idea that mechanical homeostasis might be related to the evolution of the critical power, we computed both the spatial mean of the average mechanical power for each time point, as well as the critical power, p_0. A comparison of p_0 with the spatial mean of the average power in the tissue in Figure 3.3c shows that p_0 oscillates around the spatial mean of the average power with a period of approximately 1-2 hours. This suggests that p_0 is continuously adjusted to restore the fraction of the tissue with positive jerk to its homeostatic value of 0.5.

In summary, we have established that there are characteristic patterns of energy expenditure in freely migrating epithelia, and characteristic patterns of mechanical power in the tissue, with the edges of the tissue being most mechanically active, as is consistent with previous data that performed metabolic analysis during collective cell migration [94,95]. By carefully relating temporal changes in force production to the time-averaged mechanical power, we show that average mechanical power expended influences the future forces produced within tissue parcels and gives rise to a regulatory mechanism that suppresses sustained outputs of mechanical power above that of the spatial mean of the mechanical power in the monolayer. We propose this novel mechanism of force regulation as a key mechanism involved in maintaining mechanical homeostasis throughout the tissue.

3.4.2 External drivers of collective migration are regulated by tissue power generation

Having established a framework that describes how forces and average mechanical power are related and regulated in freely expanding epithelial monolayers, the next question to explore is how the joint dynamics of force and power change when the analysis of the previous section is extended to epithelia driven out of equilibrium by an external stimulus. Concerning migration specifically, there exist myriad stimuli that elicit directed collective migration in epithelia, including chemotaxis [7], durotaxis (frictiotaxis) [115–117], and electrotaxis [1, 39, 90, 118]. For this study, we leveraged electrotaxis due to its biological prevalence [38, 119, 120] and experimental tractability [40, 100]. When exposed to an external electric field with field density $\mathcal{O}(1\text{V/cm})$, MDCK cells exhibit a mode of collective migration characterised by a near order-of-magnitude increase in speed and a total alignment of migration with the external field direction. For a graphical overview of this highly directed mode of migration, see Figure 3.4a. Motivated by the range of physiologically relevant field strengths recorded *in vivo* [121], electrotaxis in this experimental set-up included three different voltage traces: a 3 V/cm pulse, a 1.5 V/cm pulse, and a 3 V/cm linear ramp (voltage traces shown in Figure 3.4b).

Studies on collective electrotaxis commonly describe migration using the metrics of mean speed and directional order [39, 40, 100]. The directional order of the tissue, Φ, is defined as

$$\Phi = \frac{1}{N}\sum_{i=1}^{N}\cos\theta_i, \tag{3.18}$$

where θ_i is the angle between the velocity vector at location i and the EF, and N is the number of grid points occupied by the tissue. Since the stimulation traces in this experiment span a range of different conditions, these commonly used metrics provide a good starting point to characterise the effects of the different stimulation traces on

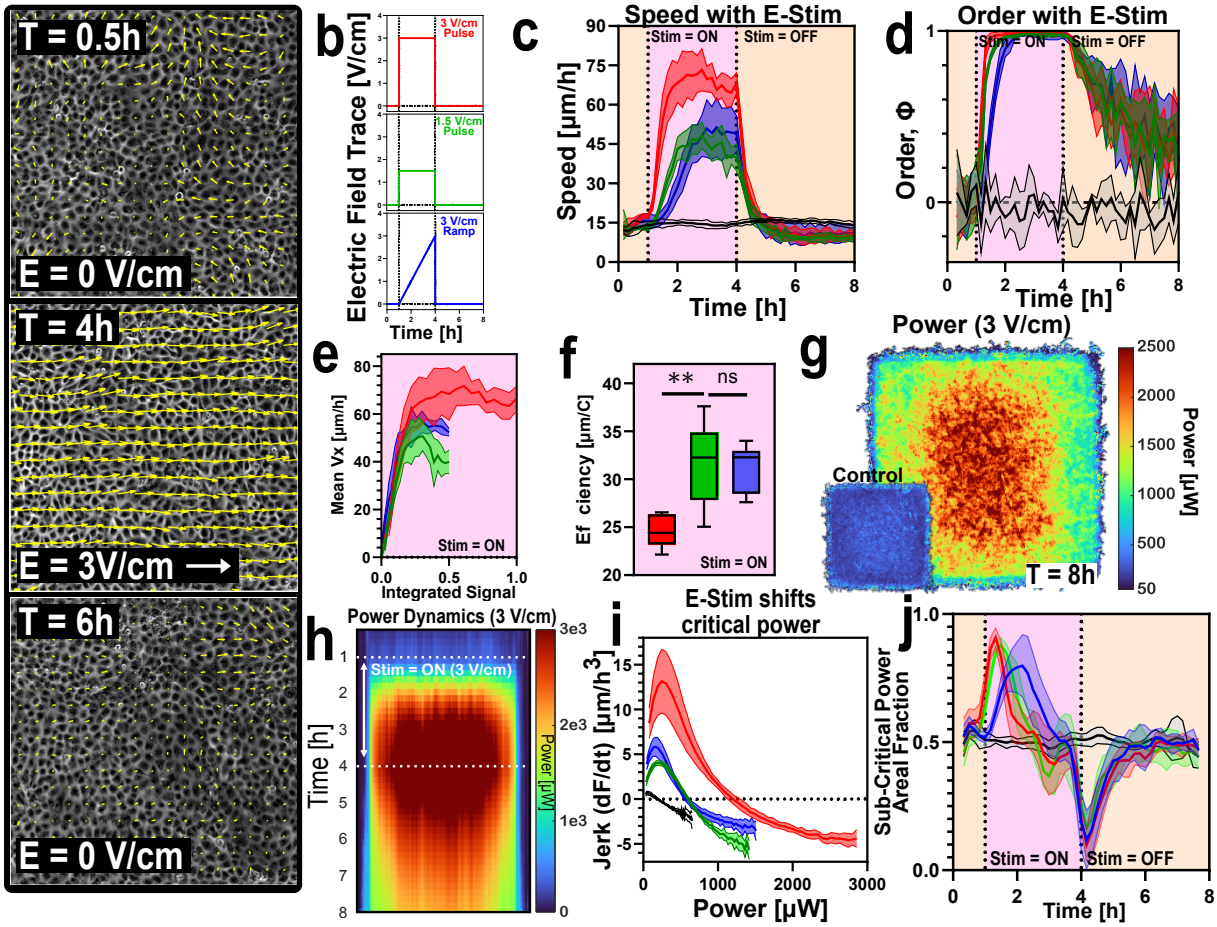


Figure 3.4: Energetics of externally driven epithelial migration. Unless otherwise mentioned, red key indicates 3V/cm pulse, green key indicates 1.5V/cm pulse, and blue key indicates 3V/cm ramp. **(a)** Phase images overlayed with PIV vectors in a $500\times500\mu m^2$ region in the bulk of an electrotaxing tissue. **(b)** The three different stimulation traces we use for the remainder of this study. **(c)** Speed dynamics in electrotaxing tissues. **(d)** Order dynamics in electrotaxing tissues. **(e)** Speed as a function of integrated area of signals shown in **(b)**. **(f)** The efficiency of different stimulation traces, expressed as mean cell displacement per unit of injected charge. **(g)** Heatmap of power averaged over all tissues stimulated with the 3 V/cm pulse trace (red). (Inset) Heatmap of power averaged over all control tissues. **(h)** Power kymograph averaged over all tissues stimulated with the 3 V/cm pulse trace (red). **(i)** Force production (jerk) vs. power for the different stimulation traces. **(j)** Areal fraction of positive jerk averaged over all tissues for different stimulation traces.

collective migration. By averaging the velocities within all tissue bulks (*i.e.*, the same $2\times2\text{mm}^2$ tissue region used to compute the critical powers in the previous section), the evolution of the mean monolayer velocity can be obtained. These velocity dynamics are shown in Figure 3.4c. With a 3 V/cm pulse, tissues exhibit a dramatic acceleration in the first hour of stimulation before plateauing and even slowing down *during* stimulation, as shown in previous publications [1]. When stimulated with a 1.5 V/cm pulse, tissues exhibit a lower peak speed, with a nearly identical slowdown in the latter half of stimulation. However, when stimulated with a 3 V/cm ramp, there is a plateau at the end of stimulation, rather than a slowdown. The question of why different stimulation traces result in different monolayer velocity profiles is not well understood, and will be addressed in the next chapter of this thesis, Chapter 4. For now, while the velocity curves in the experiments are different, we investigated the directional order of migration and found that all stimulation traces result in highly directional movement, as shown in Figure 3.4d.

So far, we have shown that different EFs can elicit different mechanical responses in the tissue. To further understand how the collective response of the monolayer depends on the injected charge through the EF, we first characterised the dynamic responses of tissues to different stimulation traces by plotting the velocity for each of the stimulation traces as a function of the injected electric charge (normalised to the total injected charge in the 3 V/cm pulse), which is shown in Figure 3.4e. Here, it can be seen that additional injected charge past a some critical initial amount leads to very moderate gains in the monolayer velocity. This invites us to think about stimulation with an electric charge to elicit a collective response as a question of efficiency. The idea that different stimulation traces might be more or less efficient with respect to velocity as a function of injected charge can be quantified using a metric for efficiency, which we define as the displacement of the constituent cells within the tissue per unit of injected charge. By computing this 'miles-per-gallon' metric, we show in Figure 3.4f that, in fact, the stimulation trace with the highest constant field density, namely the 3V/cm pulse, which naively might be thought to be the best to stimulate the tissue, is the least efficient.

Having understood that different EFs elicit different tissue responses, and that tissue responses are nonlinear as a function of the injected charge, we turn to understanding how electrotaxis changes the relationship between average mechanical power, forces, and velocities, which we found in the previous section. Again, using Equation (3.17), we quantified the average power in the electrotaxed tissues. Figure 3.4g shows a heatmap averaged over all tissues at two hours into electrotaxis (*i.e.*, at $t = 3$ hours into the experiment) for a 3 V/cm pulse (kymographs are qualitatively similar for other stimulation cases, data not shown). The power heatmap shows that cells in the bulk of the tissue are most active during electrotaxis, which is consistent with the fact that the cells in the tissue bulk are most responsive to the electric field [1, 39]. This is further corroborated by the kymograph in Figure 3.4h, which shows that there is a high-power patch in the bulk of

the tissue at the onset of stimulation. This high-power zone exhibits average mechanical powers far greater than in the freely expanding case, so we reason that the relationship between monolayer forces and average mechanical powers must be different than to that in the freely expanding case. To explore this, we compute the rate of force production as a function of average mechanical power, for all stimulation traces during electrotaxis (see Figure 3.4i). This is analogous to the analysis in Figure 3.3 of the previous section. While the scale of the curves is different, the qualitative relationship between mechanical power and jerk in all stimulation traces is the same. Similar to the freely expanding case, Figure 3.4i shows that there exists a critical power, p_0, beyond which the jerk is negative. For both the 3 V/cm ramp and the 1.5 V/cm pulse, this value of p_0 is equal, while it is greater for the 3 V/cm pulse. This indicates that the EF cue shifts the regulation curve, which can be interpreted biologically as the EF enlarging the energetic budget that the tissue has available to spend on migration. Together, our results indicate that epithelial monolayers stimulated with an external signal exhibit a regulation mechanism that is similar to that of freely expanding epithelia. Moreover, the external signal influences the strength, but not the qualitative properties, of this regulation mechanism.

To understand whether the interplay between time-averaged mechanical power and force expenditure in tissues undergoing electrotaxis plays a role in establishing mechanical homeostasis (*i.e.*, maintaining that the tissue on average is not accelerating), we again compute the fraction of the area of the monolayer that is accelerating (decelerating) in time. Figure 3.4j shows that regardless of the stimulation protocol used, tissues quickly recover mechanical homeostasis by returning to a fraction of accelerating tissue of 0.5. Strikingly, it is now the 3 V/cm and 1.5 V/cm pulses that have an identical trace, with a difference with the 3 V/cm ramp. This suggests that the temporal dynamics of the EF are directly reflected through the active forces generated in the monolayer.

To conclude, we have shown that electrotaxis is an energetically costly mode of migration in which tissues operate far outside mechanical equilibrium. However, despite being subject to a sustained EF which induces active cell forces in the direction of the EF, and hence large energy expenditures, there still exists a relationship between average mechanical power and the rate of force production which regulates force expenditure when the average mechanical power exceeds a critical threshold. In the remainder of this chapter we will look in detail at how this interplay between average mechanical power and forces translate to the observed velocities in the experiment, both at a tissue scale, and at a single-cell scale.

3.5 The energetic history of monolayers predicts future tissue velocity

3.5.1 Energetic history predicts future tissue migratory behaviour

The main finding of this chapter so far is that the time-averaged mechanical work in a tissue influences future changes in force production. The question that arises naturally is whether those changes also translate into changes in the migratory velocity of tissue parcels during collective migration, whether that be during electrotaxis or free epithelial expansion. To understand the joint dynamics of velocity and power, empirical two-dimensional phase spaces of velocity and power can be computed. Since the EF is in the x-direction, we analyze the x-component of the velocity, which we denote by v. The time-averaged mechanical power is computed as before. These phase spaces show the v-p trajectories found in the experimental data set. For the computation of each empirial phase space, we consider a dataset that consists of points within a given spatiotemporal domain, *e.g.* all spatial locations between $t = 1$ hour and $t = 3$ hours for a 3V/cm pulse. We then compute the time derivative of the velocities and power using a first-order finite difference scheme. The time derivatives of velocities and power can be used directly to compute velocities within v-p space, which are then described by a vector field, $\mathbf{R}$, defined as

$$\mathbf{R}(\mathbf{x}, t) = \frac{\mathrm{d}}{\mathrm{d}t} \left(v(\mathbf{x}, t), p(\mathbf{x}, t) \right). \tag{3.19}$$

The computed velocities in v-p space can be used to generate phase plane plots graphically, using the MATLAB function `streamslice` [122]. The phase spaces graphically represent the evolution of velocity and power and allow us to interrogate and interpret any feedback between these quantities as a time-autonomous mechanism. The phase space for a representative freely expanding (control) tissue is shown in Figure 3.5a, and shows that trajectories diverge from the $v = 0$ line towards higher velocity and power. This phase space additionally reveals a global trend in which trajectories are restored towards lower speeds and subsequently lower energy expenditures, also consistent with our hypothesis that epithelia self-regulate when cells over expend.

Having understood how average mechanical power interacts with velocity in freely expanding epithelia, we turn to the question of how these phase spaces look for tissues undergoing electrotaxis. Electrotaxis creates a striking difference in phase spaces. The phase space for a representative tissue stimulated with a 3 V/cm pulse is shown in Figure 3.5b. Notice first that this phase space spans a substantially larger domain than the phase space for a freely expanding monolayer. Despite the fact that the EF provides a persistent cue for directed migration, the vertical direction of the field lines is reversed past

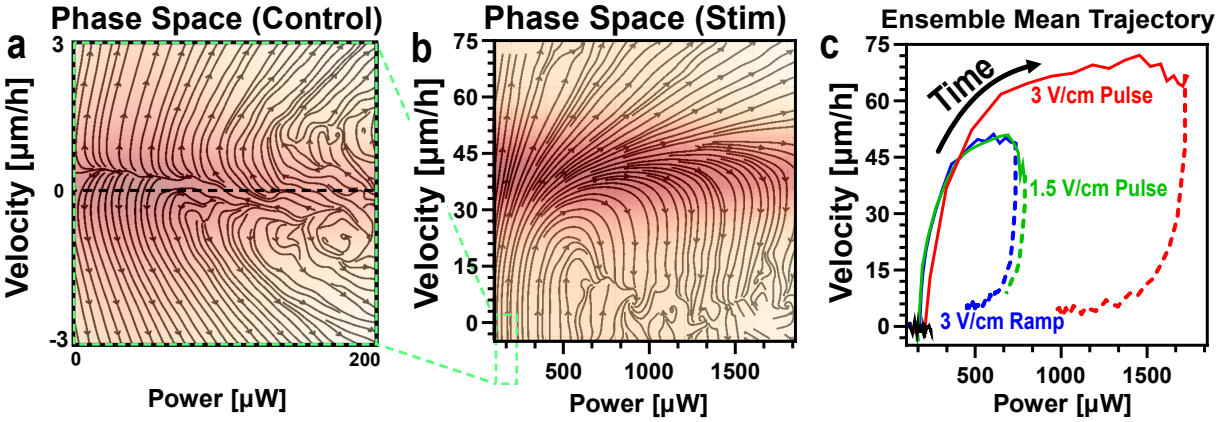


Figure 3.5: Dynamic behaviour of velocity and power within a collectively moving epithelium. **(a)** Power-velocity phase plane averaged over all control tissues. **(b)** Power-velocity phase plane averaged over all tissues under a 1.5V/cm pulse. The green box indicates extent of phase space for the control tissue in (a). **(c)** Trajectory of the ensemble mean during the course of the experiment across all experimental conditions. Note the rapidly decreasing rate of velocity increase as a higher power is attained. Please note changes in axis limits from panels **a-b**.

a certain level of average power, showing decreasing velocities beyond a critical power. This implies that as the cells approach this level of the critical power, there is a switch in their velocity. The two-dimensional histogram in the background of Figure 3.5b furthermore shows that the empirical distribution of tissue parcels in v-p space is concentrated around this switching point. Finally, the canonical behaviour of the centre of mass of the tissue accelerating in the experiment before slowing down and reaching a steady velocity is shown in Figure 3.5c, which displays how power expenditure rapidly increases during stimulation and decreases again rapidly when stimulation is ceased. Together, these data-driven phase spaces qualitatively describe how velocity and average mechanical power evolve jointly during epithelial migration, both during free (*i.e.*, unstimulated) epithelial expansion and during stimulation with an EF. These data-driven phase spaces demonstrate a strong relationship between the time-averaged mechanical power and velocity; despite the biological noise present in each biological replicate, the trends are well conserved across replicates. Strikingly, these trends appear to explain not only global dynamics, but also the behaviour of individual tissue parcels. For that reason, in the next section we will focus on the interplay between velocities and mechanical power at a cellular scale.

3.6 Power regulation dynamics of single cells

Having understood that, at the continuum scale, the time-averaged mechanical power and velocities are related through dynamics that can be described by a phase plane, it is straightforward to extend this question to the single-cell scale: do single cells show signs

of regulating their velocity in response to high average mechanical power expenditure? By using automated cell tracking, as described in Section 3.2.2, individual cell trajectories can be obtained for all of the tissues, in all experimental conditions. Since the mechanical model in Equation (3.17) leads to a continuum metric of average power, we need to relate the power as computed in Equation (3.17) to individual cells, depending on their position in time and space. This is done as follows. For each cell track, we use the location of the cell nucleus to compute the value of the mechanical power at that location, using linear interpolation of the power computed using Equation (3.17). Recall that the time-averaged power was computed using the material derivative, so each cell is assigned the average mechanical power of the tissue parcel it is contained in. For transparency, we note that the dimension of the PIV window corresponds to roughly three cell lengths. With each track now consisting of a location, velocity, and average mechanical power, we can explore the joint behaviours of velocity and power in single cells.

3.6.1 High and low power spending residence times follow a power law distribution

The first question of interest regards the temporal distribution of the times that cells spend in a state of high or low average mechanical power. To explore this, cells were classified as *high spenders* if their time-averaged mechanical power was higher than the average in the monolayer (and *low spenders* if lower). For fair comparisons between experimental conditions, all analysis was done for observations between $t = 1$ hour and $t = 3$ hours, *i.e.*, the times that correspond to the EF being turned on during the electrotaxis experiment. Additionally, in order to obtain an unbiased estimate of the switching probabilities between being in a high spending state or in a low spending state, we restrict our analysis to the tracks that contain cell locations for the entire stimulation window. As a first step to understand the dynamics of switching from a high to a low spending state, we computed the time that each cell spent between two switching events. These *residence times* show a striking empirical density, as shown in Figure 3.6.

The empirical distributions display fat tails, which suggests that the probability density of the residence times might be more similar to a power law than to an exponential distribution. To investigate the distribution of the residence times, the histograms of the residence times can be compared to a least squares fit of an exponential distribution, whose probability density function (pdf) is given by

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0, \tag{3.20}$$

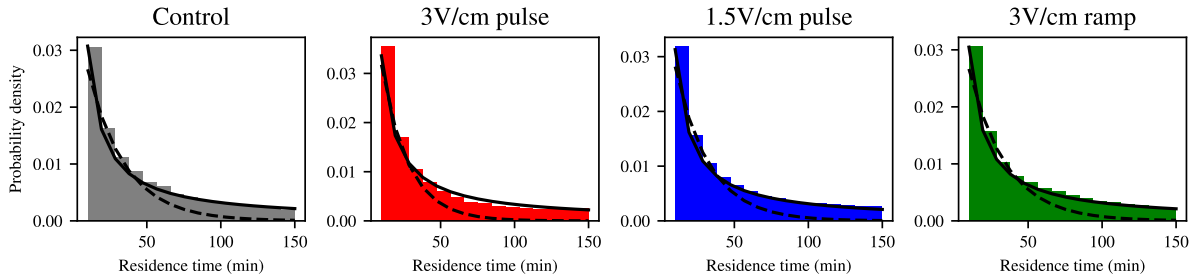


Figure 3.6: Empirical probability density function of the residence times, *i.e.*, the lengths of times spent being either a high or a low spender for each of the experiments, compared to linear squares fit for exponential distribution (dashed line) or power law distribution (solid line).

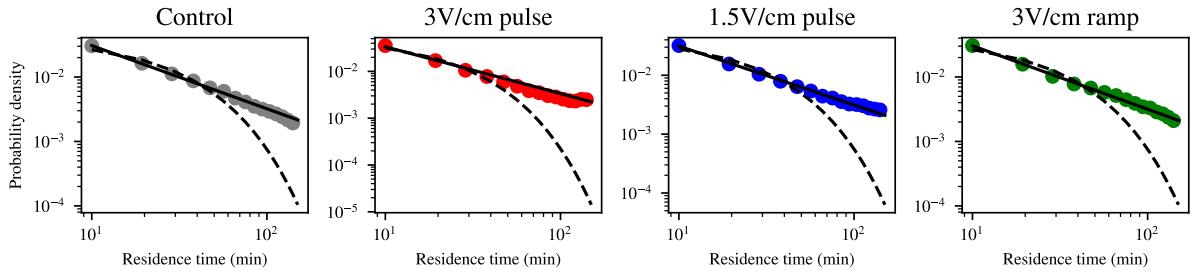


Figure 3.7: Log-log plot of empirical probability densities for the lengths of times spent being either a high or a low spender (scatter), compared to linear squares fit for exponential distribution (dashed line) or power law distribution (solid line).

where λ is the *rate*. We also consider a power law distribution, whose distribution is given by

$$f(x) = ax^{a-1}, \quad x > 0, a < 1, \tag{3.21}$$

where a is the *exponent*. It can be seen in Figure 3.6 that the power law distributions are in good agreement with the empirical distributions of the residence times, whereas the exponential distributions underestimate the probabilities that cells remain in the same state for a long times. To assess the goodness of fit of the power law distribution to the data, and to understand whether there is a good agreement with the data, we plot the empirical densities obtained from the histograms in Figure 3.6 on double-log axes in Figure 3.7. The log-log plots in Figure 3.7 demonstrate that there is excellent agreement between the empirical probability density of the residence times and the fitted power law distributions. Moreover, we find that the exponents for the least-squares fit to the empirical distributions of the residence times arising from the different stimulation traces and the control experiment are all very similar (Table 3.2). Strikingly, all distributions, regardless of EF strength, follow a very similar distribution, suggesting that the temporal dynamics of switching spending state are agnostic to external stimuli. This is surprising because it suggests that the EF stimulation protocol changes how much time-averaged power cells

Experiment	Computed exponent
No stim (control)	0.325
3 V/cm pulse	0.336
1.5 V/cm	0.313
3 V/cm ramp	0.319

Table 3.2: Computed power law exponents for the switching times pdf in the different experiments.

change, but it does not change how cells collectively organise their power expenditure. Since therefore the mechanism underlying this collective regulation appears unchanged between different EF stimulation protocols, the similarity between the distributions of the residence times points to mechanical factors inherent to epithelial migration, such as cell shape index and tissue fluidity, playing a role in influencing state switching. Finally, the fact the power law is a good fit to the empirical distribution of the times spent in either of the energetic states, and the exponential law is not, suggests that state switching is a random process with some memory in the system, as has previously been suggested in electrotaxing epithelia [1].

3.6.2 High and low spenders: velocities of single cells are regulated by average mechanical power during migration

The finding that single cells display a characteristic distribution of times spent in a high or low spending state begs the question: what effect does the energetic state of single cells have upon their force production? That is, do the continuum-scale regulatory mechanisms between velocities and time-averaged mechanical power translate to the single-cell level? We used the same cell tracks as those in the determination of power residence times in the previous section to determine which cells in the population are the top 50% and the bottom 50% spenders at $t = 2$ hours, *i.e.* after one hour of stimulation with the EF (for the control experiment of the freely expanding monolayer this is equally at $t = 2$ hours). The reason for choosing this time point in the experiment is that the peak average monolayer velocity during pulse stimulation is attained one hour after the EF is turned on, so it becomes a natural question to ask which cells have contributed the most, energetically, to that peak in velocity, and if this contribution was constant in time. *A priori*, one would expect that at the velocity peak, the velocity of high spending cells should exceed that of the low spending cells. Indeed, Figure 3.8 demonstrates that this is the case with the high spending cells leading migration. In all stimulation cases, including control, the high spending cells have the highest velocity at $t = 1$ hour post-stim (see Figure 3.8b-d). However, Figure 3.8b-d shows also that, as the experiment goes on, high spending cells

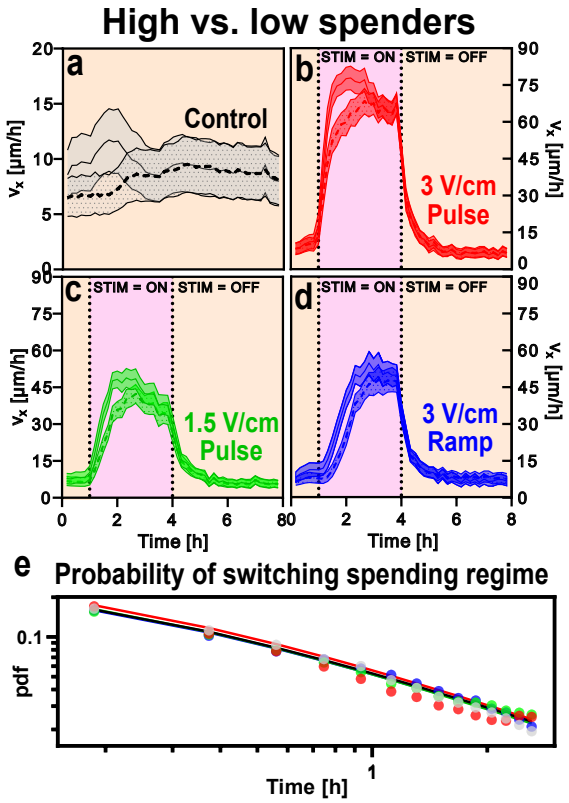


Figure 3.8: Speeds of high and low power spending cells during collective migration. Speed dynamics of 'high spenders' (solid lines) and 'low spenders' in tissues stimulated with **(a)** no electric field (control). **(b)** 3 V/cm pulse. **(c)** 1.5 V/cm pulse. **(d)** 3 V/cm ramp. **(e)** Overlayed probability density functions of switching spending regime (high to low, or vice versa) for all experiments. Points represent data, solid lines represent power law fits. Please note the changes in axis limits in panels **a**-**d**.

exhibit a dramatic slowdown, while the frugal spenders continue at approximately the same velocity throughout the experiment.

It is striking that in all plots in Figure 3.8b-d, the time at which the velocities of the high spending and low spending cells become equal is roughly the same, at the end of EF stimulation. We suggest that this might be the case due to the distribution of the times spent in either of the energetic regimes being the same across stimulation traces, as shown in the log-log plot of these residence shown in Figure 3.8e. The clear difference in migratory behaviour between high and low spenders during migration can be understood in the context of the times that cells spend in either a high or low migratory state. The highest spending cells at $t = 1$ hour after the onset of EF stimulation have likely produced the most active forces, and due to the fat tails of the power law distribution have done so, on average, for a considerable fraction of that first hour. This leads them to sustain a high power output and move quickly. However, as time goes on, these cells switch back to being low spenders, who, on average, will move more slowly than their surrounding cells. This causes their velocity to drop, and for the peak in velocity that they exhibited one hour into the experiment not to be sustained.

3.7 Regulation is dynamic in time and adapts to external stimuli

So far, we have established that epithelial migration exhibits a feature of closed-loop control in that force production is attenuated as a function of energy expenditure, both at the tissue-scale and at the individual cell scale. Given the rich dynamics of this regulation: the interplay between critical power, p_0, and mean power (Figure 3.3a-c), the dramatic shift of p_0 due to external stimulus (Figure 3.4i-j), the existence of deterministic trajectories in 'phase space' (Figure 3.5), and the existence of two obviously bimodal populations of 'high' and 'low spenders' (Figure 3.8), the final question of this chapter becomes: is regulation time-autonomous, or does it change in time? And, if so, how do external stimuli significantly affect that change?

One reason to suspect that regulation might be time dependent is that many cell types are known to adapt to constant stimuli, like during chemotaxis [123]. We hypothesise that electrotaxis might exhibit such adaptation due to the similarity in signaling pathways involved in electrotaxis and chemotaxis [124]. In this case, during the course of EF stimulation, one would expect the phase space of velocities and average mechanical power to progressively resemble that of the control experiment. Whether the phase spaces of tissues undergoing electrotaxis gradually resemble that of control conditions as the experiment goes on can be explored by computing separate phase spaces and regulation curves[1] at one hour intervals in a single representative tissue stimulated with a 3 V/cm pulse. The resulting phase spaces are shown in Figure 3.9a. During the one hour pre-stimulation, the phase space for a stimulated tissue is equal to that of a non-stimulated tissue – all trajectories flow towards higher powers and velocities (see Figure 3.5a). This suggests that the control phase space is also dynamic, as the first hour of a stimulation experiment is indistinguishable from a control experiment. Likewise the regulation curves for all experimental conditions are identical to that of the control experiment of the freely expanding monolayer.

During electrotaxis, however, the joint dynamics of velocity and average mechanical power are far from time-autonomous. While the phase space one hour into electrotaxis in Figure 3.9a is similar to that in Figure 3.5b, the phase space two hours into electrotaxis shows that a region appears in the phase plane where the entire tissue is decelerating, regardless of the magnitude of the average mechanical power. This feature is even more pronounced three hours after the onset of electrotaxis. How regulation changes in time is perhaps most apparent when observing the regulation curves in Figure 3.9b. Recall that these curves show the changes in force production as a function of average mechanical

[1]Recall that these are the relationships between jerk and average mechanical power, such as those shown in Figure 3.3 and Figure 3.4.

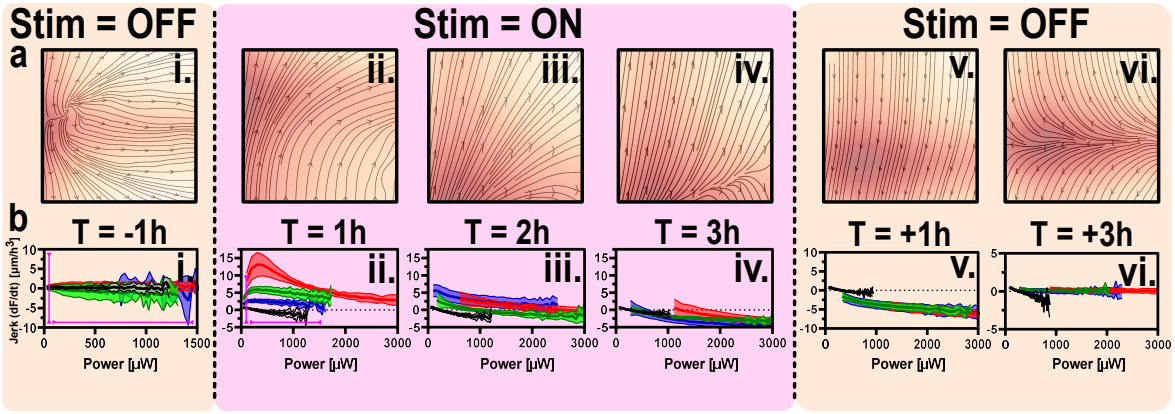


Figure 3.9: Dynamics of regulation during an external stimulus. **(a)** phase spaces for 1h intervals in a representative 3 V/cm stimulated tissue. **(b)** Jerk vs. power plots averaged over all tissues for the four different experimental conditions. Please note important changes in axis limits through the progression of i - vi.

energy. While the regulation curves conserve their shape during electrotaxis as time goes on, it can be seen that, for all experimental conditions, the critical power, p_0, *i.e.* the average mechanical power for which the jerk becomes negative, continuously shifts to the left. The biological interpretation of this is the energy budget available for responding to the electrotactic cue decreases in time. The apparent change in available energy budget can be interpreted as adaptation to the field. This is an important feature of collective migration during electrotaxis that needs to be accounted for in mathematical models of electrotaxis, which we will develop in Chapter 4.

Immediately after the EF is turned off, the corresponding phase space shows that, regardless of initial velocity or power, all trajectories lead to a decrease in both variables. This is in turn reflected by the dramatic shift of the data distribution towards lower velocities. Finally, the phase space during the last hour of the experiment appears similar to that of the first hour of the experiment, with a stereotyped flow towards the origin. Together, these data suggest that the regulation of active force expenditure as a function of time-averaged mechanical power is not time-autonomous but is regulated by processes at the signaling level, which determine the power threshold in response to the electric field.

3.8 Discussion

3.8.1 Understanding the biological origin of the discovered power regulatory mechanism

The main insight of this work is that time-averaged mechanical power plays a crucial role in influencing and regulating active cell forces during collective cell migration. This discovery begs the question of what the biophysical origin is of this regulatory mechanism. So far, we have shown that regulation of active cell forces happens as a function of time-averaged mechanical power. While regulation is shown to play an important role during migration, it is unknown which cellular and mechanical processes contribute to this mechanism. For example, is it the case that highly mechanically active regions of the monolayer are constrained by slower regions of the tissue by tissue collisions that slow them down, in turn leading them to experience less friction with the substrate? Or, alternatively, does there exist a cell-autonomous mechanism by which the active output of cells is, in fact, limited by energetic expenditure? These are two out of several possible models that can explain the dynamics of slow and fast moving regions within the monolayer, and can be answered only using new experiments. In this context, we propose that high-resolution microscopy can prove beneficial to tell these two different mechanisms apart, since by quantifying the mechanics of constituent cells and cell-cell interactions, these different hypothesis can be tested and interrogated.

3.8.2 Incorporating mechanical power in mathematical models of collective cell migration

The main hypothesis arising from this work is that time-averaged mechanical power plays a crucial role in influencing and regulating active cell forces during collective cell migration. For this reason, future mathematical models of collective cell migration should be extended to take these regulatory mechanisms into account.

A minimal mathematical model describing velocities, $\mathbf{v}$, and average mechanical power, p jointly could, for example, prescribe that the magnitude of the acceleration (and so the force per unit tissue mass) decreases whenever the power exceeds a critical power, p_0. Understanding the precise functional form relating how velocities depend on average mechanical power and *vice versa* presents an exciting opportunity to use data-driven methods to discover the structure of an appropriate model from experimental data. Furthermore, the role of different functional forms can be explored to understand the effect of various hypothetical regulation mechanisms on collective migration. This will bridge the gap between two of the aims set forth in the introduction to this thesis, namely

discovering new biophysical mechanisms – which we have done by discovering that there exists a regulatory mechanism that depends on the average mechanical power – and then translating this into viable models that have predictive power in diverse applications of epithelial collective cell migration.

Much like how this work considers cell migration at the continuum scale and the single-cell scale, mathematical models of collective cell migration can be formulated at different spatial scales. The continuum mechanics model developed in this chapter might be most easily understood and implemented in the context of other continuum models for collective cell migration. In such models, we propose that the average mechanical power can be coupled to the other state variables to introduce a regulatory mechanism that depends on the average mechanical power. Indeed, in the next chapter, Chapter 4, we will develop, calibrate, and validate a model for collective cell migration incorporating various functional forms of a possible regulatory mechanism to ultimately derive optimal control strategies for collective electrotaxis.

Besides continuum-scale models, agent-based models constitute another popular class of models of collective cell migration. Such models describe the behaviour of single cells and cell-cell interactions using cell-autonomous mechanisms. In such models, many of the emergent properties at the continuum limit, like the regulation mechanism discovered in this chapter, cannot be directly incorporated. Rather, it must be understood first which behaviours at the single-cell scale can give rise to the observed behaviours at the tissue scale. In order to do this, a careful study will be necessary to compute the detailed dynamics of single cells using cell-autonomous descriptions of their behaviours. In this arena, again, data-driven methods can be crucial to validate proposed cell-based models of collective electrotaxis, or to infer them directly from data, whenever the data are detailed enough to do so.

Chapter 4

Optimal temporal control of collective electrotaxis in epithelial monolayers

The work in this chapter uses the experimental data collected by Isaac Breinyn (Princeton University) in Chapter 3. My contributions to the work were the development of the continuum mechanics modelling, all computational implementations, and the writing of the results in this chapter.

4.1 Introduction

Electrotaxis is the process by which eukaryotic cells establish a cell polarity and move directionally in the presence of an electric field [39, 40, 124]. Electrotaxis is strongest in cellular collectives due to E-cadherin mediated adhesions [1, 3], and displays striking similarities between different cell types and electric field stimulation patterns. Previous studies have focused on using electrotaxis as a means to achieve particular patterns in collective movement [39,40]. For instance, electric fields can be programmed to give bespoke migration patterns [40], which are along the direction of the field lines in converging and diverging electric fields [39]. A common theme in all of these experiments is that the average direction of collective migration can be accurately controlled during electric field stimulation. The ease with which electric fields can establish directed migration in cellular collectives has led electrotaxis to be accepted as a robust method of steering collective migration [125]. The theoretical interest in modelling collective electrotaxis then lies in the ability to answer the following question: Given a desired spatio-temporal movement pattern, how should the electric field be designed to achieve this?

The adoption of electrotaxis as a predictable driver of collective migration has been stalled by a poor understanding of the temporal dynamics of the collective speed during

electric field stimulation. For example, Wolf *et al.* [1] found that MDCK collectives display a strong, but transient, response to the EF: while the average speed in the direction of the field increases quickly after the electric field is turned on, the collective starts to slow down after approximately one hour, even though the electric field strength remains constant. In general, the current lack of understanding of how the speed of the cellular collective will evolve during stimulation with an electric field poses difficulties in designing bespoke stimulation strategies that can be used to achieve a predictable migration outcome. This difficulty places the question of predicting collective speed during electrotaxis among the most pressing issues in collective electrotaxis. The problem we seek to address in this chapter is therefore how the collective velocity of MDCK monolayers during stimulation with a given electric field can be predicted, and how this knowledge could be exploited to design a stimulation strategy to produce a desired pattern of collective migration?

Both theoretically and experimentally, the question of how to control the temporal dynamics of the collective velocity has scarcely received attention in the literature, even though the fact that the temporal dynamics of electrotaxis are non-trivial is well-characterised. Controlling the temporal dynamics of collective velocity remains challenging, since the subcellular processes governing the evolution of cell polarity and active force production in response to stimulation with an electric field remain poorly understood [1, 39, 40, 124]. At the subcellular level, electrophoresis of charged membrane components creates a cell polarity that activates intracellular signalling pathways that are largely shared with those found in chemotaxis [124, 126, 127]. Subsequent modelling efforts [128] have been able to explain the directional movement of cells during electrotaxis, but the existing literature is not well positioned to explain the temporal dynamics of cell responses.

In this work, we are interested in constructing a deterministic, continuous-time model to predict the average velocity of the centre of mass of a cellular monolayer in response to stimulation with an electric field. To isolate the effect of speed from that of directionality, we will consider only uniaxial electric fields, *i.e.,* the field can be described solely by its field strength in the positive or negative direction, and the field direction is constant throughout the experimental setup. Under these assumptions, the goal of this chapter will be to derive a system of ODEs which predict the collective velocity and the active forces in the monolayer. A significant benefit of using ODE models is that there exists a completely developed theory on how to optimally control ODE solutions according to a given optimisation problem [12]. In the context of collective electrotaxis, this would mean that a well-calibrated and reliable model can be used to solve an optimisation problem, whereby the optimality condition is given by the desired temporal pattern of collective movement and the control is the electric field strength.

The current research on collective migration of epithelia provides a convenient starting point for the construction of such an ODE model – see [81] for a review of continuum models of collective epithelial migration. For example, since the internal signaling pathways

for collective electrotaxis are largely shared with chemotaxis, one can begin to model the internal cellular response to an electric field with a well-established continuum model for the response to a chemotactic signal. Erban and Othmer proposed an *adaptation-excitation model* for the strength of the chemotactic response of single cells, whereby cells are *excited* when initially exposed to an external signal, but slowly adapt over time when exposed for a longer period of time [123]. In their work, signal transduction is interpreted as "having two input pathways, an excitatory one that stimulates the [response], and one [...] which, in turn, shuts off the response". Similar models are frequently proposed to model chemotaxis. The advantage of such a model is that it is given by a straightforward system of scalar ODEs that describe the *effective signal* transduced within a cell upon contact with a chemoattractant [123]. This effective signal aims to capture the fact that the signal transduced within the cell is different, but related to, the real electric field that the cell is exposed to.

In this work, we will assume that similar internal signaling dynamics can be used to model the response to an applied electric field. By letting the strength of the active forces in the monolayer be described using the model of Erban and Othmer [123], we can derive a closed system of ODEs to describe the monolayer velocity. Using experimental data in conjunction with Bayesian inference, we can interrogate the validity and applicability of using this system of ODEs to model collective electrotaxis in response to stimulation with an electric field. Put together, the model can be used to develop a framework for designing electric fields that lead to desired outcomes in the form of an optimal control experiment.

This chapter will be structured as follows. In Section 4.1 we present the experimental data of collective electrotaxis that will be used for the calibration and validation of the ODE models. Then, in Section 4.3, we derive a basic system of ODEs that describes the evolution of the tissue velocity. We then use this system of ODEs to find the optimal control policy when a number of targets are considered, and compare our results to the current standards for electric field stimulation used in the literature. In Section 4.4 we present an extension to the model in Section 4.3 that takes into account not only the adaptation of the cells to the external field, but also the energetic and metabolic costs of collective migration. This section incorporates our experimental findings in Chapter 3 that point to an important role of energy expenditure in collective migration. We demonstrate that cellular responses to the electric field consist of metabolism and cellular adaptation on different time scales. Finally, we bring these insights to the problem of optimal control of collective electrotaxis in the last section, Section 4.5, where we consider the problem of how to optimally control both adaptation and cell exhaustion. Together, we offer a mathematical modelling framework to describe, understand, and optimise key cellular processes involved in the regulation of collective electrotaxis.

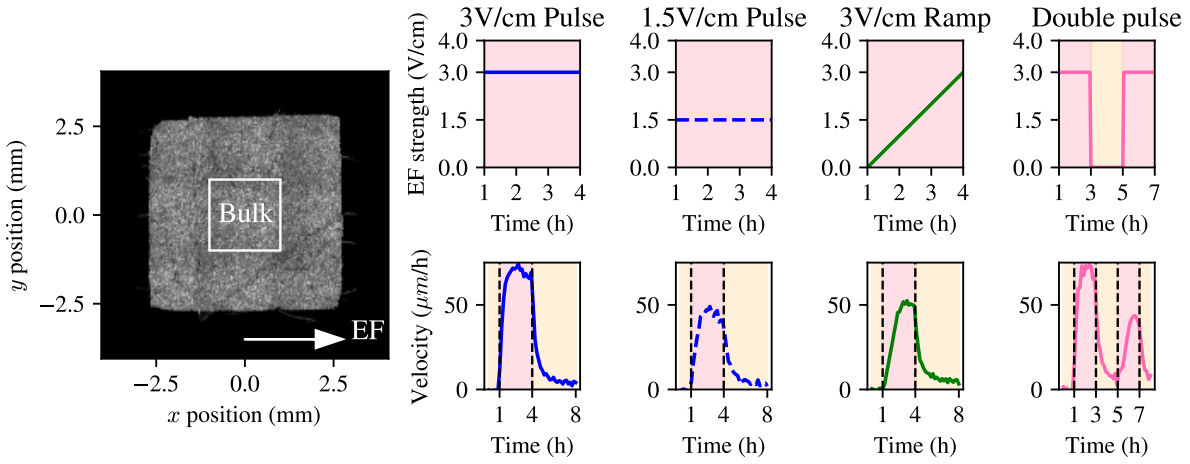


Figure 4.1: Experimental data of MDCK monolayer electrotaxis with different electric field stimulation protocols. Left panel: phase field image of an MDCK epithelial monolayer indicating direction of the uniaxial electric field and bulk region where the mean velocity is measured. Top row: electric field densities for each of the different experimental conditions. Bottom row: bulk velocities corresponding to the electric field traces in the top row.

4.2 Methods and data

In this section, we present the experimental data that will be used in this chapter. In brielectric field, all experiments involved electrotaxis of 5×5 mm^2 MDCK-II epithelial monolayers. Electrotaxis was induced by custom electro-bioreactors developed by the Cohen group at Princeton University, which are based on the SCHEEPDOG platform [40]. These bioreactors provide a continuous flow of media and can be programmed to generate electrical fields with custom electric field densities. Velocity data from each of the experiments were acquired using particle image velocimetry (PIV) on phase contrast images, where the cell nuclei are used as the fiduciary markers. Local tissue velocities were computed every 10 minutes. The PIV data provide a velocity field at a spatial resolution of 32×32 μm^2, which corresponds roughly to 3×3 cell lengths. For a sample velocity heat map, see Figure 4.1. For further details on the PIV data processing and implementation details in FIJI [98] and MATLAB [122], we refer the reader to [1].

4.2.1 Different stimulation traces

We investigated the effect of four different temporal stimulation patterns, each with a different electric field strength or temporal distribution of the charge. Data for three of these experiments, which are the 3V/cm pulse, 1.5V/cm pulse and 3V/cm linear ramp in Table 4.1, were obtained in the experiments in Chapter 3. We additionally performed an experiment with a double 3V/cm pulse for model validation in this study (see Table 4.1). In all studies, nine biological replicates were included for each of the experimental conditions,

as shown in Table 4.1. The electric field profiles and their corresponding data area are illustrated in Figure 4.1.

Experiment	Description
3V/cm electric field pulse	Constant electric field of 3V/cm between $t = 1$ hour and $t = 4$ hours, no electric field otherwise
1.5V/cm electric field pulse	Constant electric field of 1.5V/cm between $t = 1$ hour and $t = 4$ hours, no electric field otherwise
Ramp electric field to 3V/cm	electric field strength increasing linearly from 0V/cm at $t =$ 1 hour to $3V$/cm at $t = 4$ hours, no electric field otherwise
Double 3V/cm pulse	Constant electric field of 3V/cm between $t = 1$ hour and $t = 3$ hours, and between $t = 5$ and $t = 7$ hours, no electric field otherwise

Table 4.1: Overview of the different electric field stimulation experiments from Chapter 3 and additional experiments performed for this chapter.

4.2.2 Experimental methods

The double stimulation experiment in Table 4.1 was performed under the same experimental conditions as the experiments in Chapter 3, described in Chapter 3 and those performed by Wolf *et al.* [1]. Briefly, in each of these experiments, three 5×5mm^2 square tissues were seeded onto 10cm tissue-culture plastic dishes (CELLTREAT) using stencils of 250μm-thick silicone elastomer (Bisco HT6240) cut with a hobbyist razor-writer (Silhouette Cricut). Suspensions of MDCK cells were seeded into the stencil patterns at a volume of 10μL and density of 2.5 ($\pm$0.15) x 10^6 cells/mL. Cells were given one hour to adhere before the dish was flooded with media and incubated for 18 hours prior to stencil pull and insertion into the electric bioreactor. Tissue culture stencils were pulled approximately 1 hour before imaging began. Immediately post-stencil pull, dishes were refilled with fresh culture media followed by direct integration with the bioreactor. The assembled bioreactor is then placed into the microscope, situated within a custom-built cage incubator maintained at 37°C and perfused with fresh media continuously bubbled with 5% CO_2 using a peristaltic pump (Instech Laboratories) at a rate of 2.5 mL/h.

4.2.3 Data acquision at the tissue bulk

Experimental PIV data provides detailed spatiotemporal information on the velocities at a fine spatial resolution. Wolf *et al.* [1] previously used this detailed spatial resolution to characterise the responses of different tissue locations during collective electrotaxis. They found that the edges of the MDCK monolayer have a preferred direction of movement, which is outward of the tissue and is only observed within approximately 200μm of

the tissue boundary. This effect is a hallmark of the collective migration of MDCK cells [1, 39, 81, 106]. Although some attention has been given to understanding how the preferred migratory direction of the edges interferes with electrotaxis in the literature [1], it is not necessary to incorporate detailed mechanisms of edge migration when understanding the average velocity of the centre of mass of the epithelial monolayer. Since the majority of the tissue does not exhibit the preferred direction of migration that the edges show, we restrict our analysis to the *tissue bulk*, the 3×3 mm^2 region in the centre of the tissue characterised by Wolf *et al.* [1] – for a schematic, see Figure 4.1. We take a spatial average of the velocity in the bulk, which yields one-dimensional data traces that represent the monolayer velocity in each of the experiments.

4.3 An adaptation-excitation model of collective electrotaxis

In this section, we use a known model for cellular excitation and adaptation proposed by Erban and Othmer [123] to model the internal cellular response to electric field. Since all the experiments we study consider a uniaxial electric field, we describe the migration of the monolayer bulk using a one-dimensional velocity, v, which is defined in the positive direction of the electric field. The goal of this section is to connect a model of cellular adaptation and excitation to the temporal evolution of the bulk velocity in response to the electric field and furthermore, to seek to use this model to address the question of how to optimally control the collective migration of MDCK epithelia by varying the field strength.

4.3.1 Model derivation

We assume that cells experience a viscous friction force per unit mass which is proportional to their velocity with a friction coefficient, γ, such that the friction force per unit mass is given by $-\gamma v$. The assumption of linear friction is well-founded for epithelia, since this corresponds to the polymerisation of adhesions on the substrate [81]. We assume throughout that there will be a uniaxial electric field in the positive v-direction whose magnitude is given by a scalar signal, s. In what follows, we will introduce in detail the Erban and Othmer model [123] to describe, mechanically, how the electric field creates an intracellular *effective signal* to the cytoskeletal machinery to produce an active force in the direction of the electric field.

The canonical model for excitation-adaptation proposed by Erban and Othmer considers an external stimulus, s, an *effective signal*, s_{eff}, and an *inhibitor*, I. The effective signal, s_{eff}, is a scalar quantity that describes the extent to which an external signal is transmitted

within the cell. There is therefore a difference between the constant electric field supplied and the resulting internal signal. The inhibitor, in contrast, represents the strength of the internal signalling pathways that try to inhibit this response to the external stimulus. The dynamics are given by

$$\dot{s}_{\text{eff}} = \frac{s - (s_{\text{eff}} + I)}{\tau_e}, \tag{4.1}$$

$$\dot{I} = \frac{s - I}{\tau_a}. \tag{4.2}$$

In this model, one assumes that the excitation timescale, $\tau_e > 0$, is far shorter than the adaptation timescale, $\tau_a > 0$. The different timescales in this system create a characteristic response to a constant signal [123]: when the external signal, s, is a Heaviside function, the solution is of the form

$$s_{\text{eff}} = \frac{\tau_a}{\tau_a - \tau_e}\left(e^{-t/\tau_a} - e^{-t/\tau_e}\right), \quad I = 1 - e^{-t/\tau_a}. \tag{4.3}$$

With other words, as $t \to \infty$, the effective signal wanes, *i.e.*, $s_{\text{eff}} \to 0$. In particular, since $\tau_e \ll \tau_a$,

$$s_{\text{eff}} \sim Ce^{-\tau_a t}, \tag{4.4}$$

whenever $t \gg \tau_e$. This implies that on time scales larger than the excitation time, τ_e, the field response decays exponentially, which means that a constant electric field will not produce an internal signal that leads to migration beyond the initial excitation time, τ_e. In this work, we consider s, s_{eff}, and I to be unitless, as we rescale the field strength, s, by dividing by the canonical field strength of 3V/cm in the experimental set-up. Finally, we make the assumption that whenever $s \geq 0$, the effective signal, s_{eff}, equally satisfies $s_{\text{eff}} \geq 0$. The reason for making this additional assumption is that Equations (4.1) and (4.2) allow for s_{eff} to become negative when the field is turned off. MDCK epithelial monolayers are not observed to move in the direction opposite to the electric field when the electric field is turned off, and this phenomenon is otherwise not physically realistic. For this reason, going forward, unless explicitly noted, we consider the quantity

$$s_{\text{eff}}^{+} = \max(s_{\text{eff}}, 0). \tag{4.5}$$

To close the system of equations, we now relate the system of effective signal and inhibitor to the velocity. We assume that the active force per unit mass in the direction of the electric field is proportional to the effective signal. This assumption follows canonical models in the literature for coupling polarity and force magnitude [81], and results in the

following expression for the active force per unit mass, $\tilde{F}_{\text{active}}$,

$$\tilde{F}_{\text{active}} = \alpha s_{\text{eff}}^{+}, \tag{4.6}$$

where $\alpha > 0$ is a parameter that controls the responsiveness of the bulk to the input signal. Note that since s is unitless, α is given in units of $\mu\text{m/h}^2$. In this section, α is assumed to be a constant, but we will explore different functional forms for the field response in the next section. By then assuming that the resultant force on the bulk is given by the sum of the active force per unit mass and the friction force per unit mass, one obtains a single equation for the velocity,

$$\dot{v} = -\gamma v + \alpha s_{\text{eff}}^{+}. \tag{4.7}$$

4.3.2 Linear stability analysis

We begin by noting that Equations (4.1), (4.2), and (4.7) can be written as a linear system of equations such that

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} v \\ s_{\text{eff}} \\ I \end{pmatrix} = \begin{pmatrix} -\gamma & \alpha & 0 \\ 0 & -\tau_e^{-1} & -\tau_e^{-1} \\ 0 & 0 & -\tau_a^{-1} \end{pmatrix}\begin{pmatrix} v \\ s_{\text{eff}} \\ I \end{pmatrix} + \begin{pmatrix} 0 \\ \tau_e^{-1} s(t) \\ \tau_a^{-1} s(t) \end{pmatrix}. \tag{4.8}$$

We assume that the monolayer is stationary at the beginning of stimulation and the electric field is turned off, so that the initial condition is given by

$$v(0) = s_{\text{eff}}(0) = I(0) = 0. \tag{4.9}$$

Note that the eigenvalues of the matrix in the homogeneous part of the right-hand side (RHS) of Equation (4.8) are $-\gamma, -\tau_e^{-1}, -\tau_a^{-1}$, making the matrix invertible. At steady state, one has that the left-hand side of Equation (4.8) is zero, and that the steady state solution, $(v^\star, s_{\text{eff}}^\star, I^\star)$, for a constant field strength, $s(t) \equiv s$, is given by taking the inverse,

$$\begin{pmatrix} v^\star \\ s_{\text{eff}}^\star \\ I^\star \end{pmatrix} = \begin{pmatrix} -\gamma^{-1} & -\frac{\alpha\tau_e}{\gamma} & -\frac{\alpha\tau_a}{\gamma} \\ 0 & -\tau_e & \tau_a \\ 0 & 0 & -\tau_a \end{pmatrix}\begin{pmatrix} 0 \\ \tau_e^{-1} s \\ \tau_a^{-1} s \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ s \end{pmatrix}. \tag{4.10}$$

Therefore, Equation (4.8) has a single steady state. Since this steady state is given by zero velocity and zero effective signal, it can be interpreted as corresponding to a tissue

that has fully adapted to the electric field and exhibits no response. Moreover, this steady state is stable since all eigenvalues of the matrix on the RHS of Equation (4.8) are strictly negative. This explains the waning response with constant stimulation, as there is no steady state where a non-zero velocity is attained.

4.3.3 Bayesian inference of model parameters

In this section, our objective is to calibrate the simple adaptation-excitation model to the experimental bulk velocity data of Chapter 3 by performing Bayesian inference on the unknown model parameters $\gamma, \alpha, \tau_e, \tau_a$ from the available bulk data.

First, we note that the velocity decay parameter, γ, can be estimated directly from the data, since Equation (4.7) predicts that, given an initial velocity, v_0, the bulk velocity will follow an exponential decay with v_0 as its initial condition in the absence of an electric field. Since the experimental data contain information on the bulk velocity after stimulation with the electric field was turned off, it follows that the observed velocity decay can be used to immediately infer the decay rate, γ. From these considerations, we seek to fit an exponential of the form

$$v(t) = Ce^{-\gamma(t-t_{\text{end}})}, \tag{4.11}$$

to the bulk velocity data in the 3V/cm pulse stimulation case in the 2.5 hours after stimulation with the electric field is turned off. Here, t_{end} is the time at which the electric field is switched off, and C is given by

$$C = v(t_{\text{end}}). \tag{4.12}$$

We perform a least-squares fit for the decay rate, γ, to give an estimate of $\gamma = 1.765\text{h}^{-1}$. Figure 4.2 shows that the fitted exponential decay is in excellent agreement with the experimental data after pulse stimulation. This motivates us to take the value of γ as fixed going forward. The benefit of directly estimating γ from data is that it reduces the dimensionality of the parameter space, thus simplifying the inference of the remaining parameters.

We now keep $\gamma = 1.765\text{h}^{-1}$ fixed and perform Bayesian inference of the remaining parameters, α, τ_e and τ_a. We use Markov chain Monte Carlo with a Haario-Bardenet adaptive covariance and four chains under the assumption of a Gaussian error implemented in Python with the PINTS pacakge [129]. As a metric of convergence we use the $\hat{R}$ statistic [129], which summarises mixing and stationarity of the chains. We use the reference value of $\hat{R} = 1.05$ for four chains [27] and set a maximum number of MCMC iterations to $2 \cdot 10^4$. To make sure that the inferred model parameters yield model predictions for the various different stimulation conditions, we use as data the velocity

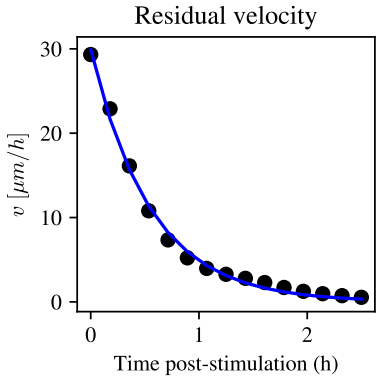


Figure 4.2: Experimental data of bulk velocity decay post-stimulation with an electric field pulse of 3 V/cm (black dots) together with least-squares fit of exponential decay in the form $Ce^{-\gamma t}$ (solid blue line). The least-squares solution gives an excellent fit to the data and can be used to identify the decay rate, γ.

profiles from the 1.5 V/cm pulse and the 3 V/cm ramp experiments from Chapter 3. The MCMC procedure yields a well-identified posterior distribution, with marginal posterior means 149.92 μm/h^2, 0.260h, and 2.038h, for α, τ_e and τ_a, respectively. These posterior means are in good agreement with previous intuition that the time scale for adaptation is longer than the time scale for excitation, as well as with the data, suggesting that the bulk slows down on a timescale of the same order of magnitude as τ_a. Plots of the marginal posterior distributions of the model parameters are shown in Figure 4.3. These show that the parameters can be confidently identified from the data, given their well-defined posterior distributions.

To assess the predictive capability of the model given the posterior distribution of the model parameters, we compute the posterior predictive intervals of the model, *i.e.* a 95% confidence interval for the model outputs given the posterior distribution of the model parameters. These posterior predictive intervals are shown in Figure 4.3. It can be seen that the posterior predictive intervals are tightly concentrated around the mean, showing little model uncertainty in predicting the experimental data, given the posterior distribution of the model parameters. By overlaying the posterior predictive intervals for the model onto the experimental data, we observe that the model predictions in both the pulse-stimulation and the ramp-stimulation case are in good agreement with the experimental data and lie comfortably within the standard deviation of the experimental data. We conclude that the adaptation-excitation model presented in this section, when calibrated to the experimental data, is able to recapitulate the phenomenon of attenuating velocity during electrotaxis to the field when the electric field is held constant and reaching a plateau when the electric field is increased during the course of the experiment.

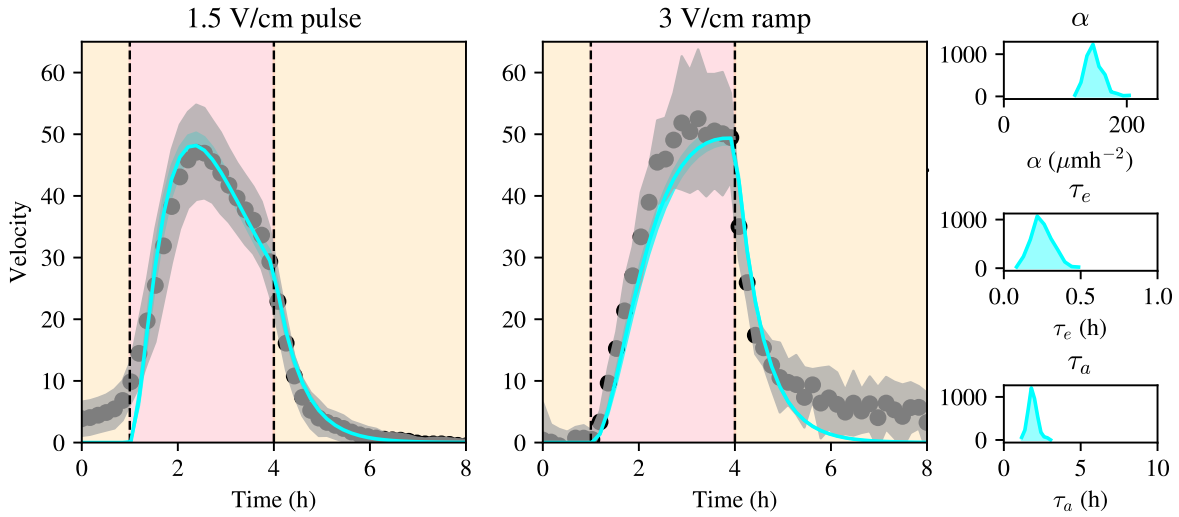


Figure 4.3: Bayesian inference on a simple model of adaptation and excitation given by Equation (4.8) using experimental data from an electrotaxis experiment. Left: bulk velocity using a constant field of 1.5 V/cm. Posterior mean of the model together with a confidence interval between the 5% and 95% quantile of the model posterior predictions in cyan, experimental data and their confidence interval in grey. Notice the small width of the posterior predictive intervals. Middle: bulk velocity using a linear ramp from 0 to 3 V/cm. Right panels: posterior distributions for model parameters, α, τ_e, τ_a, respectively.

4.3.4 Optimal control

Having established that excitation-adaptation dynamics can be used to describe the velocity of the tissue bulk during electrotaxis, both when the electric field strength is held constant, and when it is ramped in time, we turn to the question of designing a stimulation protocol that is in some way optimal for the experiment. We previously established that the question of how to design electric field stimulation protocols has received little attention due to the difficulties in predicting tissue response *a priori*. Instead, experiments have been carefully calibrated through trial and error by the experimental laboratories that have performed them over the past few years.[1]

In this section, we will consider the problem of deriving an optimal pattern of stimulation using the adaptation-excitation model. We will consider that the optimal control problem can be cast as an optimization problem of a functional, E, that depends continuously on the velocity as well as the signal,

$$E(v, s) = \int_0^T f(v, s)\mathrm{d}t, \tag{4.13}$$

and that there is a fixed start and end time of stimulation, given by 0, T, respectively.

[1]It is worth noting that these experiments are costly and time-consuming, making a grid search or other brute-force methods to determine optimal temporal stimulation patterns unappealing.

Furthermore, we need to account for the total amount of electric current administered during the experiment, as applying large currents throughout the cell medium for large amounts of time yields damage to the tissue and the experimental set-up. For now, we will incorporate this constraint by requiring that a proxy of the total amount of charge delivered to the tissue during stimulation remains constant, *i.e.*

$$\int_0^T s(t)^2 \mathrm{d}t = s_{\text{budget}}. \tag{4.14}$$

Later in this section, we will investigate the role of introducing lower and upper bounds on the field strength instead. Note also that in Equation (4.14) we have introduced a quadratic to account for the fact that the electric field can be positively or negatively charged, and that fields of both signs will affect the tissue. Under the condition of Equation (4.14), we set up the first optimal control problem to maximise the total distance travelled by the tissue. This is a straightforward question to ask when one is concerned with increasing the controllability of collective migration. We therefore consider the optimization problem

$$\max_S \int_0^T |v(t)| \mathrm{d}t = \max_S \int_0^T \sqrt{v^2(t)} \mathrm{d}t, \tag{4.15}$$

which, as the integrand is monotonic in the velocity, can be cast as

$$\max_S \int_0^T v^2(t) \mathrm{d}t, \tag{4.16}$$

for mathematical convenience. This optimisation problem is then subject to the condition in Equation (4.14). In optimal control theory, this is known as an *isoperimetric problem*. In this formulation, we introduce a new variable,

$$z(t) = \int_0^t s(t)^2 \mathrm{d}t, \tag{4.17}$$

such that the problem can be described as

$$\max_S E = \int_0^T v^2 \mathrm{d}t, \quad \text{subject to} \tag{4.18}$$

$$\dot{z}(t) = s(t)^2, \tag{4.19}$$

$$v(0) = 0, \tag{4.20}$$

$$z(0) = 0, \quad z(T) = S_{\text{budget}}. \tag{4.21}$$

Recall that the isoperimetric contraint above has been introduced to enforce a physical limit on the amount of charge to which cells in the monolayer are exposed. One might also interpret the isoperimetric constraint as a manner of interrogating how, given a fixed amount of charge, one might optimally distribute the delivery of this charge so that the total distance travelled by the monolayer is optimal. To make direct comparisons with the experiments performed in Chapter 3 and Wolf *et al.* [1], we wish to make the stimulation pattern have the same integrated charge for constant stimulation at 1.5 V/cm, so we set $s_{\text{budget}} = 6.75\text{V}^2\text{h/cm}^2$. For this problem the Hamiltonian is given by

$$H = v^2 + \lambda_v(t)\dot{v} + \lambda_{s_{\text{eff}}}(t)\dot{s}_{\text{eff}} + \lambda_I \dot{I} - \mu_z \dot{z} \tag{4.22}$$

$$= v^2 + \lambda_v(t)\left[-\gamma v + \alpha s_{\text{eff}}\right] + \tau_e^{-1}\lambda_{s_{\text{eff}}}(t)\left[s - s_{\text{eff}} - I\right] + \tau_a^{-1}\lambda_I(t)\left[s - I\right] - \mu_z s^2, \tag{4.23}$$

where $\mathrm{d}\lambda_\bullet/\mathrm{d}t = -\partial H/\partial \bullet$, such that

$$\frac{\mathrm{d}\lambda_v}{\mathrm{d}t} = -\frac{\partial H}{\partial v} = -1 + \gamma\lambda_v, \tag{4.24}$$

$$\frac{\mathrm{d}\lambda_{s_{\text{eff}}}}{\mathrm{d}t} = -\frac{\partial H}{\partial s_{\text{eff}}} = -\alpha\lambda_v + \tau_e^{-1}\lambda_{s_{\text{eff}}}, \tag{4.25}$$

$$\frac{\mathrm{d}\lambda_I}{\mathrm{d}t} = -\frac{\partial H}{\partial I} = \tau_e^{-1}\lambda_{s_{\text{eff}}} + \tau_a^{-1}\lambda_I, \tag{4.26}$$

$$\frac{\mathrm{d}\mu}{\mathrm{d}t} = -\frac{\partial H}{\partial z} = 0, \tag{4.27}$$

subject to the terminal conditions $\lambda_\bullet(T) = 0$. Additionally, $\mu = \mu_\star$ is a constant. The optimality condition is given by

$$0 = \frac{\partial H}{\partial s} = \tau_e^{-1}\lambda_{s_{\text{eff}}} + \tau_a^{-1}\lambda_I - 2\mu_\star s. \tag{4.28}$$

The solution to Equation (4.28) yields the optimal stimulation pattern, $s^\star$, which is given by

$$s^\star = \frac{1}{2\mu_\star}(\tau_e^{-1}\lambda_{s_{\text{eff}}} + \tau_a^{-1}\lambda_I) = \frac{\tau_e^{-1}}{2\mu_\star}\left(\lambda_{s_{\text{eff}}} + \frac{\tau_e}{\tau_a}\lambda_I\right) = \frac{\tau_e^{-1}}{2\mu_\star}\left(\lambda_{s_{\text{eff}}} + \epsilon\lambda_I\right), \tag{4.29}$$

where $\epsilon = \tau_e/\tau_a$. To find the optimal stimulation pattern, $S^\star$, one must seek to solve the linear inhomogeneous system of equations for $\lambda_\bullet$. Note that Equation (4.24), the equation for λ_v, decouples from the other adjoint equations, so we can solve directly to find

$$\lambda_v(t) = \frac{(\gamma\lambda_v(0) - 1)e^{\gamma t}}{\gamma} + \frac{1}{\gamma}. \tag{4.30}$$

Applying the terminal condition, $\lambda_v(T) = 0$, yields that $\lambda_v(0) = -(e^{-\gamma T} - 1)/\gamma$, and so

we find

$$\lambda_v(t) = (e^{\gamma(t-T)} - 1)/\gamma. \tag{4.31}$$

Direct substitution of the solution for λ_v into Equation (4.25), the equation for $\lambda_{s_{\text{eff}}}$, then yields

$$\lambda_{s_{\text{eff}}} = \frac{\alpha}{\gamma(\gamma - \tau_e^{-1})} - \frac{\alpha}{\tau_e^{-1}(\gamma - \tau_e^{-1})} + c_1 e^{\tau_e^{-1} t} - \frac{\alpha e^{\gamma(t-T)}}{\gamma(\gamma - \tau_e^{-1})} \tag{4.32}$$

$$= -\frac{\alpha}{\tau_e^{-1}\gamma} + c_1 e^{\tau_e^{-1} t} - \frac{\alpha e^{\gamma(t-T)}}{\gamma(\gamma - \tau_e^{-1})}. \tag{4.33}$$

We can now fit the terminal condition, $\lambda_{s_{\text{eff}}}(T) = 0$, such that

$$0 = \lambda_{s_{\text{eff}}}(T) = -\frac{\alpha}{\tau_e^{-1}\gamma} + c_1 e^{\tau_e^{-1} T} - \frac{\alpha}{\gamma(\gamma - \tau_e^{-1})}, \tag{4.34}$$

which gives

$$c_1 = e^{-\tau_e^{-1} T}\left(\frac{\alpha}{\tau_e^{-1}\gamma} + \frac{\alpha}{\gamma(\gamma - \tau_e^{-1})}\right) = \frac{\alpha\gamma}{\tau_e^{-1}(\gamma - \tau_e^{-1})} e^{-\tau_e^{-1} T}. \tag{4.35}$$

Therefore,

$$\lambda_{s_{\text{eff}}} = \frac{\alpha\gamma}{\tau_e^{-1}(\gamma - \tau_e^{-1})}\left(e^{\tau_e^{-1}(t-T)} - \frac{\tau_e^{-1}}{\gamma^2} e^{\gamma(t-T)}\right) - \frac{\alpha}{\tau_e^{-1}\gamma}. \tag{4.36}$$

We can then find λ_I by solving Equation (4.26),

$$\lambda_I = e^{-\tau_a^{-1} t} \int_0^t \lambda_{s_{\text{eff}}}(\xi) e^{\tau_a^{-1}\xi} \mathrm{d}\xi + C_I e^{\tau_a^{-1} t}, \tag{4.37}$$

where C_I is determined by setting the terminal condition, $\lambda_I(T) = 0$, which gives

$$C_I = -e^{-\tau_a^{-1} T} \int_0^T \lambda_{s_{\text{eff}}}(\xi) e^{\tau_a^{-1}\xi} \mathrm{d}\xi. \tag{4.38}$$

Therefore,

$$S^\star = \frac{\tau_e^{-1}}{2\mu^\star}\lambda_{s_{\text{eff}}} + \epsilon\frac{\tau_e^{-1}}{2\mu^\star}\left[e^{-\tau_a^{-1} t} \int_0^t \lambda_{s_{\text{eff}}}(\xi) e^{\tau_a^{-1}\xi} \mathrm{d}\xi + C_I e^{\tau_a^{-1} t}\right]. \tag{4.39}$$

To determine the solution of the system, it remains to find the constant $\mu^\star$ such that the optimal solution integrates to S_{budget}. We use the marginal posterior means for the model parameters α, τ_e, τ_a to first compute $S^\star$ numerically and use numerical integration to determine the integral during the stimulation window. We show a prediction for the optimal stimulation pattern in Figure 4.4. Interestingly, the optimal stimulation pattern

begins at a nonzero electric field value and increases for the first half of the stimulation window, when it exceeds the constant electric field strength of 1.5V/cm, before dropping below that field strength in the second half of the stimulation. The optimal stimulation pattern found in Equation (4.39) shows only a modest increase in the distance travelled by the bulk of the monolayer compared to either the constant 1.5V/cm pulse or 3V/cm ramp experiments from Table 4.1, since the integrated numerical solution of the velocity is 2.17% greater than the distance travelled using a constant electric field (and 19% greater than that of ramp stimulation). This reflects the fact that electrotaxis experiments have been iterated many times experimentally, and stimulation with a constant electric field seems to be a sensible stimulation strategy. Strikingly, the optimal stimulation pattern also results in a bulk velocity that still exhibits a slowdown, which runs counter to the view in the electrotaxis community that the slowdown of the bulk during stimulation represents nonoptimality of the stimulation procedure.

We incorporate parameter and prediction uncertainty by computing the posterior distribution over the optimal control solution. We take $2 \cdot 10^3$ samples from the MCMC posterior distribution and for each sample compute the optimal solution according to Equation (4.39) using that given parameter combination and recomputing the normalising constant, $\mu^\star$, for each new parameter combination to obtain a normalised density. We again compute the 5% and 95% percentiles of the velocity distribution at each time point and plot them in Figure 4.4. We see that the posterior distribution is tight around the posterior mean, which confirms that the above analysis holds true across the entire posterior distribution.

4.3.5 Optimal control for the terminal velocity

Optimising for the total distance travelled shows that there is only a very moderate theoretical improvement in the optimised outcome compared to stimulation with a constant electric field of 1.5 V/cm. A striking feature of the optimal stimulation curve shown in Figure 4.4 is that the velocity curve under the optimal stimulation protocol, like the pulse stimulation, displays a diminishing response to the field at later time points. At the same time, the velocity in the case of the linear ramp increases and plateaus at its maximum velocity at the end of stimulation. Therefore, we pose the additional question of how to maximise the terminal velocity of the tissue, in other words: How can the stimulation be designed so that the bulk velocity at the terminal time, T, is maximal? *A priori*, one would expect this to be a different stimulation pattern than the optima found previously. We change the objective function so that we maximise the terminal velocity, $v(T)$,

$$E(v) = \int_0^T \dot{v} \mathrm{d}t = \int_0^T [-\gamma v + \alpha s_{\text{eff}}] \mathrm{d}t. \tag{4.40}$$

Note that this change in objective function only changes the equations for $\mathrm{d}\lambda_v/\mathrm{d}t$ and $\mathrm{d}\lambda_{s_{\text{eff}}}/\mathrm{d}t$ in the adjoint equations, Equations (4.24)-(4.27), since the only additional terms introduced into the Hamiltonian include the velocity, v, and the effective signal, s_{eff}. We obtain

$$\frac{\mathrm{d}\lambda_v}{\mathrm{d}t} = \gamma + \gamma\lambda_v, \tag{4.41}$$

$$\frac{\mathrm{d}\lambda_{s_{\text{eff}}}}{\mathrm{d}t} = -\alpha\lambda_v + \tau_e^{-1}\lambda_{s_{\text{eff}}}, \tag{4.42}$$

such that, directly fitting the terminal condition, $\lambda_v(T) = 0$, we obtain

$$\lambda_v = e^{\alpha(t-T)} - 1, \tag{4.43}$$

which we substitute into Equation (4.25) to find

$$\lambda_{s_{\text{eff}}} = \frac{\alpha}{\alpha - \tau_e^{-1}} \left(e^{\tau_e^{-1}(t-T)} - e^{\alpha(t-T)} \right). \tag{4.44}$$

The same computation as in Equation (4.39) now yields the optimal electric field, which results in a terminal velocity of 62.9 μm/h compared to 52.5μm/h for a constant field. The electric field that optimises the terminal velocity is shown in Figure 4.4. Note that, like in the linear ramp case, the field strength is monotonically increasing to its maximum intensity at the end of stimulation. We obtain a higher terminal velocity than in the other experiments, at the cost of a very high field strength at the end of stimulation, approximately 6 V/cm, which is physiologically and experimentally feasible, albeit not for long periods of time. The fact that the optimal control problem for the terminal velocity yields a ramping pattern suggests that total distance travelled and terminal velocity seem like two incompatible requirements to fulfill simultaneously, given the model and evidence presented. Finally, to address parameter and prediction uncertainty, we repeat the computation of posterior distributions over the velocities, as done in Section 4.3.4, which we display in Figure 4.4. Again, the posterior distribution of the velocities is tight around the posterior mean.

4.3.6 Optimal control for constant velocity

The optimal control problems for the maximum distance travelled and the maximum terminal velocity seem to suggest that maximising the distance comes at the cost of a decrease in velocity during stimulation, while optimising the terminal velocity does not maximise the total distance travelled. The finding that the optimal stimulation pattern for

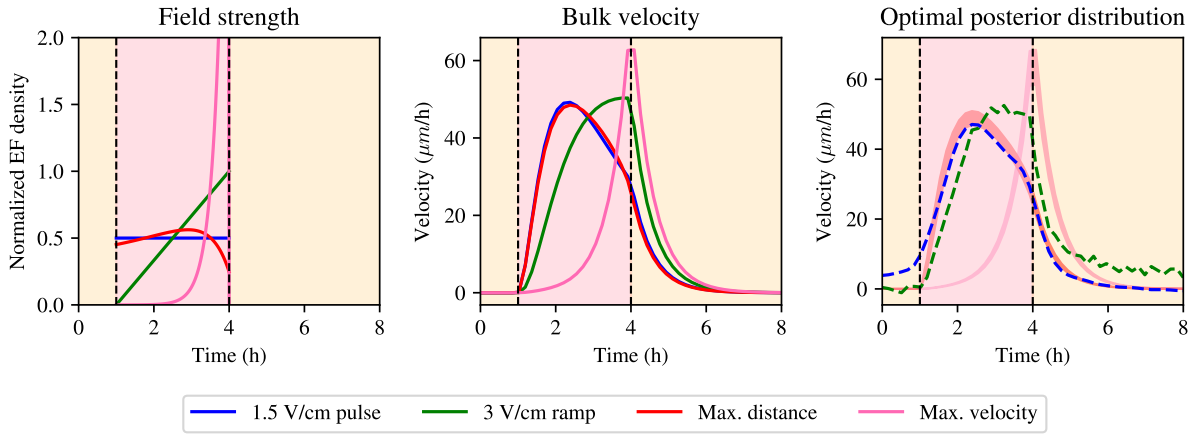


Figure 4.4: Stimulation patterns for optimal control of collective electrotaxis. In all plots, red corresponds to maximum tissue displacement, pink corresponds to maximum terminal velocity. Blue and green curves represent pulse and ramp stimulation, respectively. Left: the different stimulation patterns in normalized units. Middle: corresponding velocity curves using the different stimulation patterns. Right: data from 1.5V/cm pulse (dashed line, blue) and ramp (dashed line, green), compared to posterior distribution for optimal solution for distrance travelled (red) and terminal velocity (pink).

total distance travelled does not yield a nondecreasing bulk velocity during the course of stimulation with the electric field runs counter to intuition in the electrotaxis community, as the current opinion in the field tends to view a drop in bulk velocity as a source of nonoptimality in stimulation. From an experimental viewpoint, controlling field strength so that migratory velocity is constant is attractive, since electric fields are often used for 'cruise control'. Therefore, we set up a final optimal control problem designed to keep variations in the velocity minimal. We do this by once more modifying the functional so that we seek solutions that minimise

$$E(\dot{v}) = \int_0^T \dot{v}^2 \mathrm{d}t = \int_0^T (-\gamma v + \alpha s_{\text{eff}})^2 \mathrm{d}t. \tag{4.45}$$

As in the previous section, the only adjoint equations affected are those for λ_v and $\lambda_{s_{\text{eff}}}$. These become

$$\frac{\mathrm{d}\lambda_v}{\mathrm{d}t} = 2\gamma(-\gamma v + \alpha s_{\text{eff}}) + \gamma\lambda_v, \tag{4.46}$$

$$\frac{\mathrm{d}\lambda_{s_{\text{eff}}}}{\mathrm{d}t} = -2\alpha(-\gamma v + \alpha s_{\text{eff}}) - \alpha\lambda_v + \tau_e^{-1}\lambda_{s_{\text{eff}}}. \tag{4.47}$$

In addition, we impose a terminal condition for the velocity, so that we prescribe a desired velocity at the end of stimulation, *i.e.*, $v(T) = v^\star$ for some pre-determined constant, $v^\star$, which can be set by the practitioner in accordance with a practical need. Together with Equation (4.8), Equations (4.46), (4.47) and the usual transversality conditions,

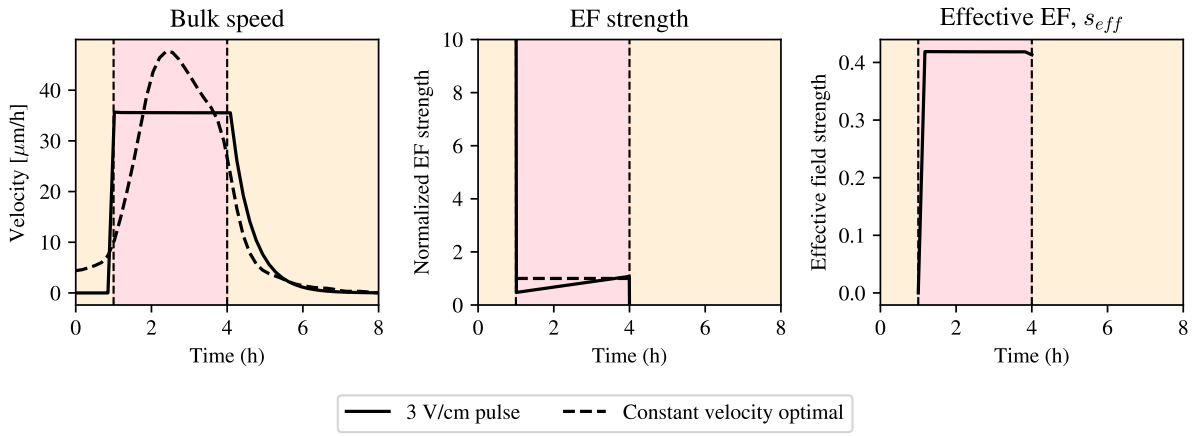


Figure 4.5: Solution to optimal control problem for constant bulk speed during stimulation with an electric field. Left: bulk speed for constant velocity (solid line) plotted against experimental data with constant electric field pulse (dashed line). Middle: numerical solution of the optimal control problem for constant velocity (solid line) plotted against reference normalized electric field strength (dashed line). Right: effective signal, s_{eff}.

$\lambda_\bullet(T) = 0$, the optimisation of the function in Equation (4.45) creates a nonlinear first-order boundary value problem (BVP). The complicated coupling between the equations requires a numerical solution to the problem. In this simulation, we set $v^\star$ so that it is equal to the average bulk velocity during stimulation with a single pulse of 3 V/cm.

For the numerical solution of the BVP, we use the bvp4c scheme of Kierzenka *et al.* implemented in Python through the SciPy package [130]. This scheme implements a fourth-order collocation algorithm with a control of the residuals and uses a damped Newton method with an affine-invariant criterion function. For the implementation, we set a maximum relative tolerance of 10^{-3}, an absolute boundary value tolerance of 10^{-6} and specify a maximum of 10^6 nodes for the collocation algorithm.

Figure 4.5 shows that the electric field profile needed to produce a constant velocity profile is very strong initially and rapidly erodes to create a constant effective signal, s_{eff}. In fact, the field strength predicted here is 75 V/cm, which is not physiologically realistic, as it would lead to toxicity and rapid cell death. This finding suggests that given the physiological constraints of the problem, it is not possible to design an optimal control experiment that creates a constant speed profile with the same average velocity as those found during pulse stimulation. This large initial peak in electric field strength arises from the requirement that the tissue hits a constant velocity quasi-instantly upon stimulation. This finding invites the question of which constant velocity values are feasible to achieve given that the solution profile needs to be realistic, *i.e.* given the physical constraints on the electric field strength. To this end, we perform a grid search, whereby we solve the optimal control problem for a range of target velocities, $v^\star$, ranging from 0 to 40 μm/h. For each of these terminal velocities, we solve the BVP for optimal control and compute

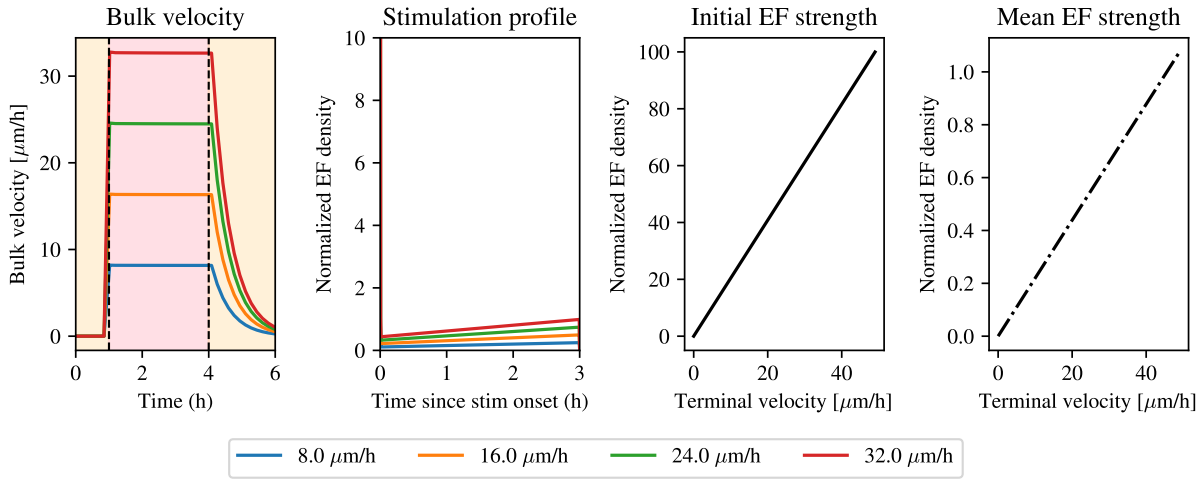


Figure 4.6: Stimulation profiles and electric field strengths for varying terminal velocities. From left to right. A: Bulk velocity when using optimal control solution for constant stim using different values of terminal velocity (see legend). B: Optimal stimulation profile corresponding to the different terminal velocities in panel A. C: electric field strength at the start of stimulation. D: electric field strength averaged during stimulation. Note the difference in the magnitude of the axis between C and D.

.

the resulting bulk velocities and optimal stimulation strategy. We summarise the results in Figure 4.6.

Figure 4.6 shows that at each of the values for the terminal velocity, $v^\star$, the model predicts that an optimal stimulation strategy can be designed so that the bulk moves at this constant speed. Each of these stimulation profiles has the same shape, whereby a very large current is applied at the beginning of the experiment, which slowly decreases during the course of the experiment. We investigate the relationship between the desired terminal velocity and the field strength at the onset of stimulation as required by the model. We see that there is a linear increase in the normalised electric field density and that the initial electric field strength is prohibitively large for velocities higher than 5 μm/h. This is surprising and experimentally frustrating given that the mean electric field strength predicted by the model only shows a marginal increase with the desired terminal velocity, whereby all terminal velocities experience a normalised electric field strength that is physiologically tolerable. We conclude, thus, that designing a realistic electric field to control electrotaxis at constant velocity, while keeping the electric field strength within realistic bounds, is not possible.

Since the source of the unphysically strong electric field at the start of stimulation is the requirement that the field reaches the desired velocity immediately at the start of stimulation, we ask if we can stimulate ignoring the non-realistic initial prediction by setting a maximum field strength, since the velocity in the period after the onset of stimulation is indeed constant. We expect this might come at the cost of reaching the

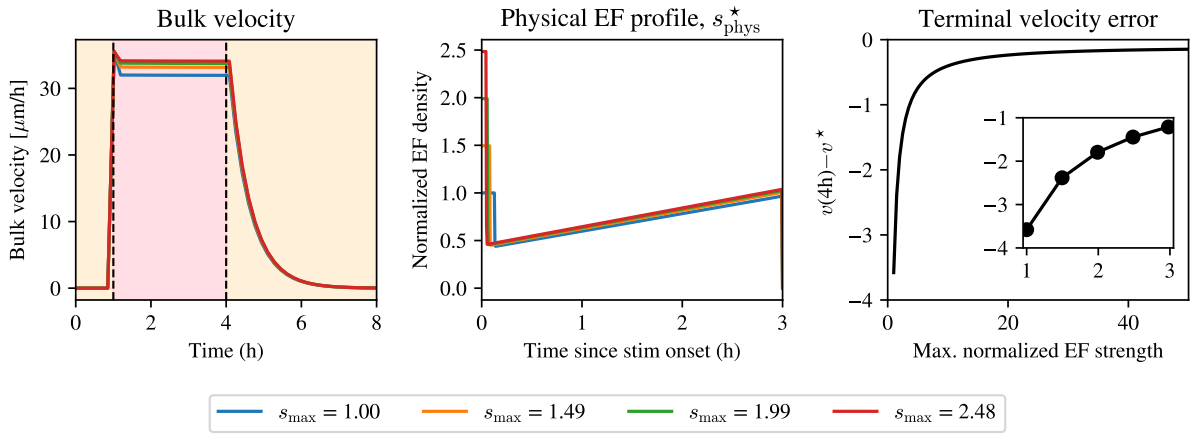


Figure 4.7: Optimal stimulation to keep the velocity constant during stimulation. Left: bulk velocity using different maximum electric field strengths (see legend). Middle: physical electric field profile showing similar quantitative solutions to the field. Right: terminal velocity error decreases when the maximum normalized electric field strength increases.

target velocity at the beginning of stimulation. We allow for the field to have a field strength of at most 9V/cm, which is at the limit of the order of magnitude at which electric fields are tolerated *in vitro*. For the optimal control problem in Equation (4.45), we require that the *physical* optimal solution, $s^\star_{\text{phys}}$, satisfies

$$s^\star_{\text{phys}} = \min\left(s^\star, s_{\max}\right), \tag{4.48}$$

where $s_{\max}$ is a parameter we can vary. By considering $s_{\max}$ across a range between 3 and 9 V/cm, we solve the optimal control problem and show the results in Figure 4.7. We set $v^\star$ as the mean velocity in the 1.5V/cm pulse experiment and now vary $s_{\max}$. While the solutions are qualitatively similar and the velocity is constant during most the experiment, the solutions are not identical. A closer inspection of the terminal velocity error shows that solutions with a weaker initial field have a larger error at the end, implying that the large initial pulse is what is necessary to create the terminal velocity, but it is the temporal evolution of the field which is necessary to create a plateau. While this error decreases rapidly for small values of the maximum field strength, $s_{\max}$, a close agreement with the desired terminal velocity is only achieved, once again, with unfeasibly high electric field strengths.

4.3.7 Constant velocity during a small time interval

When the goal was to optimise to keep the velocity constant, as in the previous section, we observed that the electric field profile had a non-physically high field density at the start of stimulation, making this optimal control strategy infeasible in practise. This comes from

requiring that the velocity hits the target velocity, $v^\star$, quasi-instantly upon the start of stimulation. Here, we ask the question of how solutions change when instead the velocity is only required to be constant on an interval $[t_1, t_2]$ with $0 < t_1 < t_2 < T$. Here, we will explore how the optimal electric field profile is affected by this requirement. We will also study how solutions are affected by altering the values for the start and end of stimulation, t_1 and t_2, respectively.

The functional in Equation (4.45) needs to be modified to address the optimality condition that the velocity is constant on the interval. This is because minimising the square of acceleration $\dot{v}^2$ as in Equation (4.45) leads to a trivial solution for the velocity: when the velocity is assumed to be zero by the initial conditions and the velocity is continuous, the optimal solution will be zero throughout the stimulation time frame. Therefore, one must consider a functional that directly relates the velocity during the interval $[t_1, t_2]$ to the target velocity. Such a functional can be defined as

$$I(v) = \int_0^T \mathbb{I}(t \in [t_1, t_2]) \cdot (v - v^\star)^2 \mathrm{d}t, \tag{4.49}$$

where $\mathbb{I}$ is an indicator function. This indeed restricts the constant velocity requirement to the desired interval, but the integrand is no longer a continuous function of time. To make the optimisation problem admissible, we choose to approximate the indicator function in the integrand by a smooth bump function, Φ, which we define as

$$\Phi(t) = \frac{1}{1 + \exp\left(-\frac{t - t_1}{\epsilon}\right)} \cdot \frac{1}{1 + \exp\left(-\frac{t_2 - t}{\epsilon}\right)}, \tag{4.50}$$

where ϵ determines the length scale of the region where the function Φ transitions from zero to one. We set $\epsilon = 50$ so that

$$\epsilon \ll t_1, \tag{4.51}$$

in all of our considered optimisation problems. Now we obtain an admissible optimal control formulation, given by

$$E(v) = \int_0^T \Phi(t) \cdot (v - v^\star)^2 \mathrm{d}t, \tag{4.52}$$

which yields a modified version of the adjoint equations for λ_v and $\lambda_{s_{\text{eff}}}$, which are

Equations (4.46) and (4.47), respectively. These become

$$\frac{\mathrm{d}\lambda_v}{\mathrm{d}t} = \Phi(t) \cdot 2(v - v^\star) + \gamma\lambda_v, \tag{4.53}$$

$$\frac{\mathrm{d}\lambda_{s_{\mathrm{eff}}}}{\mathrm{d}t} = -\alpha\lambda_v + \tau_e^{-1}\lambda_{s_{\mathrm{eff}}}, \tag{4.54}$$

subject to the terminal conditions $\lambda_\bullet(T) = 0$. Note that, in contrast to the optimal control problem for constant velocity in the previous subsection, there is no need to set up a terminal condition for the velocity. This means that the zero initial conditions of Equation (4.9) can be imposed. As before, we solve this boundary value problem using the bvp4c numerical scheme introduced in Section 4.3.6.

To determine the range of admissible values for the interval $[t_1, t_2]$, we make the assumption that $t_2 = T - t_1$, *i.e.* the stimulation interval is symmetric within the interval $[0, T]$. *A priori*, t_1 should be chosen sufficiently large that a physically realistic electric field is optimal, but not so large that the field is off at $t = 0$. To understand how to choose t_1, then, we note that there exist two characteristic times in this optimal control problem. First, since constant stimulation for times longer than the adaptation time scale, τ_a, results in a waning response from the field, choices for t_1 that result in the stimulation time being too long, *i.e.*

$$\frac{T - t_1}{2} \gg \tau_a, \tag{4.55}$$

will require a large value of the electric field initially to prevent cellular adaption to the field and will hence provide unphysical solutions that are not realistic stimulation protocols in practice. At the same time, one might expect that there is a time scale that dictates when the field is initially on at $t = 0$h and when it is off: if the waiting time from the beginning of stimulation, t_1, is shorter than this critical time scale, the electric field should be turned on immediately, since otherwise the monolayer will not have reached the target velocity during the optimality time window. Intuitively, this time scale should be of the same order as the time it takes for the velocity to reach its maximum under a constant electric field. We can estimate of this time scale, $\tau_{\max}$, by maximising the expression in Equation 4.3, which yields

$$\tau_{\max} = \frac{\log(\tau_a) - \log(\tau_e)}{\tau_e^{-1} - \tau_a^{-1}}. \tag{4.56}$$

For $t_1 > \tau_{\max}$ one would therefore expect that the optimal electric field density is zero at $t = 0$, while for $t_1 < \tau_{\max}$ one would expect that the density is nonzero. Therefore, one would expect that the choices of t_1 in the region

$$t_1 \in \left[\frac{T - \tau_a}{2}, \tau_{\max}\right] \tag{4.57}$$

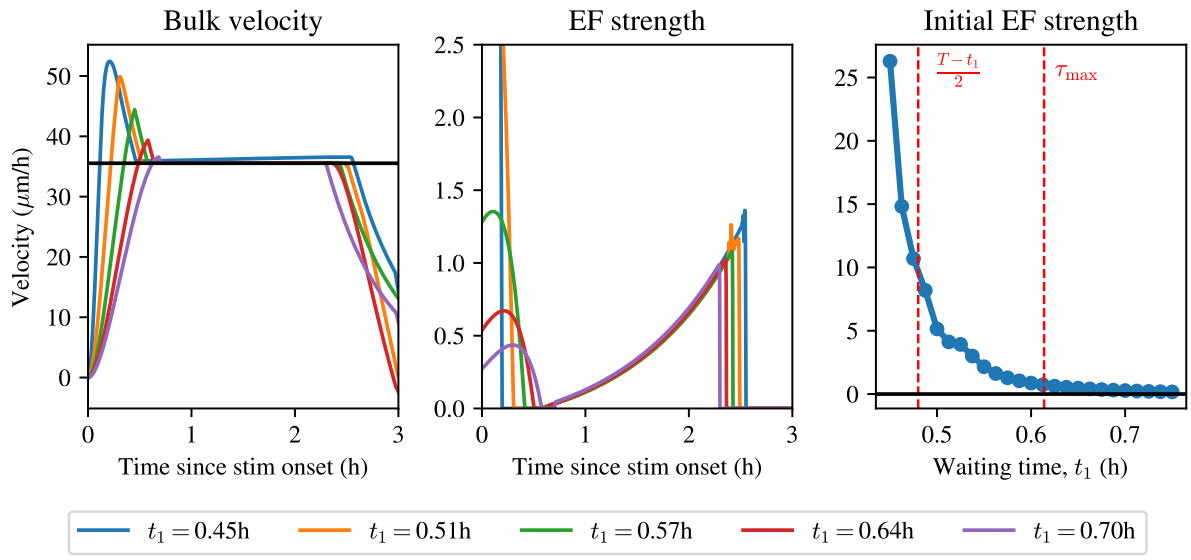


Figure 4.8: Optimal control to keep velocity constant in a small time frame. Left: bulk velocity traces using different waiting times for stimulation. Middle: electric field strengths with different stimulation waiting times. Right: initial electric field strength in the optimal control solution as a function of waiting time.

will yield optimal electric field protocols that are high at the onset of stimulation when t_1 is at the left boundary of the interval, or electric field protocols that slowly ramp after the onset of stimulation. We also expect values of t_1 smaller than the values in this range to yield physically unrealistic electric field protocols, and solutions for t_1 larger than the values in this range to have a near-zero initial electric field strength. In Figure (4.8), we show how the optimal electric field protocol changes when different values for $t_1, t_2, v^\star$ are used.

Figure (4.8) shows that the optimal electric field protocols change qualitatively as t_1 is varied. For small values of t_1, close to τ_a, the optimal electric field protocol exhibits a sharp overshoot of the velocity prior to reaching a plateau in the desired interval. This is achieved by the optimal electric field protocols exhibiting a sharp peak initially, bringing the velocity above the target, and then letting the velocity drop down so that it can later be kept at the target. This overshoot disappears whenever $t_1 > \tau_{\max}$. In this regime, the field allows the velocity to slowly increase to the desired target velocity. All optimal electric field protocols exhibit a ramp. This effect is strongly visible in the final plot of Figure (4.8), which shows that there is a sharp decrease in the initial field strength whenever $t_1 < \tau_{\max}$. We conclude that a constant velocity can be achieved without having an overshoot when the waiting time, t_1, is long enough. Finally, we note that by performing a grid search across possible values for the target velocity, $v^\star$, we found that these results do not change qualitatively with the value of $v^\star$, as the only apparent change is that the magnitude of the optimal electric field protocol scales linearly with $v^\star$.

4.3.8 Comparison to bang-bang control

The modest gain in distance travelled by the tissue when the optimal stimulation protocol is used suggests that naive stimulation by keeping the electric field at constant field density might be an effective policy. This is not surprising given that it is the policy converged upon after years of experimental iteration, and that the key variable here is when to switch off the field. This is reminiscent of bang-bang control in linear optimisation theory. Bang-bang control occurs when the Hamiltonian, H, of the optimal control problem is linear in the control variable, so that it disappears from the optimality condition altogether [12].

In our problem the Hamiltonian is nonlinear due to the $\mu_z S^2$ term included to enforce the condition that a fixed amount of charge is injected into the system. If one now assumes instead that the field is always positive, the condition could be relaxed so that instead S is bounded by some maximum strength, $0 \leq S \leq S_{\max}$. When the isoperimetric constraint is removed, the Hamiltonian becomes

$$H = v^2 + \lambda_v(t)\left[-\gamma v + \alpha s_{\text{eff}}\right] + \tau_e^{-1}\lambda_{s_{\text{eff}}}(t)\left[s - s_{\text{eff}} - I\right] + \tau_a^{-1}\lambda_I(t)\left[s - I\right], \tag{4.58}$$

with optimality condition

$$\frac{\partial H}{\partial S} = \tau_e^{-1}\lambda_{s_{\text{eff}}} + \tau_a^{-1}\lambda_I, \tag{4.59}$$

so that the optimal control is given by $s^\star = s_{\max}$, whenever the RHS of Equation (4.59) is positive by the Pontryagin optimality criterion. This, we have established in the previous numerical experiments, holds whenever $t < T$. In sum, we find that naive stimulation with a pulse is the optimal stimulation policy when the condition of total injected charge is relaxed and replaced instead with an upper bound on the electric field strength.

4.3.9 Discussion

So far, we have developed and calibrated a canonical model for excitation-adaptation to describe the cellular response to collective electrotaxis, whereby we have identified that the mechanism of excitation and adaption can faithfully describe the temporal dynamics of the bulk response to electric field stimulation when different stimulation protocols are applied. We have used this model to optimise for different experimental outcomes, including maximising the total distance travelled by the tissue bulk, maximising the terminal velocity, and keeping the migration velocity constant.

Our model shows that maximising the distance travelled by the tissue bulk is possible by changing the temporal distribution of electric field strength, but it only produces a marginal increase in performance relative to stimulation with an electric field with constant field strength. In fact, this increase is well within the standard deviation of the experimental data. We related the near-optimal performance of the constant stimulation

experiment to the fact that constant stimulation is the solution to a bang-bang control problem. A similar insight arises from controlling the maximum terminal velocity of the tissue, which recovers a ramp, but has an unphysically large field strength at the end of electric field stimulation. Together, this challenges the idea in the electrotaxis community that a rapidly increasing bulk velocity that is not maintained during stimulation is not optimal, as this velocity profile is in close agreement with the optimal solution for the maximum distance travelled for the tissue bulk.

Most importantly, our analysis shows that it is not possible to create constant velocity during electrotaxis when the experimental time exceeds the adaptation time scale, τ_a. This is reflected in the prediction that an unphysical field strength is necessary to create a constant velocity profile for the duration of the experiment. When we relaxed the constraint to allow the tissue to slowly increase to the desired velocity and then remain constant, such optimal electric field protocols only yielded velocity profiles without an overshoot whenever the stimulation time was within certain constraints. Together, these findings show that there exist physical limits to how we can stimulate, both in terms of duration of the experiment, and physiological constraints on electric field strength.

4.4 Incorporating cell metabolism in electrotaxis

In the previous section, we established that a mechanism by which cells passively adapt to a constant signal explains the temporal dynamics of the bulk velocity of MDCK monolayers under stimulation with an electric field. Crucially, these experiments consisted of one short window of stimulation, in which the field strength was positive for the entire duration. The validity and applicability of the model derived in the previous section can be tested for more complicated temporal electric field stimulation patterns by considering an experiment in which the tissue is exposed to two rounds of stimulation by the electric field, separated by a period of rest. This experiment can test the model, since it allows to explore possible non-linearities in the model which would arise, from example, from cellular memory of previous activity. The motivation to investigate the presence of such an effect is that experimental results in Chapter 4 established that MDCK cells actively modulate their force production as a function of expended mechanical work: the cells that have expended most mechanical power during the course of the experiment are the least responsive to the field, and the cells that have expended the least amount of mechanical energy are the ones producing the largest amount of active force in the direction of the field. Although this is a cell-level property that determines heterogeneity at the cell – and not the tissue – level, we wish to include this active regulation mechanism in our model to understand whether it can equally be used to explain the slowdown that occurs during the course of the experiment.

4.4.1 Adaptation alone cannot explain temporal dynamics in a double-stimulation experiment

To identify the validity and applicability of the model presented in Section 4.3, data are needed that run over a longer time frame than the data for a single pulse of constant stimulation. As such, we performed a new experiment in which two rounds of electric field stimulation, with a rest interval in the middle, are administered to the tissue (see also Table 4.1). In these experiments, the tissue was allowed to settle for one hour after seeding and then exposed to two hours of electric field stimulation at 3V/cm, followed by two hours of no electric field, followed by another two hours of electric field stimulation at 3V/cm. Data traces from this experiment are shown as the scatter in Figure 4.9. The data plots in Figure 4.9 show that in this two-pulse experiment the bulk did not only exhibit a slowdown in each of the two pulses – which the considerations of the previous section would suggest are due to passive adaptation to the signal – but also that the peak velocity during the second window of stimulation is much lower than during the first. We wish to understand if this slowdown can be captured by the model of the previous section. Since this model included passive adaptation to the field only, we will refer to this model as the *passive adaptation* model for electrotaxis.

We perform Bayesian inference of the model parameters using the two-pulse experimental data. As before, we use the sum of squared residuals between model and data in the inference and use the same MCMC settings as in Section 4.3.3 – including the same convergence criteria. The results of the inference procedure are shown in Figure 4.9. We find that the model parameters can be confidently identified given the experimental data, as seen by the well-defined posterior distributions in Figure 4.9. The posterior means for α, τ_e, τ_a are given by 212.77μm/h, $9.16{\cdot}10^{-2}\text{h}^{-1}$, and 2.965 h^{-1}, respectively. While these parameters can be identified confidently from the data, we find that the posterior distribution over model parameters yields model predictions that do not accurately describe the data from this new experiment. Figure 4.9 shows that the model cannot accurately capture the large difference between the two velocity peaks: the best fits of the model underestimate the velocity in the first window but overestimate the velocity in the second window. This is evident from the fact that the confidence intervals for the posterior predictions and the experimental data are disjoint for all time during stimulation. This discrepancy between model and data suggests that other sources of activity history might be at play, and their role needs to be investigated carefully.

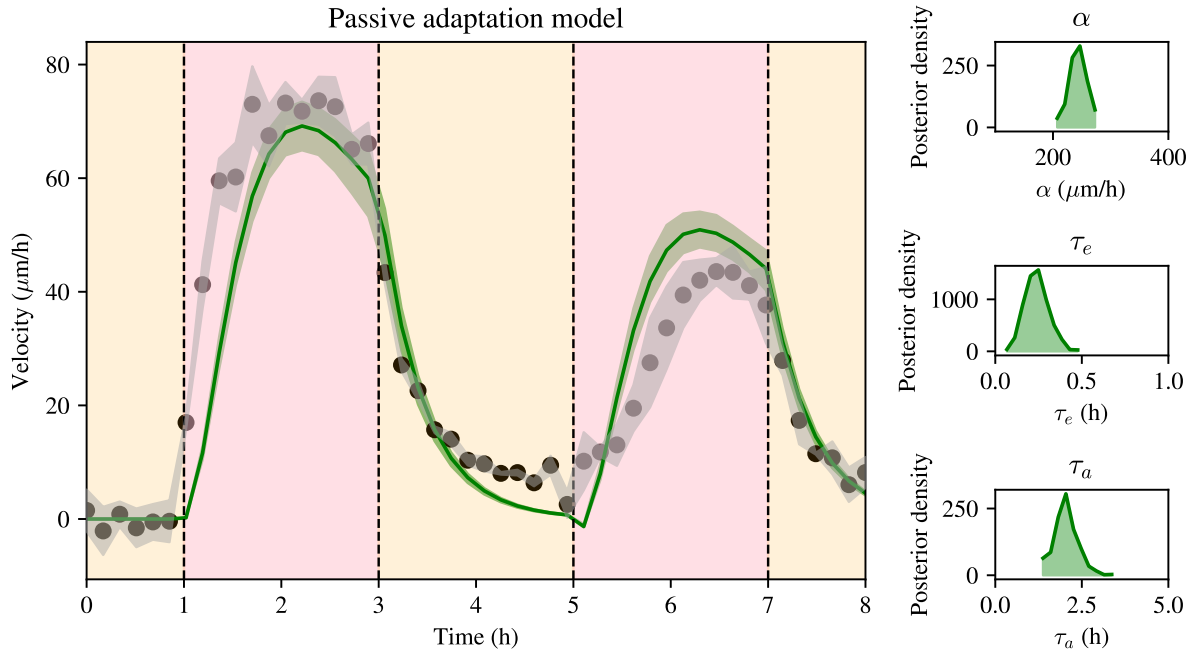


Figure 4.9: Bayesian inference for the model using using only passive adaptation to the field. Left: posterior distribution of model predictions with posterior 95% confidence interval (green). Experimental data and its corresponding confidence interval in grey. Right panels: posterior distributions over model parameters.

4.4.2 A model for active regulation mediated by mechanical work

Recall that at the beginning of this section we hypothesised that the amount of mechanical work expenditure might act as a source of nonlinearity that contributes to a lower second peak during the double pulse experiment. As a first step to capture the effect of mechanical work, we formalise the concept of energy per unit mass spent during migration by introducing a new state variable for the average mechanical work done, p. We introduce this quantity to capture the energetic history of the tissue, by using a weighted time integral of the energy dissipated as a result of friction against the substrate. Here, we use the mechanical work per unit mass performed by the cells to overcome friction with their substrate as the source of mechanical dissipation. A consequence of this is that a sustained high average velocity during electrotaxis will lead to large friction forces, which lead to a high amount of mechanical work performed during migration. Interpreted biologically, one might think that this represents the amount of energy from their energy stores cells have expended during migration. This idea can be captured in a phenomenological ODE model for the mechanical work,

$$\dot{p} = \gamma v^2 - \beta p. \tag{4.60}$$

The first term in Equation (4.60) models the rate at which mechanical energy per unit mass is expended, which is the product of the friction per unit mass, γv, and the velocity, v. The second term models that cells are continuously replenishing their energy stores at a rate that is proportional to their net work performed. This rate is given by β, which determines the time scale on which cells can recover from energy expenditure, and is given in units of h^{-1}. Solving Equation (4.60) gives

$$p(t) = e^{-\beta t} \int_0^t e^{\beta \tau} \gamma v(\tau)^2 \mathrm{d}\tau. \tag{4.61}$$

This expression for the work variable, p, justifies the interpretation that cells have a temporal memory that allows them to remember their power expenditure at short time scales but forget their activity at longer time scales, which corresponds to cells being able to recover and replenish their energy supplies from the medium during a period of lower activity. We capture this mechanism through the choice of exponential decay for the power state variable, which is consistent with the kinds of activity history that abound in viscous mechanics. Putting these formulations together creates a hereditary integral that keeps track of the quantity v^2, *i.e.* the kinetic energy, and renews at an exponential rate.

The state variable p was introduced with the aim of modelling the phenomenon that when cells have dissipated more energy, they are less responsive to the field. Therefore, we now introduce a phenomenological mechanism that modulates the influence of electric field on the monolayer velocity. Whereas in Equation (4.7), the electric field influences the velocity through the dynamical system with a constant coefficient, α, we will now allow for this term to be a monotonically decreasing function of the active power, p, so that

$$\dot{v} = -\gamma v + \alpha(1 + \tanh(-\xi \cdot p))s, \tag{4.62}$$

where ξ scales the magnitude of regulation as a function of the work. Note that ξ is given in units of $(\mu\mathrm{m/h})^{-2}$. In this model, Equations (4.60) and (4.62) are subject to the initial conditions

$$v(0) = p(0) = 0, \tag{4.63}$$

respectively. Since this model is regulated by mechanics alone, we refer to it as the *mechanically regulated model*.

4.4.3 Linear stability analysis

We seek to understand whether active regulation, in the form of Equations (4.60) and (4.62), alone is sufficient to explain the drop in velocity even as the field is held constant. For

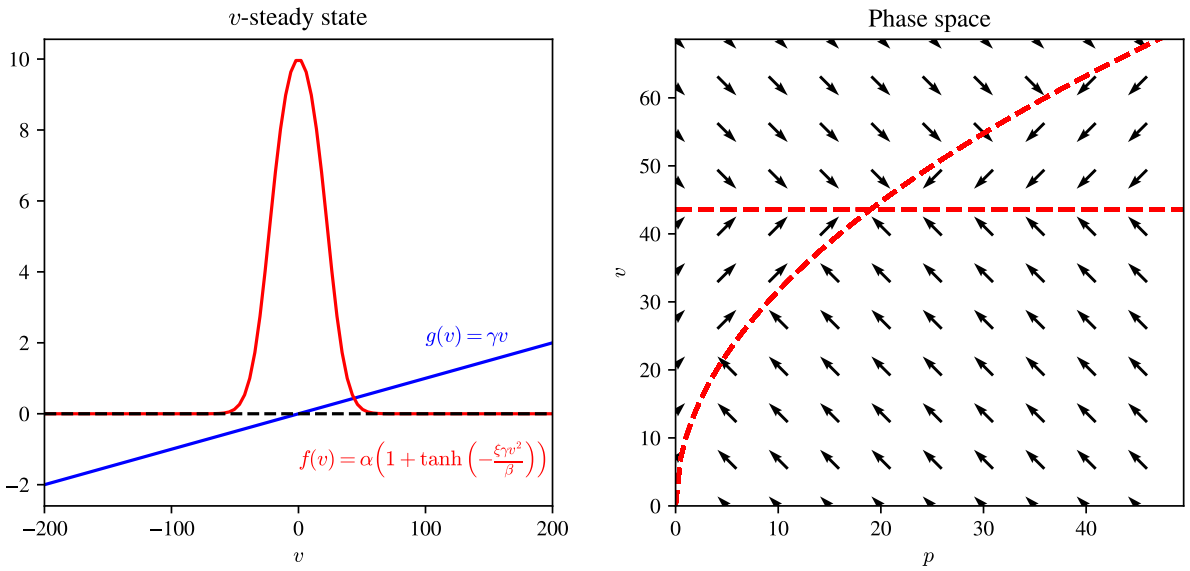


Figure 4.10: Graphical summary and phase plane of the energetics model consisting of Equations (4.60) and (4.62). Left: graphical summary of the two terms in Equation (4.62), showing there is one unique intersection which corresponds to the v steady state. Right: phase plane of the model denoting the p-nullcline given by $p = \gamma v^2/\beta$, and the constant v-nullcline. Parameters used for this phase space are $\gamma = 0.01\text{h}^{-1}$, $\beta = 1\text{h}^{-1}$, $\xi = 0.1(\mu\text{m/h})^{-2}$, and $\alpha = 10\ \mu\text{mh}^{-1}$.

this, we turn to a linear stability analysis to find that the steady state of the system when $s = 0$ is given by $v = 0$ and $p = 0$, and when $s > 0$ we obtain that $p = \gamma v^2/\beta$, and

$$v = \frac{\alpha}{\gamma}\left(1 + \tanh\left(-\frac{\xi\gamma v^2}{\beta}\right)\right)s. \tag{4.64}$$

A graphical inspection of Equation (4.64) to find this steady state shows that the LHS, as a monotonically increasing function of v, has exactly one intersection at $v = v^\star$ with the monotonically decreasing right-hand side RHS (since it equals 1 at $v = 0$). Moreover, since the LHS is strictly smaller than the RHS for $v < v^\star$, we deduce that this steady state is linearly stable. We formalise this analysis by drawing the phase plane in Figure 4.10. From the phase plane, we therefore expect that, given a positive value of the signal, the system is driven towards the stable equilibrium. The phase plane also demonstrates that the system cannot exhibit a slowdown when the electric field is turned on, as the velocity will increase monotonically until it reaches the equilibrium, which excludes the possibility of the bulk velocity rising and falling during the course of stimulation with the electric field. As a consequence, a system with energetic regulation alone is not sufficient to explain the experimental finding that the response to constant electric field wanes over time.

4.4.4 Bayesian inference of the mechanically regulated model

Having understood that the model consisting of Equations (4.60) and (4.62) cannot predict a slowdown, we wish to understand which characteristics of the data can be captured by the model, and which elements of the model might need adjusting to account for this phenomenon. To do this, first, we revisit the data from the stimulation experiments using pulse and ramp stimulation from Section 4.3. The model consists of Equations (4.60) and (4.62), subject to the initial conditions in Equation (4.63). Like in Section 4.3 we use the sum of squared residuals between the data and the model predictions as the distance function. As it concerns the same data, we use the same estimate for the velocity decay rate, γ, which was found in Section 4.3 by fitting a decaying exponential. This allows us to perform Bayesian inference on a subset of the unknown parameters α, β, and ξ. We use the same MCMC implementation as in Section 4.3.3, again using the $\hat{R}$ statistic as a diagnostic for convergence.

We find that the model parameters α and ξ can be confidently identified using the data, with posterior means 192.33 μm/h^2 and 0.0731 h^{-1}, respectively. The posterior distribution for β is not identifiable, as shown by the broad and uninformative marginal posterior distribution in Figure 4.11. The non-identifiability of the parameter β is not due to a restrictive prior range, since extending the upper limit of the prior range an order of magnitude to $5 \cdot 10^3(\mu\text{m/h})^{-2}$ does not change the fact that the posterior is broad and uninformative. The non-identification of β should not be surprising, given the posterior fits of the model to the experimental data in Figure 4.11. In the pulse stimulation case, the posterior fits show that pulse stimulation results in the bulk velocity quickly reaching the steady state velocity and then staying constant throughout the stimulation, regardless of the value of the parameter β.

Put together, these findings suggest that a model that modulates the tissue response to the electric field as a function of its energy dissipation history is on its own not sufficient to explain the drop in velocity when the electric field is held constant. Instead, the passive adaptation mechanism detailed in Section 4.3 might be better suited to model this problem. The biological implications of this are that, while single cells moderate their energetic expenditure as a function of past activity history, this trend cannot be observed at a collective level, at least on short time scales like those in the experiments discussed in this section.

4.4.5 Active regulation working in tandem with passive adaption

In the last section, we found that a model in which cells actively modulate their force expenditure as a function of their expended energy does not account for the observed temporal dynamics of the slowing bulk when the electric field is held constant. Meanwhile,

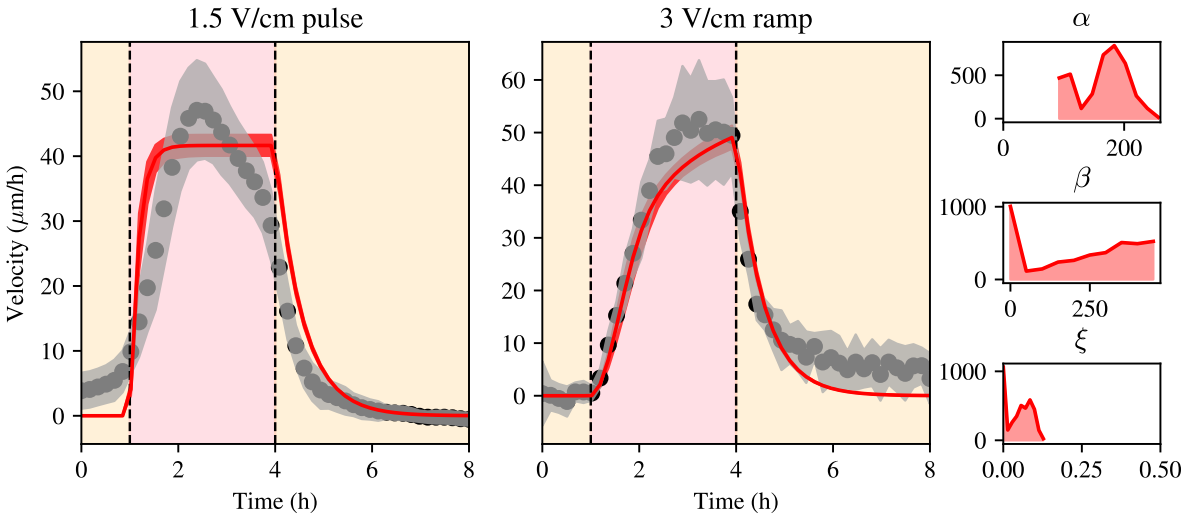


Figure 4.11: Bayesian inference on a model of active regulation using experimental data from an electrotaxis experiment. Left : bulk velocity using a constant field of 1.5 V/cm. Posterior mean of the model and 95% confidence interval in red, experimental data and confidence interval in grey. Middle: bulk velocity using a linear ramp from 0 to 3 V/cm. Right panels: posterior distributions for model parameters, α, β, ξ, respectively. Note the broad posterior for β and the multimodal posterior for α.

at the cellular level, there is strong experimental evidence to support the hypothesis that active regulation of force expenditure as a function of past energy dissipation indeed plays a role in the electric field response. To reconcile this finding with the fact that regulation of active forces through energy dissipation is present at the single-cell level, we wish to explore whether energy expenditure and passive adaptation to the electric field might both be at play, on different time scales that cannot be observed in the single pulse experiment of the preceding section. We combine the two different mechanisms by letting the work state variable directly influence the strength of the effective signal. That is, we let the equations for the work, p, Equation (4.60), the effective signal, s_{eff}, Equation (4.1), and the inhibitor, I, Equation (4.2) be as before, and consider

$$\dot{v} = -\gamma v + f_i(p) s_{\text{eff}}^+, \tag{4.65}$$

$$\dot{p} = \gamma v^2 - \beta p, \tag{4.66}$$

$$\tau_e \dot{s}_{\text{eff}} = s - (s_{\text{eff}} + I), \tag{4.67}$$

$$\tau_a \dot{I} = s - I, \tag{4.68}$$

subject to the initial conditions

$$v(0) = p(0) = s_{\text{eff}}(0) = I(0) = 0. \tag{4.69}$$

Here, $f_i(p)$ represents the different functional forms that can be given to the magnitude of the active force per unit mass. For example, in Section 4.4.1, we considered

$$f_1 \equiv \alpha \tag{4.70}$$

to find that this model did not yield an accurate fit to the data. For the new model, which we denote the *adaption-regulation* model since it combines both phenomena, we consider the functional form chosen in the start of this section,

$$f_2(p) = \alpha(1 + \tanh(-\xi \cdot p)). \tag{4.71}$$

Hence, the adaption-regulation model assumes that active regulation and passive adaptation work in tandem by the active regulation mechanism modulating the effective signal in the presence of high previous energy expenditure. In other words, one could *a priori* expect that the power will reduce future sensitivity to the signal, even when the effective signal transduced within the cell is the same. We seek to understand whether this mechanism is sufficient to explain the existence of a lower second peak when two consecutive pulses of stimulation are given to the bulk.

4.4.6 Bayesian inference for the adaption-regulation model: regulation is needed to explain long time scale behaviour

We perform Bayesian inference on the adaptaptive response model, *i.e.*, using f_2. Using the same MCMC implementation as in Section 4.4.1, we find that the posterior means for α, β, ξ, τ_e, and τ_a, are given by 235.48μm/h, 3.78·10^{-2}h^{-1}, 2.68·10^{-5} (μm/h)$^{-2}$, 3.33·10^{-2}h, and 7.97h. Inspection of the posterior predictive distributions of these model parameters in Figure 4.12 shows that all model parameters can be confidently identified from the data. Moreover, the adaptive model can faithfully recapitulate the large difference in response between the two electric field pulses, as can be seen by the fact that the posterior distributions of the model are in good qualitative agreement with the uncertainty bounds of the experimental velocity data. We assess the quality of the model fits of the new, adaptive model, and the simple adaptation-only model using the Bayesian information criterion (BIC), which is a formal comparison of how well the model fits the data. The BIC penalises models that have more parameters and tries to find a parsimonious model that can fit the data well. The BIC is given by

$$\mathrm{BIC} = k \ln(n) - 2 \ln(\hat{L}), \tag{4.72}$$

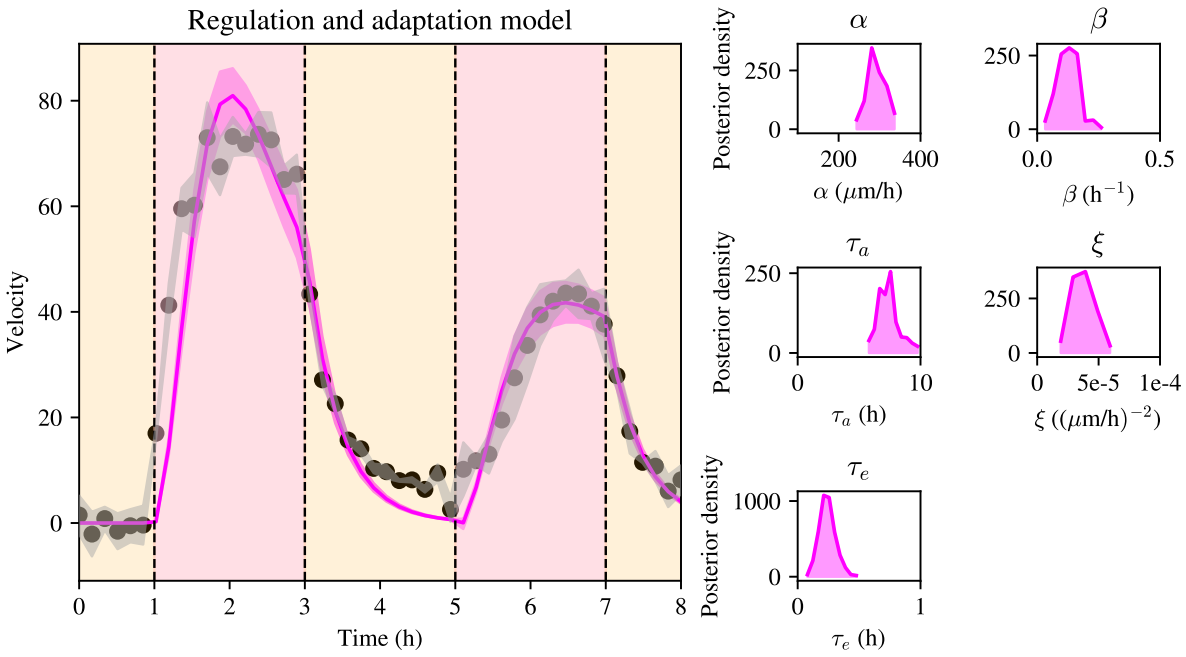


Figure 4.12: Bayesian inference for the model using including active regulation and passive adaptation. Left: posterior distribution of model predictions with posterior 95% confidence interval (magenta). Experimental data and its corresponding confidence interval in grey. Right panels: posterior distributions over model parameters.

where $\hat{L}$ is the maximum likelihood of the data under the model, k is the number of estimated parameters estimated in the model, and n is the number of datapoints. In general, a lower value of the BIC is preferred. Here, we use a Gaussian likelihood function to estimate the likelihood, $\hat{L}$, with the posterior mean for the variance found in the MCMC inference. This yields a value of 358.85 for the BIC of the simple model and a value of 347.67 for the BIC of the full model. Therefore, despite the full model being more complex with two additional parameters, we find that it is better suited to describe the data, which is in agreement with our qualitative finding that the simple model cannot explain the large difference between the two peaks in the double stimulation experiment.

Together, these findings suggest that, while at short stimulation times when the electric field is on for an uninterrupted period of time, the mechanism of passive adaptation is an appropriate description of the underlying dynamics, both active regulation and passive adaptation are necessary to explain the dynamics in the two-pulse experiment.

4.4.7 Separating the contributions of regulation and adaptation

The posterior distributions found using Bayesian inference in the previous subsection reveal the roles played by passive adaptation and active regulation in creating the slowdown in the second peak observed in the model simulations. Table 4.2 displays the posterior means for each of the model parameters.

Model	α (μm/h)	τ_a (h)	τ_e (h)	β [h^{-1}]	ξ [(μm/h)$^{-2}$]	σ^2
Adaptation and regulation	302.5	5.49	0.253	0.129	$3.89 \cdot 10^{-5}$	7.54
Adaptation only	251.7	2.142	0.262	NA	NA	9.77

Table 4.2: Posterior means of key inferred model parameters for the two models considered in this section.

In both models, the parameter α, which corresponds to the maximum response to the electric field, is of a magnitude similar to that of the short stimulation experiments in Section 4.3. Likewise, the excitation timescale, τ_e, is similar across the two posterior distributions. The marginal posterior mean for the adaptation timescale, τ_a, however, is different in the two models. The marginal posterior mean for the adaptation time scale, τ_a, is greater in the model involving active regulation than in the model containing only passive adaptation. In addition, the adaptation time scale in the model that includes energetic regulation is also longer than the time scale for the decay of the energy dissipation history, β, which is of a smaller magnitude than the excitation and adaptation time scales, τ_e and τ_a, respectively. Mechanistically, this suggests that in the active regulation model, the cellular adaptation to energy expenditure and cellular adaptation to the field indeed occur on different time scales. At long time scales, the power expended leads to a lower response in the second pulse of stimulation since the tissue has not 'forgotten' its energy expenditure in the first window of stimulation, *i.e.* there is accumulated fatigue preventing a strong tissue response when the second window of electric field is administered. Through this mechanism, the term in Equation (4.65) coupling the effective signal to the field is smaller than its equivalent in the passive adaptation only model. In contrast, in the model containing only passive adaptation, the second peak is lower because the tissue is still 'adapted' to the electric field.

We further explore the idea that passive adaptation and active regulation work in tandem to influence the instantaneous response to the field, as well as a slowly adapting propensity to respond to the field at a longer time scale. We do this by analysing the different terms in the active regulation model during the course of one forward model simulation. We parameterise the active regulation model with the marginal posterior means in Table 4.2 and simulate the double stimulation experiment, while keeping track of all state variables. In Figure 4.13, we plot the simulation results with the values computed for the activity history, p, the response function, f_1, the effective signal, s_{eff}, and the effective response, which is given by $\alpha(p) \cdot f_1$.

In the double-pulse experiment, the most significant difference in the state variables between the first pulse of electric field and the second is that the responsiveness to the field, $\alpha(p) \cdot f_1$, has been greatly reduced. Note that this is a direct consequence of the increased activity history from the first electric field pulse to the second, as shown in Figure 4.13. Put together, the tissue is much less well placed to respond to the second

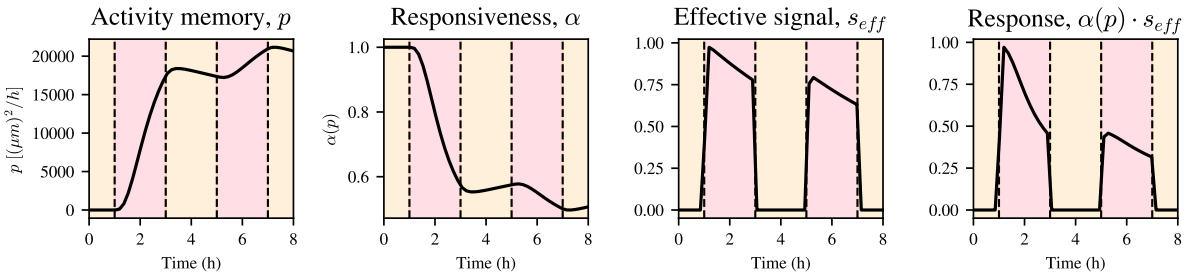


Figure 4.13: Plots of the activity history, effective signal, and responsiveness function to the electric field, $\alpha(p)$, showing that the responsiveness wanes and as a result the tissue is pre-conditioned to be less responsive in the second electric field pulse.

pulse and, as a consequence, the response during the second pulse of stimulation is much lower than it is during the first. This finally also means that we can conclude that the two effects of adaptation and regulation are both identified in this experiment, and that they act on different time scales, which can be seen moreover by comparing the time scales of β, τ_e, and τ_a.

4.4.8 Discussion

In this section, we used a model derived using recently generated experimental data to capture the phenomenon that cells deplete their energy stores during collective electrotaxis and that this leads to a diminishing response to the electric field. We showed with Bayesian inference that this phenomenon can explain the temporal dynamics of the evolution of the bulk velocity when the tissue is exposed to two pulses of electric field stimulation, while a simple model that does not take cellular energetics into account fails to do so.

We note that the choice for the tanh function in the active regulation term in the model of Equation (4.62) was phenomenological, so a relevant question that arises might be to what extent these results are sensitive to a different choice of response function. First, we note that one might seek to restrict the choice of the response function in Equation (4.65) to any monotonically decreasing function of p, which can be normalised to be 1 at $p = 0$. Given these assumptions, the system would still exhibit a single stable steady state, which means that the system would exhibit the same qualitative feature that there is no adaptation to the field, and our conclusion that active regulation works in tandem with passive adaptation would be equally valid.

4.5 Optimal control of a double stimulation experiment

The central conclusion of the previous section was that active regulation as a function of the activity history and passive adaptation work in tandem to create complicated temporal patterns in the bulk response of collective electrotaxis. Notably, it was found that past mechanical activity effectively pre-conditions the tissue to become less responsive to the electric field. While this effect was negligible for short periods of electric field stimulation and therefore did not need to be taken into account in the optimal control problem considered in Section 4.3, it becomes relevant when considering stimulation on the time scales at which active regulation is observed at a collective level. In this section we therefore seek to develop an optimal control framework for this problem and derive an optimal policy that addresses these issues. In all analyses, we will again consider the isoperimetric problem controlling the total injected charge. The main problem of this section is how to identify a stimulation pattern that ensures that the velocity in both the peaks is constant at a certain velocity.

4.5.1 Optimal control problem

To enforce that the velocity between the two different peaks in the double simulation experiment (see also Table 4.1) is the same, and constant during stimulation, we consider the same functional as for constant stimulation during an interval as in Equation (4.52), *i.e.*, we seek to minimise the total integrated value of the function

$$f(v) = (v - v^\star)^2, \tag{4.73}$$

during the stimulation windows, given a target velocity, $v^\star$. Similarly, we require that the field is constant on the domain $[t_1, t_2] \cup [t_3, t_4]$, where

$$1\text{h} < t_1 < t_2 < 3\text{h}, \quad 5\text{h} < t_3 < t_4 < 7\text{h}, \tag{4.74}$$

and, for simplicity, we choose that

$$t_1 = 1 + t^\star, \quad t_2 = 3 - t^\star, \quad t_3 = 5 + t^\star, \quad t_4 = 7 - t^\star, \tag{4.75}$$

i.e., all the intervals are symmetric. By making a similar approximation as in Equation (4.52), we approximate the indicator function for the interval in question by a smooth

bump function, Φ, which we define as

$$\Phi(t) = \frac{1}{1+\exp\left(-\frac{t-t_1}{\epsilon}\right)} \cdot \frac{1}{1+\exp\left(-\frac{t_2-t}{\epsilon}\right)} + \frac{1}{1+\exp\left(-\frac{t-t_3}{\epsilon}\right)} \cdot \frac{1}{1+\exp\left(-\frac{t_4-t}{\epsilon}\right)}, \quad (4.76)$$

where ϵ determines the length scale of the transition of the bump from zero to one, and as in Equation (4.8), we set $\epsilon = 50$. Now we obtain an admissible optimal control problem given by

$$I(v) = \int_0^T \Phi(t) \cdot (v - v^\star)^2 \mathrm{d}t. \quad (4.77)$$

Since the regulation and adaptation model consists of a larger system of ODEs than previously considered, we obtain a larger system of adjoint equations than in Section 4.3, which now include the constraints imposed by the differential equations for v, p, s_{eff}, and I. This problem includes a Hamiltonian with four different multipliers, λ_v, λ_p, $\lambda_{s_{\text{eff}}}$, and λ_I. Using the same optimal control theory method as that used to derive the Hamiltonians in the optimal control problems in Section 4.3.6, one obtains the Hamiltonian for this problem:

$$\begin{aligned} H = \Phi(t) \cdot (v - v^\star)^2 + \lambda_v \left[-\gamma v + \alpha(1 + \tanh(-\xi p)) s_{\text{eff}}\right] + \lambda_p \left[\gamma v^2 - \beta p\right] \\ + \tau_e^{-1} \lambda_{s_{\text{eff}}}(t) \left[s - s_{\text{eff}} - I\right] + \tau_a^{-1} \lambda_I(t) \left[s - I\right]. \end{aligned} \quad (4.78)$$

Then, as usual, we find the differential equations for the multipliers by taking derivatives of the Hamiltonian with respect to the state variables,

$$\frac{\mathrm{d}\lambda_v}{\mathrm{d}t} = -\frac{\partial H}{\partial v} = -2\Phi(t) \cdot (v - v^\star) + \gamma \lambda_v - 2\gamma \lambda_p v, \quad (4.79)$$

$$\frac{\mathrm{d}\lambda_p}{\mathrm{d}t} = -\frac{\partial H}{\partial p} = \xi \alpha \lambda_v \cdot (1 - \tanh^2(-\xi p)) + \beta \lambda_p, \quad (4.80)$$

$$\frac{\mathrm{d}\lambda_{s_{\text{eff}}}}{\mathrm{d}t} = -\frac{\partial H}{\partial s_{\text{eff}}} = \alpha \lambda_v \cdot (1 + \tanh(-\xi p)) + \tau_e^{-1} \lambda_{s_{\text{eff}}}, \quad (4.81)$$

$$\frac{\mathrm{d}\lambda_I}{\mathrm{d}t} = -\frac{\partial H}{\partial I} = \tau_e^{-1} \lambda_{s_{\text{eff}}} + \tau_a^{-1} \lambda_I, \quad (4.82)$$

subject to the terminal conditions that $\lambda_\bullet(T) = 0$ and zero initial conditions for all state variables. As in the simple adaptation-excitation optimal control problem in Section 4.3, we obtain that the optimal control protocol for the electric field, $S^\star$, is given by

$$S^\star \propto \frac{\partial H}{\partial S} = \tau_e^{-1} \lambda_{s_{\text{eff}}} + \tau_a^{-1} \lambda_I, \quad (4.83)$$

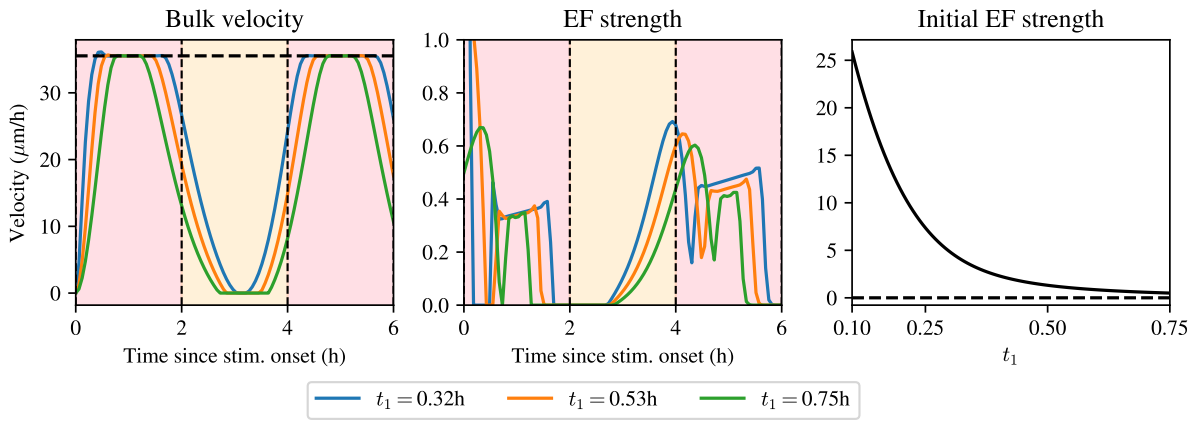


Figure 4.14: Optimal control protocol for a double-pulse electric field experiment. Left: optimal solutions for the velocity, v, under different choices for t_1. Middle: corresponding optimal electric field profiles. Right: initial field strength.

subject to $0 \leq S^\star \leq S_{\max}$. We solve the system of Equations (4.79)-(4.82) with the same bvp4c numerical scheme [130] which we introduced in Section 4.3.6.

4.5.2 Role of different stimulation interval lengths

Like in Section 4.3.6, we will focus on the problem of controling the velocity during the stimulation windows only in a sub-window of the stimulation window. That is, we will again allow for the tissue velocity to increase during an interval of some length, $t^\star$, and only require the velocity to be constant after $t = t^\star$. As a result, we wish to understand how optimal control solutions change when the parameter $t^\star$ changes. As before, there are some characteristic time scales that determine the admissible range of the parameter $t^\star$. First, note that the waiting time, $t^\star$, should reasonably be larger than τ_e so that the tissue can respond to the electric field. Likewise, we should allow for the stimulation time to be shorter than the time scale on which active regulation through energy expenditure slows the field down, which is expressed mathematically as requiring that $2 - 2t^\star < \beta$. Put together,

$$t^\star \in \left[\tau_e, 1 - \frac{\beta}{2}\right]. \tag{4.84}$$

With this in mind, we consider values of $t^\star$ in this range and observe how the different optimal electric field protocols resulting from Equation (4.77) vary with different values of $t^\star$. Figure 4.14 shows how the optimal electric field protocols change when $t^\star$ is varied. Changing $t^\star$, yields a family of optimal control electric field protocols whose shapes are qualitatively similar. A common feature of these electric field profiles is that the electric field is stronger in the second window of stimulation than in the first (which is in agreement with the intuition that the tissue is more 'tired' and adjusted at the second peak so it requires a stronger electric field to migrate at the same velocity) and that the electric

field strength is increasing for most of the stimulation window, which counteracts the adaptation and cellular exhaustion.

Moreover, it can be seen that to achieve the target velocity within the sub-intervals, it is necessary to start stimulating outside of the stimulation window, as shown by the fact that the optimal electric field protocols in Figure 4.14 are nonzero during the time between the two stimulation windows. For small values of $t^\star$ it is even the case that the maximum electric field strength is attained before the start of the stimulation window. There is a feature reminiscent of the optimal electric field solutions of Section 4.3.6 where there is an unphysically high electric field at the beginning of stimulation. Together, we find that it is possible to find an optimal stimulation pattern to achieve constant velocity during a sub-interval, and that there is a typical, characteristic optimal electric field protocol that can be used to achieve this.

4.6 Discussion and outlook

In this work, we developed a detailed continuum model for the evolution of the bulk velocity of an epithelial monolayer during electrotaxis. By using Bayesian methods to fit the model to experimental data obtained in different canonical electrotaxis experiments, we showed that the temporal dynamics of the bulk velocity can be explained by passive adaptation to the field at short time scales, and that active regulation of force production as a function of expended energy has a dramatic effect on collective electrotaxis on long time scales. This modelling framework enabled us to formulate an optimal control problem for several different desired movement patterns, including maximising the total distrance travelled by the monolayer, maximising the velocity of the monolayer, and enforcing constant velocity of the monolayer.

Our framework establishes a reproducible and predictive model for collective cell migration that can be used to develop new experimental approaches. Conversely, experimental data can help validate and develop the optimal control framework proposed in this work. More generally, our modelling approach can be generalised and extended to be used in other contexts where there exists a desired velocity trace and one must find an optimal electric field protocol to achieve this pattern of migration. This can be achieved directly in our framework by expressing the desired movement pattern as the optimal solution to a linear control problem and the optimal control framework put forth in this chapter.

In this work we have only considered uniaxial electric fields, while experimental set-ups exist to design spatially dependent fields [39, 125, 126]. Our framework can be extended to optimise for desired migration outcomes that vary in both time and space. For the optimisation and control of spatially varying fields, one must first develop a spatial model of collective electrotaxis that takes into account the different migratory cues that are present

in the different parts of the monolayer. Using such a spatially dependent model can enable the design and implementation of physically relevant applications of electrotaxis, such as those used in wound healing or steering collective migration. We turn to considering this problem in the next chapter.

Chapter 5

Optimal spatial control of collective electrotaxis in epithelial monolayers

The work in this chapter uses experimental data originally published by Wolf *et al.* [1]. My contributions to the work were the development of the continuum mechanics modelling, all computational implementations, and the writing of the results in this chapter.

5.1 Introduction

In Chapter 3 and Chapter 4, we have seen that an electric field can act as an external driver of collective cell migration and have a dramatic impact on collective behaviour by providing the constituent cells in a tissue with a persistent cue to migrate in a given direction. For this reason, electrotaxis has become known as a method to reliably 'steer' a cellular collective [40]. We explored the possibilities for optimal steering of an epithelial monolayer in Chapter 4 by considering the velocity of the centre of mass of the monolayer, and deriving an optimal stimulation protocol such that the velocity of the centre of mass satisfies certain optimality conditions. In this chapter, we seek to refine this method by considering the question of how to use electric fields to steer entire tissues. That is, rather than controlling just the movement of the centre of mass, how can mathematical modelling be used to design electric fields to move entire tissues, and making sure that the entire tissue moves according to some desired velocity?

Controlling the movement of an entire population of cells is significantly more challenging than controlling the movement of the centre of mass of a collective, for the following reason. In a large epithelial monolayer, there exist several endogenous cues that lead cells in different positions to have a preferential polarity, and exert active forces in different directions. For example, in MDCK monolayers, as we reviewed in Chapter 3, there exists

a clear polarity at the edges that acts as a cue for cells to migrate outward. Therefore, when studying how to use an electric field to steer an entire epithelial monolayer, one must have a good understanding of how the endogenous migratory cues inside the monolayer compete with the external migratory cue of the electric field to determine the cell polarity and active forces that drive collective migration.

The question of how the electric field interferes with local endogenous cues has been documented experimentally but is still not well understood. For example, Wolf *et al.* find that the electric field reprograms long-term supracellular dynamics, and that the edge regions of the tissue have significantly different velocity dynamics than the bulk of the tissue [1]. Modest theoretical advances exist to understanding mechanistically how the electric field interferes with endogenous migratory cues to create collective migration. For example, Zhang *et al.* [131] develop an on-lattice model that phenomenologically models the collective response to the electric field as a linear superposition of different forces, including cell-cell repulsion, and a force in the direction of the electric field. While these results can capture some of the features of collective electrotaxis, they do not take into account the temporal dynamics of adaptation to the field, which we demonstrated in Chapter 4 have a dramatic effect on the controllability of the tissue. Likewise, the on-lattice model cannot be used directly to design optimal electric fields for any given velocity pattern, which makes this framework difficult for practical applications in tissue engineering. To overcome both these difficulties, this work will be concerned with developing robust continuum models that can capture the spatial heterogeneity of velocities in MDCK monolayers undergoing electrotaxis, and using these models to propose the design of electric fields that induce collective migration.

As a first contribution, this chapter will propose a model extension to the ODE model for the tissue bulk velocity in Chapter 4. This model extension will introduce terms that that relate the position of cells in the monolayer to endogenous cues and their sensitivity to the field. This model extension is needed because experiments by Wolf *et al.* [1] suggest that cells at the tissue edges appear to be less responsive to the field than cells in the bulk. How this weaker sensitivity to the field then interacts with the endogenous cues inducing outward migration at the edges is not understood. Therefore, to understand the spatial and temporal heterogeneity of the responses in the tissue to electric field stimulation, there is a need to have detailed and reliable modelling of how cells integrate cues. Taking the insights from a continuum model of cue competition in different tissue regions, we develop a novel reaction-convection-diffusion model for collective electrotaxis that can predict the movement of an entire MDCK monolayer as it undergoes electrotaxis. This continuum model can readily be used to design bespoke electric fields for desired migratory patterns. We also use the model to predict how tissue size and geometry affect the collective migration of MDCK monolayers during electrotaxis.

This work is structured as follows. In Section 5.2 we briefly present the data from Wolf *et al.* [1] on electrotaxis in different regions of MDCK epithelial monolayers and introduce some of the most notable findings regarding the experimentally observed spatial heterogeneities. We proceed in Section 5.3 by analysing the question of how different migratory cues are interpreted together to give rise to different velocity distributions in different regions in the tissue. In Section 5.4 we use these insights to derive and validate a diffusion-convection-reaction equation that describes the movement of an epithelial monolayer undergoing collective electrotaxis. We use this continuum model to propose several ways in which electric fields can be designed such that they give rise to a desired spatial and temporal pattern of collective migration.

5.2 Experimental data and methods

Experimental data are publicly available from Wolf *et al.* [1]. In these experiments, a total of nine MDCK epithelial monolayers were grown to confluence and stimulated with an electric field of 3V/cm with the experimental protocol described in Chapter 4. Velocities during the experiment were computed in the source publication [1] using PIV with the same computational implementation as described in Chapter 4. PIV data was acquired in each of the different tissue regions in Figure 5.1 (left). PIV data from the experiment for each tissue are shown in Figure 5.1 (right). For the computation of local cell densities, we used the segmented cell nuclei data available from Wolf *et al.* [1] and computed two-dimensional histograms with a grid of size 150×150. A one-dimensional density profile, averaged in the y-direction, in the leading and trailing edge of a representative tissue, is shown in Figure 5.1 (bottom row).

The experimental velocity data in Figure 5.1 (right) contain a number of notable features. The first notable feature is that each of the edges has its own representative behaviour, with the trailing, leading, and top regions each having a different velocity trace, which are all distinct from the bulk. The second notable feature is that the edges of the tissue all exhibit a far smaller velocity during electrotaxis than in the tissue bulk. Interestingly, all the edge regions experience a slowdown during electrotaxis over a time scale similar to that of the slowdown in the bulk. The final notable feature is that both the top edge and the trailing edge seem to experience *recoil* after the electric field is turned off. That is, for a period of time after the electric field is turned off, the velocity in the top edge is negative, and the velocity in the trailing edge exhibits a sharp undershoot with respect to its velocity during electroxais and free epithelial expansion. This finding was also reported by Wolf *et al.* [1], who related it to mechanical factors. In the remainder of this chapter, we will explore these three phenomena and investigate the possible mechanisms responsible for creating them.

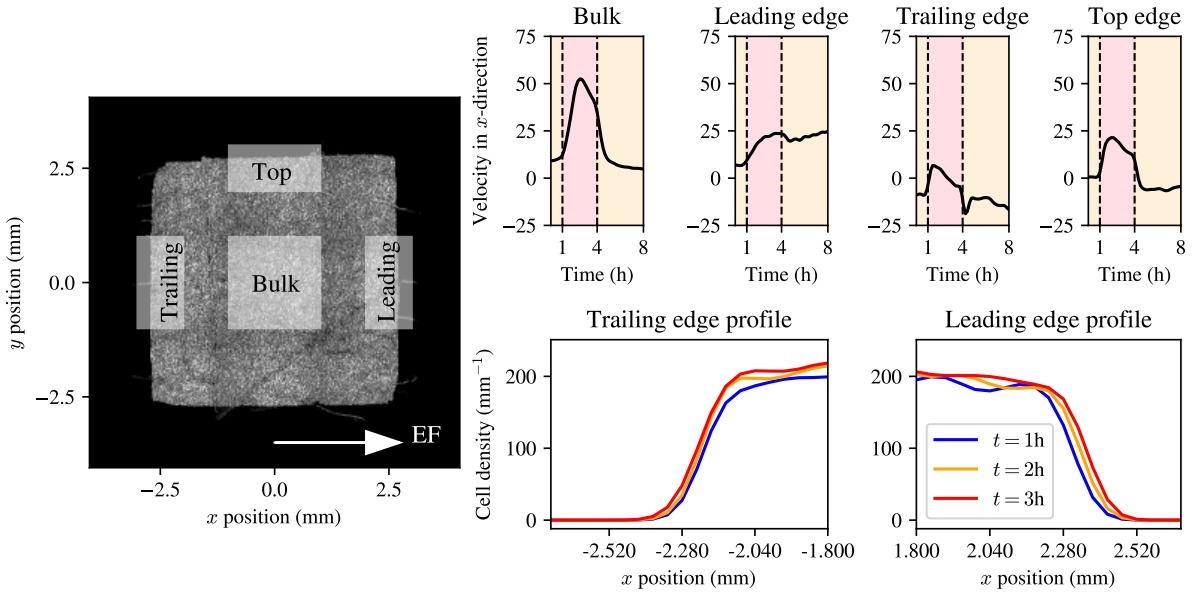


Figure 5.1: Collective electrotaxis data from Wolf *et al.* Left panel: phase image of a representative square MDCK-II epithelial monolayer, with different tissue regions marked. Top row: average velocity data in the bulk, leading edge, trailing edge, and top edge, respectively. Red shaded times indicate electric field on. Bottom row: cell density profile in space at three time points during the experiment.

5.3 Modelling spatial heterogeneity during electrotaxis

Having understood that the edge regions of the tissue exhibit velocity dynamics that are different from those of the bulk of the tissue, in this section we propose a model extension from the ODE model in Chapter 4 to account for the observed heterogeneity in velocities between the different tissues. In Section 5.3.1 we propose the model extension as a system of ODEs describing the different tissue regions. In Section 5.3.2 we perform Bayesian inference on the model, with a view to understanding how well the model can describe the experimental data.

5.3.1 An ODE model for vector superposition of competing migratory cues

Figure 5.1 shows that different regions of the epithelial monolayer migrate at different speeds during electrotaxis: the bulk of the tissue migrates at the highest speed of all tissue regions; and the trailing and top edges migrate slower than the leading edge. The fact that different regions of the monolayer migrate at different speeds should not be surprising, given that each tissue region is subject to different migratory cues: MDCK-II epithelial

monolayers are well-known to exhibit persistent spreading behaviours, even in the presence of electrotactic or chemotactic cues [1, 39, 40]. As such, there is a competition between different cues inducing migration in different directions at every tissue location, and tissue migration is the result of how these cues are interpreted together.

In order to construct a continuum model of collective migration during electrotaxis, we first seek to understand the mechanisms driving these differences in observed migratory velocity. We do this by modelling the average velocity of the different tissue regions using an extension of the one-dimensional ODE model for the velocity in the direction of the electric field derived in Chapter 4. To extend the model introduced in Chapter 4, we formulate a phenomenological ODE model for the velocity in the direction of the field in each of the edge regions. In Chapter 4, the velocity of the bulk of the tissue was modelled by the system in Equation (4.8), which described the average acceleration of the bulk of the tissue as a balance of active forces in the direction of the electric field against forces resulting from friction with the substrate. In the tissue bulk, this simple force balance is an appropriate model of the constituent forces since, on average, cells do not exert forces in any preferential direction. At the edges of MDCK-II epithelial monolayers, however, it is well-characterised that cells in a region within $\sim 100\mu$m of the edge exert active forces in the direction of the normal at the tissue boundary [81]. For this reason, a first extension in this model of acceleration in different tissue regions is to include an active force per unit mass, $\mathbf{F}_{\text{edge}}$, in the direction of the normal at the tissue edge, *i.e.*,

$$\mathbf{F}_{\text{edge}} = \alpha_{\text{edge}} \hat{\mathbf{n}}, \tag{5.1}$$

where $\hat{\mathbf{n}}$ is a unit vector and α_{edge} is a constant that controls the magnitude of the force per unit mass. Note here that in the specific example of the square tissues considered in this study, the direction of the outward normal is constant throughout each of the edge regions.

A second extension we formulate takes into account cell polarity. Both *in vivo* and *in vitro* cell polarity plays an important role in collective migration, since the physical orientation of cells in collectives has an effect on the magnitude and the direction of active force generation [4, 132–134]. At the tissue edges with a normal vector perpendicular to the direction of the electric field, *i.e.*, the top and bottom edges, cells have a preferred polarity in the direction of the normal due to free-space sensing of cells at the edge, as well as contact inhibition of locomotion [1, 39, 106]. Given that cell polarity in the edge regions whose normal is perpendicular to the electric field might influence the orientation of the cytoskeleton, and this might have a direct effect on the magnitude of the active force in the direction of the electric field, we propose a model extension to take this into account. In the original model in Equation (4.8), the magnitude of the active force per unit mass was given by a single constant, α. As a model extension in this chapter, we

allow the magnitude of the active force per unit mass to vary, depending on the position in the monolayer. To do so, we introduce a constant for the magnitude of the field response in regions of the tissue whose cell polarity is parallel to that of the electric field, which we denote by $\alpha_{\parallel}$, and another for the tissue regions whose polarity is perpendicular to that of the electric field, which we denote by $\alpha_{\perp}$. Since cells in the tissue bulk have no preferential polarity and polarise quasi-instantaneously in the direction of the electric field upon electric field application [1, 39], we model active forces per unit mass in the direction of the electric field as

$$\mathbf{F}_{\text{active}} = \alpha_{\parallel} s_{\text{eff}}^{+} \hat{\mathbf{s}}, \tag{5.2}$$

for the bulk of the tissue, where here $\hat{\mathbf{s}}$ is a unit vector pointing in the direction of the electric field, which we assume to be constant, uniaxial and spatially uniform. This framework can be naturally extended to time-varying and spatially heterogeneous electric fields. Finally, to model the apparent phenomenon that cells at the tissue edges are less sensitive to the electric field than cells in the bulk, we model that the active forces in the leading and trailing edge regions are given by

$$\mathbf{F}_{\text{active}} = \delta \alpha_{\parallel} s_{\text{eff}}^{+} \hat{\mathbf{s}}, \tag{5.3}$$

and by

$$\mathbf{F}_{\text{active}} = \delta \alpha_{\perp} s_{\text{eff}}^{+} \hat{\mathbf{s}}, \tag{5.4}$$

for the top region. Here, $\delta \in [0, 1]$ is a constant which scales the active force per unit mass in the direction of the electric field to take into account that cells at the tissue edges are less sensitive to the electric field. Putting the forces per unit mass in dynamic equilibrium finally yields that

$$\gamma \dot{\mathbf{v}} = \mathbf{F}_{\text{friction}} + \mathbf{F}_{\text{edge}} + \mathbf{F}_{\text{active}}, \tag{5.5}$$

for each of the tissue regions as described above. In this case, we have assumed that the different migratory cues act additively on the monolayer regions, as we take the resultant force to be the sum of each of the active forces. As in the previous chapter, we take the friction force per unit mass to be linear in the velocity. Taking components in the direction of the electric field, *i.e.*, in the x-direction, gives a system of one-dimensional ODEs for

γ (h^{-1})	$\alpha_{\parallel}$ ($\mu m/h^2$)	$\alpha_{\perp}$($\mu m/h^2$)	δ	α_e ($\mu m/h^2$)
1.765	171.95	217.08	0.31	24.93

Table 5.1: Posterior means for the extended edge velocities model, Equations (5.6)-(5.11). Note that the value for the velocity decay rate, γ, like in the previous chapter, is estimated directly from the data.

the velocities in each of the different tissue regions:

$$\dot{v}_{\text{bulk}} = -\gamma v_{\text{bulk}} + \alpha_{\parallel} s^{+}_{\text{eff}}, \tag{5.6}$$

$$\dot{v}_{\text{leading}} = -\gamma v_{\text{leading}} + \delta\alpha_{\parallel} s^{+}_{\text{eff}} + \alpha_{\text{edge}}, \tag{5.7}$$

$$\dot{v}_{\text{trailing}} = -\gamma v_{\text{trailing}} + \delta\alpha_{\parallel} s^{+}_{\text{eff}} - \alpha_{\text{edge}}, \tag{5.8}$$

$$\dot{v}_{\text{top}} = -\gamma v_{\text{top}} + \delta\alpha_{\perp} s^{+}_{\text{eff}}, \tag{5.9}$$

$$\tau_e \dot{s}_{\text{eff}} = s - (s_{\text{eff}} + I), \tag{5.10}$$

$$\tau_a \dot{I} = s - I, \tag{5.11}$$

where, like in the previous chapter,

$$s^{+}_{\text{eff}} = \max(s_{\text{eff}}, 0), \tag{5.12}$$

and s is the field strength in the x-direction.

5.3.2 Bayesian inference of model parameters in the superposition model

In this section, we aim to validate the hypotheses presented in the previous section by calibrating the model in Equations (5.6)-(5.11) to the experimental data from Wolf *et al.* [1] presented in Section 5.2. As in Chapter 4, we fix the velocity decay parameter, γ, from the data directly as $\gamma = 1.765 h^{-1}$. Since the bulk stimulation dynamics in the simple adaptation-excitation model of Chapter 4 gave predictions that were in excellent agreement with the experimental data, we equally fix the values for $\tau_a = 2.04$h and $\tau_e = 0.26$h. We perform Bayesian inference of the remaining parameters, $\alpha_{\parallel}, \alpha_{\perp}, \delta$, and α_{edge}, using Markov Chain Monte Carlo with a Haario-Bardenet adaptive covariance under the assumption of a Gaussian error implemented in PINTS [129], with the same implementation parameters as in Chapter 4. The numerical integration scheme used for simulating from the model is identical to that of Chapter 4. The resulting posterior means are given in Table 5.1.

The top row in Figure 5.2 shows the resulting marginal posterior distributions for these estimated model parameters. It can be seen that the marginal posterior distributions are all well-identified and concentrated around their means, showing that, given the available experimental data, the model parameters can be accurately estimated. The posterior means reported in Table 5.1 suggest that the three different edge regions are indeed less sensitive to the field, given the small posterior mean value of δ. Note also that $\alpha_{\parallel}$, the strength of the active forces in the direction of the electric field, is greater than $\alpha_{\perp}$, suggesting that cells with a preferred polarity that is not in the direction of the electric field have a stronger response to the electric field than the cells whose preferred polarity, and therefore also that of their competing migratory cues, is in the direction of the electric field.

To understand the validity of the model in reproducing the experimental data of Wolf *et al.* [1], and hence the underlying assumption that the edges have a weaker response to the field on the one hand, and that the field combines with edge endogenous migratory cues in an additive way, on the other, we calculate the posterior predictive intervals for the data resulting from the posterior distributions of the model parameters. These intervals include the middle 95% percentiles of model values found by simulating using parameters distributed according to the posterior distribution. Inspection of these posterior predictive intervals show that the model predictions are generally in good agreement with the experimental data and that the model can explain the different behaviours of the bulk of the tissue and the edges during stimulation. An important feature of the experimental data in Figure 5.2 is that the top edge and the trailing edge display an 'undershoot' of the velocity [1] after stimulation is halted, *i.e.* the average velocity in the direction of the field in the top edge is negative for several hours post-simulation. We comment on this phenomenon in the discussion, Section 5.6, where we show that it has important implications for how to model cellular transduction of the electric field signal as a function of position inside the monolayer.

Finally, we remark that, at the leading edge, Figure 5.2 shows that the model moderately overestimates the velocity as a result of vector superposition of the electric field cue together with the endogenous cue to migrate rightward and out of the tissue. This shows us that, while the model extension presented in this section can capture large scale features, there are some finer details that are not well explained. Addressing these details will require a finer quantification of velocity, force and polarity at the monolayer edge and should be the focus of new research into collective migration during electrotaxis.

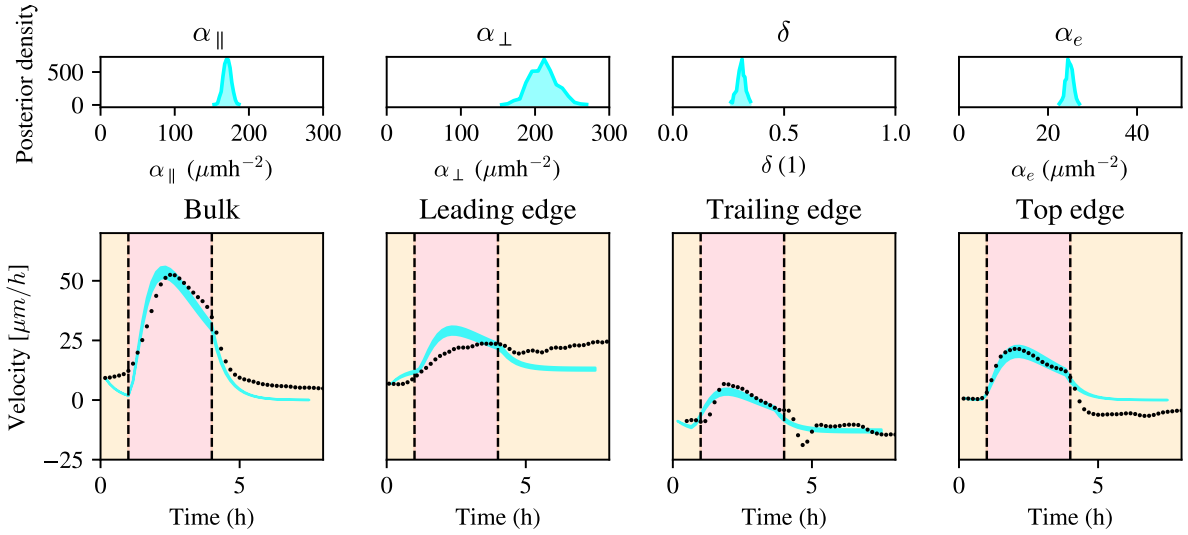


Figure 5.2: Bayesian inference with a model using vector superposition of the electric field with endogenous cues according to Equations (5.6)-(5.11). Top row: posterior distributions for model parameters α, δ, and α_e, respectively, from left to right. Bottom row: posterior 95% confidence interval for the vector superposition model in cyan with experimental data (black scattered points).

5.4 Construction and validation of a continuum model for collective electrotaxis

So far, we have studied control of collective electrotaxis in MDCK monolayers by controlling the velocity of the center of mass (in Chapter 4), or by describing collective electrotaxis by the velocities of distinct regions in the tissue (*i.e.*, the bulk and the edges in the previous section). In practical applications of electrotaxis, however, one is interested in moving the tissue in its entirety to a desired final spatial extent. For this reason, understanding how the entire monolayer moves is of the greatest interest in controlling collective electrotaxis in epithelial monolayers. For this reason, we will focus in this section on developing and validating a continuum model for collective migration during electrotaxis that describes the evolution of monolayer velocities. Using the coupled system of Equations (5.6)-(5.11), we found that the behaviour of the edges during electrotaxis can be described as the result of reduced sensitivity to the electric field together with linear superposition of the cue to migrate outward at the edges. In this section, we seek to use these insights to formulate and analyse a continuum model for cell migration.

5.4.1 Model development

The goal of this section is to formulate a continuum model for the evolution of cell density during electrotaxis using a reaction-diffusion equation. At a continuum level, the migration

of a population of cells is commonly modelled using a PDE for cell densities, ρ, of the form

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\mathbf{J}(\rho)) + f(\rho), \tag{5.13}$$

where $\mathbf{J}$ represents the *flux* of cells, and f is a function that represents cell proliferation [135]. Continuum models of cell migration are widely used [135] and provide physically or phenomenologically derived expressions for the fluxes and proliferation terms in Equation (5.13). For instance, in the context of chemotaxis, the standard Keller-Segel model decomposes the flux into a diffusive term due to random motion and one due to chemoattraction, *i.e.*,

$$\mathbf{J}(\rho) = \mathbf{J}_{\text{diffusion}}(\rho) + \mathbf{J}_{\text{chemoattraction}}(\rho), \tag{5.14}$$

and a wide range of models exists to describe the different (nonlinear) diffusion and chemoattraction terms. Likewise, many different forms exist for the proliferation function, f [135]. In this work, we will assume the flux due to random movement of cells to be given by non-linear diffusion of the form

$$\mathbf{J}_{\text{diffusion}}(\rho) = D\rho\nabla\rho, \tag{5.15}$$

where D is a diffusion coefficient. This model has been previously shown to well describe the edge spreading of epithelia [136]. We also choose the proliferation term to be described by the standard logistic growth model, which is a canonical model for cell proliferation and takes the form

$$f(\rho) = r\rho\left(1 - \frac{\rho}{K}\right), \tag{5.16}$$

where p is the proliferation rate, and K is the carrying capacity [135]. Having fixed the flux due to random cell movement and the proliferation function, we turn to proposing a model for the tissue flux due to electrotaxis, which is the main contribution of this chapter. In his seminal work, Murray, drawing an analogy with chemotaxis, proposed that the flux due to electrotaxis be represented as

$$\mathbf{J}_{\text{electrotaxis}}(\rho) = \rho G(\phi)\nabla\phi, \tag{5.17}$$

where ϕ is an electric potential, and G is a function that encodes the strength of the response of cells to the electric field. This model assumes that the response to the electric field is in the direction of the gradient of the potential, and that the strength of the response is dependent on the local magnitude of the potential. While the first assumption

is reasonable, given the experimental evidence presented in this chapter, the preceding analysis demonstrating that the tissue edges are less sensitive to the field gives an indication that edge-effects must be included in a continuum model for collective electrotaxis. In a continuum model describing tissue density, the difference in density at the edge versus in the bulk gives a good proxy to quantitatively establish the location of the tissue edge. For this reason, we incorporate the spatial heterogeneity in cell responses to the electric field by choosing a functional form that makes responses to the electric field be strongest in the bulk of the tissue, where $\rho \approx K$, and weakest at the boundary where $\rho \approx 0$. Since calibrating the system of Equations (5.6)-(5.11) shows that the ratio of the sensitivity at the edges to that in the bulk can be described by a fixed parameter, δ, a simple model for the dependence of the electrotaxis flux on local cell density is given by

$$\mathbf{J}_{\text{electrotaxis}}(\rho) = \rho\alpha s_{\text{eff}} \left(\delta + \frac{\rho}{K}(1-\delta)\right) \nabla\phi, \tag{5.18}$$

such that

$$\mathbf{J}_{\text{electrotaxis}}(\rho)/\rho \approx \alpha s_{\text{eff}} \nabla\phi, \tag{5.19}$$

in the tissue bulk and

$$\mathbf{J}_{\text{electrotaxis}}(\rho)/\rho \approx \delta\alpha s_{\text{eff}} \nabla\phi, \tag{5.20}$$

at the tissue edges, which corresponds with the analysis in the Bayesian inference of Equations (5.6)-(5.11). We note here that s_{eff} is now a spatially varying field, described using Equations (5.22) and (5.23). Bringing together the fluxes and the proliferation term, we propose the following continuum model for collective electrotaxis,

$$\frac{\partial\rho}{\partial t} = \nabla \cdot \left[\rho D \nabla\rho + \rho\alpha s_{\text{eff}} \left(\delta + \frac{\rho}{K}(1-\delta)\right) \nabla\phi\right] + r\rho\left(1 - \frac{\rho}{K}\right). \tag{5.21}$$

So far, for the evolution of the effective signal, we have used the governing equations in Equations (5.6)-(5.11), since the field was assumed to be uniaxial. In a fully two-dimensional model, however, the electric field can locally point in any arbitrary direction, and its magnitude can also vary in space. For that reason, we must define an extension for the governing equations for the effective signal, s_{eff}, that takes into account a vector field for the velocity, whose magnitude can vary in space. Here, we make the assumption that the strength of the effective signal only depends on the magnitude of the electric field, and so rewrite the governing equations for the effective signal and the inhibitor

$$\tau_e \dot{s}_{\text{eff}} = \|\mathbf{s}\| - (s_{\text{eff}} + I), \tag{5.22}$$

$$\tau_a \dot{I} = \|\mathbf{s}\| - I. \tag{5.23}$$

To conclude, we have proposed Equations (5.6)-(5.11) as a model extension to the model in Equation (4.8) that takes into account several edge effects at the tissue boundaries. This model contains three additional parameters: the strength of the response to the electric field when the preferred polarity of the tissue is perpendicular to the electric field, $\alpha_\perp$; the difference in field responsiveness, δ; and the strength of the endogenous cue, α_{edge}.

5.4.2 Bayesian inference of model parameters

We wish to calibrate the model in Equation (5.21) to the experimental cell density data of Wolf *et al.* [1]. However, fitting a full two-dimensional density model to the data is challenging due to the very large local fluctuations in cell density, which are characteristic of MDCK epithelial monolayers during expansion [91, 96]. For this reason, we select a region consisting of a horizontal strip of the monolayer which is one mm away from the top and bottom edges and average in the y-direction. This averaging removes any artefacts that arise from migration from the top and bottom edges and provides an averaged one-dimensional density profile across the monolayer. In one dimension, Equation (5.21) reduces to

$$\frac{\partial \rho}{\partial t} = \frac{\partial}{\partial x}\left[\rho D \frac{\partial}{\partial x}\rho + \rho \alpha s_{\text{eff}}\left(\delta + \frac{\rho}{K}(1-\delta)\right)\frac{\partial}{\partial x}\phi\right] + r\rho\left(1 - \frac{\rho}{K}\right), \tag{5.24}$$

subject to initial conditions given by the experimentally measured tissue density at $t = 0$h, and Neumann boundary conditions. The domain is the interval spanned by the x-coordinate in the field of view of the experiment shown in Figure 5.1, *i.e.*, the range -3.5mm to 3.5mm from the tissue centre. The full parameter set, θ, for the model in Equation (5.24) is given by

$$\theta = (\tau_a, \tau_e, \alpha, D, \delta, r, K). \tag{5.25}$$

In the experiments of Wolf *et al.* [1], the field was uniaxial and constant in space and time, meaning that the temporal dynamics of the effective signal, s_{eff}, are homogeneous in time and are identical to those described in Section 5.3. For this reason, since the parameters governing the temporal dynamics of the effective signal, τ_a and τ_e, could be confidently identified given the data, we choose the posterior means for $\tau_a = 2.04$h and $\tau_e = 0.26$h from Section 5.3 and perform Bayesian inference on the remaining model parameters, α, D, δ, r, and K.

Using the same Bayesian inference procedure as in Section 5.3, we obtain the posterior distributions for the model parameters α, D, δ, r, and K, with their posterior means recorded in Table 5.2. The marginal posterior distributions of the model parameters are shown in Figure 5.3; all model parameters can be confidently estimated given the

α (mm/h)	D (mm^{2}h^{-1})	δ	r (h^{-1})	K (mm^{-1})
0.837	13.30	0.625	0.180	260.29

Table 5.2: Posterior means for the extended edge velocities model. Note that the value for the velocity decay rate, γ, like in the previous chapter, is estimated directly from the data.

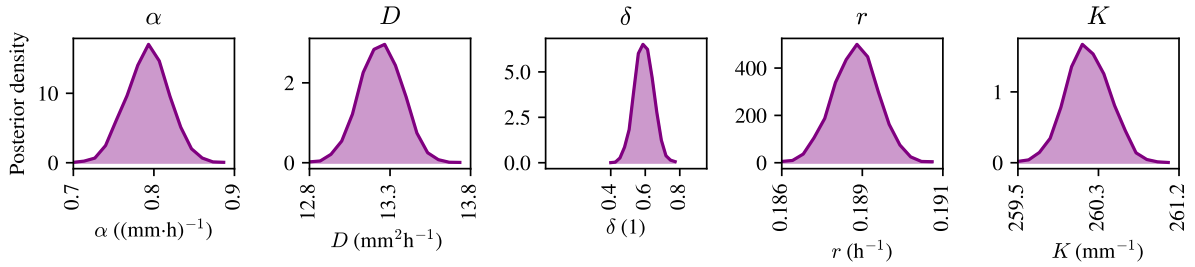


Figure 5.3: Marginal posterior distributions for a one-dimensional continuum model of collective electrotaxis in Equation (5.24).

experimental data with each of the distributions supported on a narrow interval around its mean.

To understand the quality of the resulting model predictions given the posterior distributions over model parameters, we simulate from the model in Equation (5.24) using the posterior means of the distributions shown in Figure 5.3, using the initial condition from the data as the initial condition in the simulations. Figure 5.4 shows a comparison between the experimental data and the predictions generated by the model. The kymographs in Figure 5.4 show a density profile across the tissue that is in excellent agreement with the experimental data. Figure 5.4 also shows the profiles of the leading and the trailing edge at one, two, and three hours during electrotaxis as simulated by the model, compared to the experimental data, again showing excellent agreement between the model predictions of the densities and the experimental data. We conclude that the model in Equation (5.24) can faithfully capture the tissue migratory behaviours of the monolayer during electrotaxis, and can be used to predict cell densities during collective migration when the tissue is subject to an electric field.

5.4.3 Tissue size affects migratory speed

The model in Equation (5.21) provides a good description of the evolution of cell densities during electrotaxis. In this section, we outline how the model can be used to obtain predictions for the local migratory velocity. We will use this framework to show how tissue size and geometry affect the magnitude, and distribution, of migratory velocity in MDCK monolayers. For a scalar field, c, a general expression for the convection-diffusion-reaction equation is given by

$$\frac{\partial c}{\partial t} = \nabla \cdot (D \nabla c) - \nabla \cdot (\mathbf{v} c) + f(c), \tag{5.26}$$

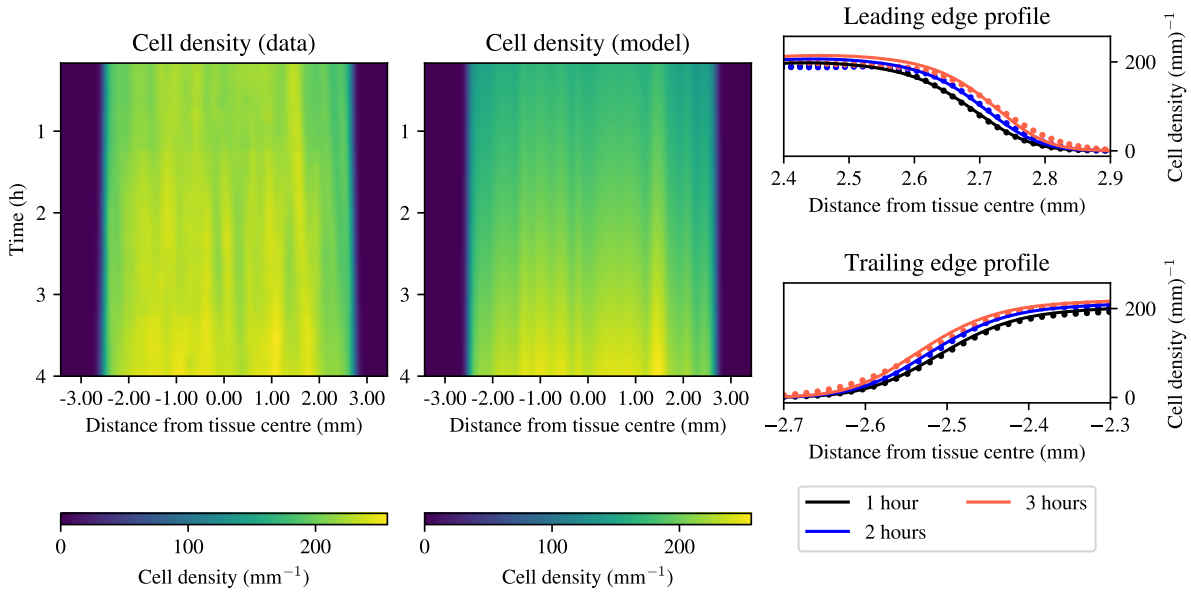


Figure 5.4: Comparison between model predictions and experimental data of one-dimensional cell density profiles during collective electrotaxis. Left panels: kymographs for cell density (experimental data left, model solutions right). Right panel: profiles of the leading and trailing edge at one, two, and three hours of electrotaxis.

where D is the diffusion coefficient, $\mathbf{v}$ is the advective velocity, and $f(c)$ is a reaction term, which in the context of epithelia can represent cell proliferation and death. Inspection of Equation (5.21) shows that the velocity is given by

$$\mathbf{v} = D\nabla\rho + \alpha s_{\text{eff}} \left(\delta + (1-\delta)\frac{\rho}{K}\right) \nabla\phi. \tag{5.27}$$

The expression in Equation (5.27) allows us to compute the velocity field from simulations of the model in Equation (5.21), and it allows us to probe the role of tissue geometry and size in the resulting tissue velocities during collective electrotaxis.

To study the velocity of the monolayer during electrotaxis, we simulate tissues with two different geometries: circular and triangular. These shapes provide a good starting point for understanding the role of tissue geometry, given that they have curvature and acute corners. For each shape, we simulate tissues with a diameter between 0.25mm and 4mm, corresponding roughly to the range between the smallest tissue sizes in which directional collective electrotaxis has been reported [137] and twice the size of the experiments of Wolf *et al.* [1], thus providing a good range of realistic tissue sizes for collective electrotaxis. For each tissue shape, we generate a grid of tissue diameters consisting of 100 equally spaced points between this minimum and maximum tissue diameter, and compute the maximum electrotaxis velocity in the simulation, using the means of the posterior distributions of the model parameters obtained in Figure 5.3. For all model simulations, a uniform initial condition of $\rho = K$ is used. We solve Equation (5.21) numerically in two spatial

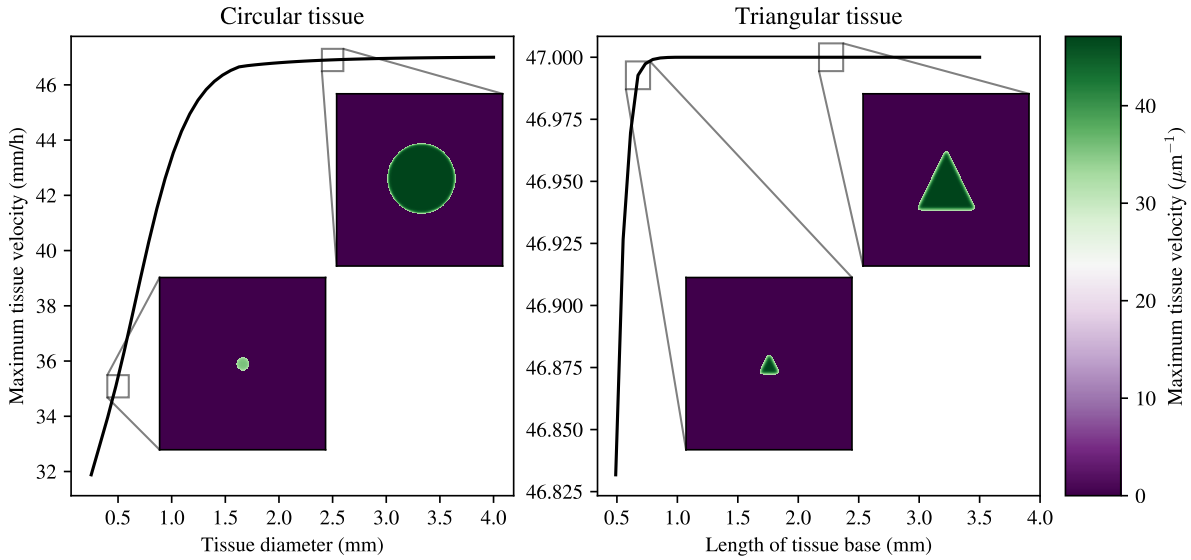


Figure 5.5: Effect of tissue size and geometry on maximum migratory velocity during electrotaxis. Insets show circular (triangular) tissues with a tissue base of 0.5mm and 2.5mm, respectively. The spatial distribution of velocities is qualitatively different in the circular tissues, but not in the triangular tissues.

dimensions with the finite-volume numerical scheme described in [136, 138].

Figure 5.5 shows that tissue size has an impact on the maximum velocity during electrotaxis: with increasing tissue size, the maximum velocity during electrotaxis increases before reaching a plateau. The scales of the variation in tissue velocity are vastly different between the two different geometries, as shown by the limits of the y-axes in Figure 5.5. The difference in the maximum velocity between the smallest triangular tissues and the largest triangular tissues amounts to less than $0.5\mu\text{mh}^{-1}$, whereas it is more than $10\mu\text{mh}^{-1}$. The insets in Figure 5.5 show that the spatial distribution of velocities is qualitatively different in the circular tissues of different sizes, with the large circular tissue and both triangular tissues having a well-defined edge with a lower magnitude velocity. In summary, we find that tissue geometry geometry, together with tissue size, informs the distribution, and magnitude of the velocity in the electrotaxing monolayer.

5.5 Optimal design of electric fields to achieve spatially uniform migration velocity

Having understood that the continuum model from Equation (5.21) provides an expression for the tissue velocity, $\mathbf{v}$, through Equation (5.27), we can turn to the question of how to use these insights to design bespoke electric fields to achieve a target spatial velocity distribution during collective electrotaxis. While the possible applications of electrotaxis to

guide collective cell migration are myriad [40], we focus on one specific task in controlling collective migration, which is to create a spatially uniform migratory velocity. Our motivation for this is that overcoming the large differences in migration velocities between the edges and the bulk of the tissue, which are a result of the interplay between endogenous and exogenous cues that cells at different tissue locations receive, would represent an important first example of spatial control of collective velocity during electrotaxis. In this section, we first investigate the spatial distribution of an optimal electric field predicted by a one-dimensional model, which can be analysed analytically, and conclude with two numerical examples in two dimensions to demonstrate practical applications of the method.

5.5.1 Physical insights from a one-dimensional model

Recall that in one dimension, the electrotaxis model for the evolution of cell densities is given by Equation (5.24), and the velocity can be found analogously to that in Equation (5.27),

$$v = D\frac{\partial}{\partial x}\rho + \alpha s_{\text{eff}}\left(\delta + \frac{\rho}{K}(1-\delta)\right)\frac{\partial \phi}{\partial x}. \tag{5.28}$$

We now turn to the following problem. Given a target velocity, $v^\star > 0$, design an electric potential, ϕ, such that the monolayer velocity is uniform, $v = v^\star$. When $s_{\text{eff}} > 0$, Equation (5.28) rearranges to

$$\frac{\partial \phi}{\partial x} = \frac{1}{\alpha s_{\text{eff}}\left(\delta + (1-\delta)\frac{\rho}{K}\right)}\left(v - D\frac{\partial \rho}{\partial x}\right). \tag{5.29}$$

While tractable, the expression for the field in Equation (5.29) comes with the limitation that it depends on s_{eff}, which is zero at $t = 0$. As we have discussed previously, s_{eff} also exhibits $s_{\text{eff}} \to 0$ as $t \to \infty$. Therefore, Equation (5.29) does not prescribe how to start stimulation at $t = 0$, and predicts unphysiologically strong electric potential gradients as $s_{\text{eff}} \to 0$ for large t. For this reason, the expression for the field in Equation (5.29) does not constitute a full optimal control solution. At this point, we also note that controlling the velocity, v, in Equation (5.29) so that it matches some desired velocity is not a straightforward problem, conceptually, or computationally due to the highly nonlinear coupling with the effective signal, s_{eff}. This is made more complicated by the fact that in applications more generally, one might want the target velocity, $v^\star$, to vary in space. For that reason, we make progress by simplifying the problem substantially. Instead of considering a full optimal control problem, we propose to overcome the difficulty that $s_{\text{eff}} = 0$ at $t = 0$ by stimulating with a constant electric field until s_{eff} is at a sufficient level that the predicted gradient in Equation (5.35) can be used. As s_{eff} wanes, we propose that

the electric potential gradient is capped by a maximal field strength, *i.e.*,

$$\frac{\partial \phi}{\partial x} = \max\left(\frac{1}{\alpha s_{\text{eff}}\left(\delta + (1-\delta)\frac{\rho}{K}\right)}\left(v - D\frac{\partial \rho}{\partial x}\right), s_{\max}\right), \tag{5.30}$$

for some $s_{\max} > 0$ that describes the maximum field strength that is physiologically tolerable. With this in mind, we can turn to understanding how the one-dimensional model of electrotaxis in Equation (5.24) can be used to obtain electric fields to achieve a constant velocity of migration throughout the monolayer. By substituting $v = v^\star$ for the velocity and integrating over the domain, one obtains the following direct expression for the potential,

$$\phi(x,t) = \int_0^x \frac{1}{\alpha s_{\text{eff}}\left(\delta + (1-\delta)\frac{\rho}{K}\right)}\left(v^\star - D\frac{\partial \rho(\xi,t)}{\partial \xi}\right)\mathrm{d}\xi. \tag{5.31}$$

Here, the lower limit of integration can be chosen arbitrarily, since the velocity in Equation (5.28) is only influenced by the gradient of the potential, $\partial_x \phi$, meaning that the constant of integration is arbitrary. This also is equivalent to the natural boundary condition

$$\phi(0,t) \equiv 0, \quad t > 0. \tag{5.32}$$

With this in mind, the problem of keeping the velocity constant in the monolayer reduces to

$$\phi(x,t) = \frac{1}{\alpha s_{\text{eff}} K(1-\delta)}\int_0^x \frac{v^\star - D\partial_\xi \rho(\xi,t)}{\frac{\delta}{K(1-\delta)} + \rho}\mathrm{d}\xi \tag{5.33}$$

$$= \frac{1}{\alpha s_{\text{eff}} K(1-\delta)}\left[\int_0^x \frac{v^\star}{\frac{\delta}{K(1-\delta)} + \rho(\xi,t)}\mathrm{d}\xi - D\int_0^{\rho(x,t)} \frac{1}{\frac{\delta}{K(1-\delta)} + \rho}\mathrm{d}\rho\right]. \tag{5.34}$$

Evaluating the final integral yields

$$\phi(x,t) = \frac{1}{\alpha s_{\text{eff}} K(1-\delta)}\left[\int_0^x \frac{v^\star}{\frac{\delta}{K(1-\delta)} + \rho(\xi,t)}\mathrm{d}\xi - D\ln\left(1 + (K\delta^{-1} - 1)\rho(x,t)\right)\right]. \tag{5.35}$$

The two terms in the potential defined in Equation (5.35) carry a convenient physical interpretation each. The first term in Equation (5.35) describes the changes in potential needed for a tissue element at position x to migrate at the same velocity as the trailing edge. The second term in Equation (5.35) corresponds to corrections in the field strength that arise from differences in the sensitivity of cells that are at the bulk or at the edge.

5.5.2 Design of electric fields for arbitrary shapes

In two dimensions, the coupling of velocity to local gradients in the electric potential, ϕ, and cell density can also be obtained explicitly from the equation for the tissue velocity, Equation (5.27). As in the previous section, one can set the problem of designing an electric potential, in this case in two dimensions, such that the entire tissue is moving at the same velocity. We denote this velocity by $\mathbf{v}^\star$. In a more general optimal control framework, the target velocity of migration, $\mathbf{v}^\star$, could vary in time and in space. In the present framework, whenever $s_{\text{eff}} > 0$, Equation (5.27) stipulates that the optimal gradient of the electric potential, which we denote by $\mathbf{s}_\star$ is given by

$$\mathbf{s}_\star = \frac{1}{\alpha s_{\text{eff}} \left(\delta + \frac{\rho}{K}(1-\delta)\right)} \left(\mathbf{v}^\star - D\nabla\rho\right). \tag{5.36}$$

As in the one-dimensional case discussed in Section 5.5.1, there are limits on the velocities that can be physically achieved, as the system adapts to the electric field signal: Equation (5.36) fails to hold when $s_{\text{eff}} = 0$ and predicts unphysically strong potential gradients in the limit as $s_{\text{eff}} \to 0$. At the same time, having a spatially varying field, $\mathbf{s}_\star$, makes the effective signal, s_{eff}, vary in space, thereby making the computation of the optimal field through Equation (5.36) significantly more complicated. Therefore, to avoid the complications that could arise from a spatially varying s_{eff} on the one hand, and possibly unphysically strong electric fields on the other, we propose a simple heuristic to achieve migration in the direction of the desired velocity. That is, we let the direction of the field vary, but not its magnitude. This comes with the limitation that while the resulting field will be in the direction of the optimal field predicted by Equation (5.36), the magnitude and the temporal distribution might not be optimal for the desired target velocity. We suggest this problem as a starting point for future research. For now, in order to make progress in the design of two-dimensional electric fields for controlling collective electrotaxis, the heuristic we propose is to choose a fixed magnitude for the electric field, in the direction of the optimal gradient, $s_\star$, as defined in Equation (5.36). Given a fixed strength of the field, $s_{\max}$, this amounts to setting the gradient of the potential, $\nabla\phi$, as

$$\nabla\phi = s_{\max}\hat{\mathbf{s}}_\star, \tag{5.37}$$

where $\hat{\mathbf{v}}^\star$, is a unit vector in the direction of the optimal field, $\mathbf{s}_\star$. This heuristic electric field provides a means to vary the electric field in space to effectively counteract the influence of local gradients in tissue density, which can cause the tissue to spread in directions other than that of the target velocity. As a proof of principle, we take as target velocity, $\mathbf{v}^\star$, a constant vector field in the x-direction with constant magnitude given by the maximum velocity of the bulk in Figure 5.1, *i.e.*, $47\mu\text{mh}^{-1}$. We take two sample

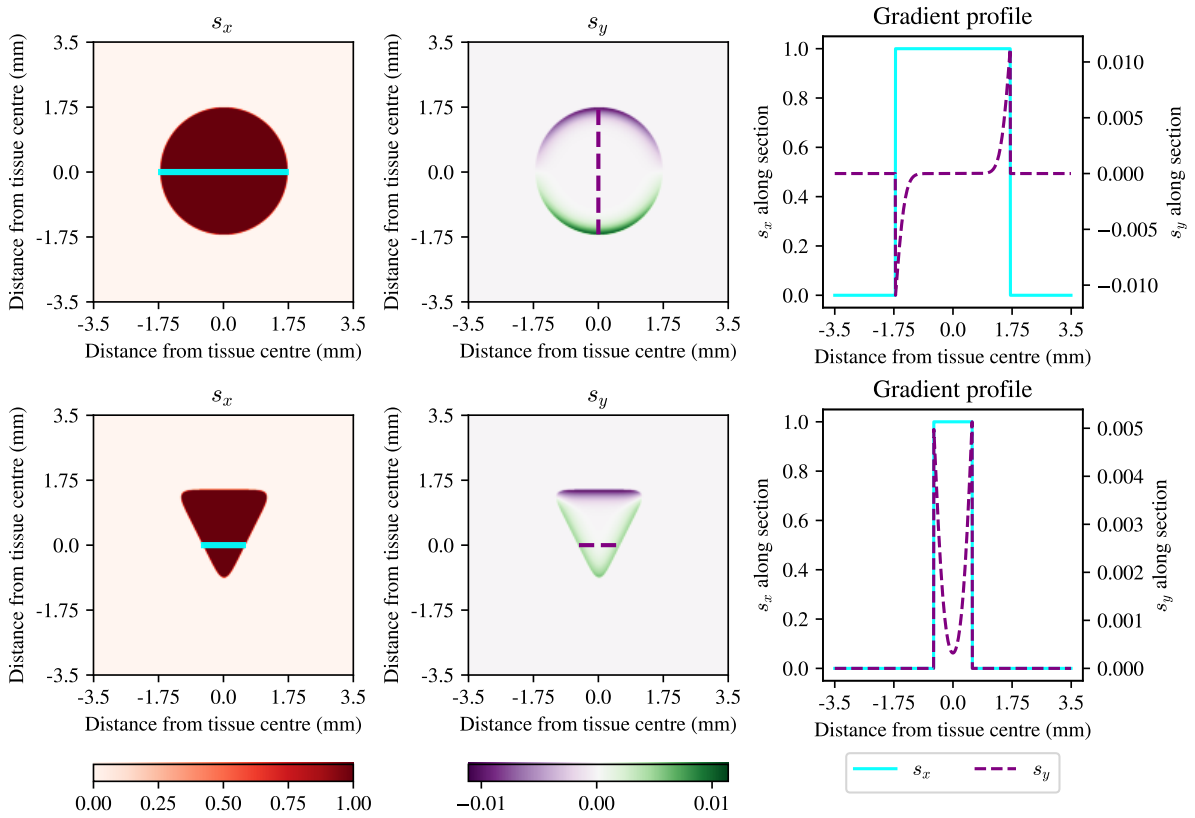


Figure 5.6: Spatial distribution of the electric field after one hour of electrotaxis to achieve uniform migration in the x-direction, for a circular tissue (top row) and a triangular tissue (bottom row). First column: x-component of the heuristic electric field. Second column: y-component of the heuristic electric field. Third column: one-dimensional profiles along solid cyan and dashed purple lines in the two-dimensional tissues, respectively.

geometries, one circular with a diameter of 3.5mm, and one triangular with a tissue base of 3mm, initialised using uniform initial condition of $\rho = K$, before applying the electric field suggested by Equation (5.37). After having observed that the field is near constant in time, we plot the gradient, $\mathbf{s}_\star$, in Figure 5.6.

Figure 5.6 shows that the electric field computed using the heuristic in Equation (5.37) mainly acts to override the outgrowth of the epithelium in directions other than the electric field. For example, in the circular tissue there is a clear edge region defined by having a y-component of opposite sign to that of the unit normal at the tissue boundary, with a similar phenomenon occuring in the triangular tissue. The one-dimensional profiles in Figure 5.6 show that the gradient acts against the outward migration of the epithelial monolayer in a small region at the tissue boundary.

5.6 Discussion and a starting point for future modelling

In this chapter, we have investigated the mechanisms behind the spatial heterogeneity in collective electrotaxis that was first observed in the experiments of Wolf *et al.* [1], and we have used these insights to develop a continuum model of collective electrotaxis. This model lends itself for the experimental design of bespoke electric fields that can be used to achieve specific aims during engineered collective cell migration. There are several ways in which our modelling approach can be enhanced for applications in bio-engineering on the one hand, and for a better understanding of the cell signalling and mechanics involved in establishing collective electrotaxis, on the other.

5.6.1 Spatially varying electric fields: consequences for modelling

The specific application of the continuum model to bio-engineering proposed in this work consisted of the design of an electric field to achieve uniform tissue velocity across the tissue. Our model can directly be applied in situations in which the desired velocity field is not spatially uniform, but varies in space. Such applications exist, for example, in wound healing, where one would want to move two opposite sides of a wound closer together to achieve healing. A second area of extension for our model would be to derive a full optimal control formulation for the problem involving the design of spatially and temporally varying electric fields. Formulating the design of spatially varying electric fields as an optimal control problem would overcome two restrictions that we introduced with the heuristic in Equation (5.37). First, an optimal control formulation would cast the design of an electric field as minimising a given objective function that can encode the desired monolayer displacement during electrotaxis. The resulting flexibility allows for a great range of possible electric field designs and applications. Second, having a coupled system of PDEs describing the optimal electric field would provide a stimulation protocol starting from $t = 0$ where $s_{\text{eff}} = 0$. This would avoid having to stimulate with an arbitrarily set field first to make sure that s_{eff} is large enough to use the field in Equation (5.36).

5.6.2 Understanding recoil in the top edge post-stimulation

The second, and final area of extension for this chapter concerns the phenomenon of undershoot at the top and trailing edges seen in Figure 5.1. While the results of Section 5.3.1 show that the system of ODEs given in Equations (5.6)-(5.11), which was an extension of the model presented in Chapter 4 can accurately represent the velocities of different tissue locations during electrotaxis, one of the phenomena in the data in Figure 5.1 that

is not well-explained by the model is the fact that the top and trailing edges display an undershoot of the velocity when the field is switched off. In this part of the discussion, we turn to understanding the phenomenon that the velocity in the top edge of the tissue becomes negative for several hours after the electric field is switched off. This finding was originally reported by Wolf *et al.* [1] who refer to it as recoil. In fact, recoil was also mentioned by Wolf *et al.* for the velocity at the trailing edge, which exhibits a sharp drop below its equilibrium value before restoring. On this phenomenon, Wolf *et al.* proposed that the observed recoil might be due to a mechanical effect induced by the visco-elastic properties of the epithelium. Although such a mechanical effect may indeed be at play at the trailing edge, we reason that it is not sufficient to explain the sustained negative velocity at the top tissue edge, as the characteristic time scale for the elastic response of MDCK epithelia is on the order of 15 to 30 minutes [87]. Therefore, the recoil as observed in the experiments of Wolf *et al.* [1] cannot be explained by visco-elastic mechanical effects alone in the top edge. We then ask the question: if mechanics are not responsible for creating recoil, might it be that cessation of the electric field signal itself could be inducing cells to move in the opposite direction of the electric field post-stimulation?

5.6.2.1 Model and constitutive equations

So far, we have assumed that the effective signal inside the constituent cells making up the monolayer is always positive given a positive field, as we have considered s_{eff}^{+} as in Equation (5.12). The effect of this is that the effective signal inside the cells can only give a migratory command in the direction of the electric field. To understand whether cessation of the electric field itself can provide a negative signal to the cells, we wonder if the model assumption that the effective signal is always in the direction of the field can be relaxed, *i.e.* by replacing s_{eff}^{+} by s_{eff} in the active force for the top edge. Since we only consider the dynamics of the top edge in this case, we can considerably simplify the equations at the top edge, since there is no need to compare the response of the top edge relative to that of the bulk. This means, for example, that there is no need for a separate δ parameter. Put together, we can study the reduced system of equations describing the dynamics of the top edge of the tissue,

$$\dot{v}_{\text{top}} = -\gamma v_{\text{top}} + \alpha s_{\text{eff}}, \tag{5.38}$$

$$\tau_e \dot{s}_{\text{eff}} = s - (I + s_{\text{eff}}), \tag{5.39}$$

$$\tau_a \dot{I} = s - I. \tag{5.40}$$

5.6.2.2 Bayesian parameter identification

As before, we keep the pre-fit value for γ constant, and seek to estimate the parameter values for α, τ_e, and τ_a using the same Bayesian inference procedure as in Section 5.3. We find that the posterior means of the model parameters are given by 0.103h, 6.11h, and 57.37h^{-1}, respectively. By computing the posterior predictive interval of the middle 95% of model solutions generated using parameter values drawn from the posterior distribution, we observe in Figure 5.7 that the model in Equations (5.38)-(5.40) is able to recapitulate the long-term undershoot of the average velocity in the top edge post-stim accurately. Given this excellent agreement with the experimental data, we turn to understanding the quality of the posterior distributions over the model parameters.

The marginal posterior distributions over the model parameters α, τ_a, and τ_e in Figure 5.7 show that while the field response, α, and the adaptation time scale, τ_a, can be confidently identified from the data, the excitation time scale, τ_e, is less confidently estimated. The marginal posterior distribution for the excitation timescale, τ_e, is broad and largely uninformative. This is different from the posterior distribution for τ_e in Section 5.3.1, where Figure 5.2 showed that τ_e did not suffer this problem. In this inference problem, while the posterior distribution at τ_e has a well defined maximum, it fails to concentrate its mass in a narrow interval around this maximum a posteriori (MAP) value. Given the fact that the model predictions obtained by simulating the model with parameters sampled from the posterior distribution are in excellent agreement with the experimental data and that the inference procedure previously yielded identifiable posterior distributions for the excitation timescale, τ_e, the question of what causes the non-identifiability of the τ_e parameter, and what this implies for the applicability of the model, warrants a closer look. For this reason, we turn to investigating the profile log-likelihood of the data under different model predictions with varying values of the model parameters.

To understand the influence of different values of the model parameters α, τ_a, and τ_e on the model predictions, we create a grid of size $50 \times 50 \times 50$ spanning the prior ranges shown for the distributions in Figure 5.7. For each parameter value in the grid, we perform one forward simulation of the model in Equation (5.38) and compute the sum of squared differences between the resulting model solution and the experimental data. Under the assumption of constant additive Gaussian noise, as was used in the MCMC implementations for Bayesian inference in this chapter, the sum of squared differences is proportional to the log-likelihood. Therefore, by computing the sum of squared distances on the grid of values, we can interrogate the profile likelihoods for the different model parameters. These profile likelihoods, in turn, allow us to understand the effect of each of the three individual parameters on the model solutions. In Figure 5.8 we show the pairwise loss functions, which are proportional to the bivariate log-likelihoods, as well as

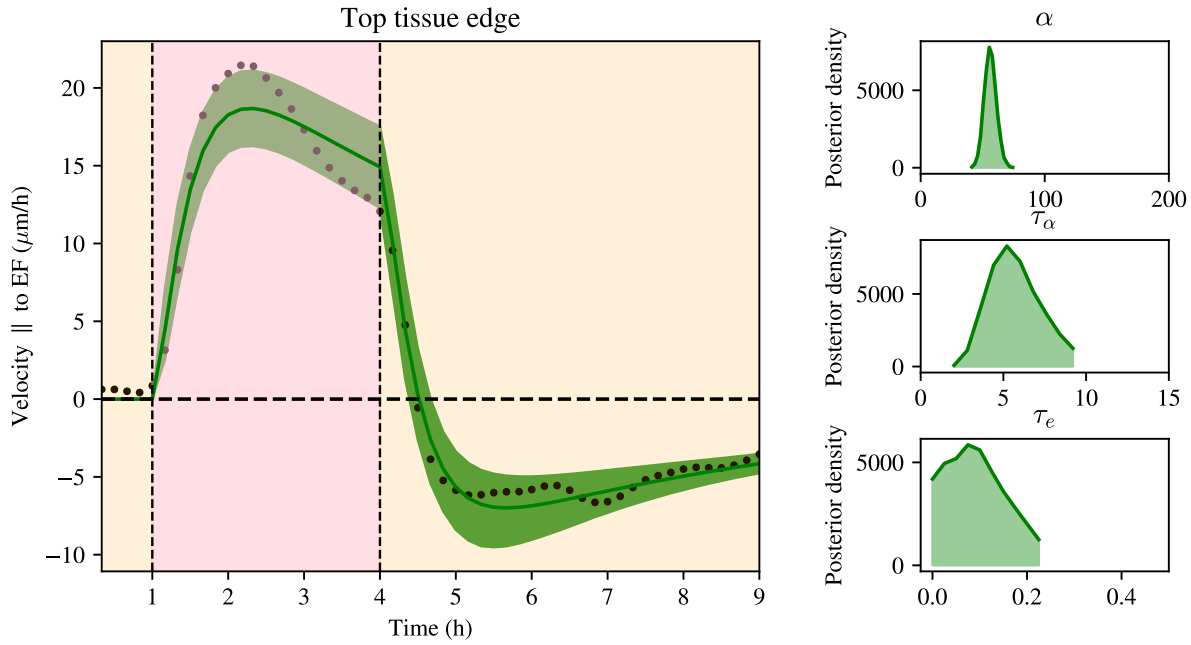


Figure 5.7: Bayesian inference for the top edge velocity model in Equation (5.38). Left panel: posterior 95% confidence interval for the top edge velocity model in green with experimental data (black scattered points). Right panel: posterior distributions for model parameters α, τ_a, and τ_e, respectively, from top to bottom.

the one-dimensional least squares profiles for each of the model parameters, which are equally proportional to the one-dimensional profile likelihoods.

The pairwise minimum loss landscapes in Figure 5.8, which show the minimum sum of squared errors when marginalised over the third parameter, show that there is a well defined minimum value for the sum of squared errors. This is also made clear in the one-dimensional profiles in Figure 5.8, which show the minimum sum of squared differences for each value of the different parameters. These profiles all have a clearly defined minimum, with the minimum least squares increasing rapidly upon deviations from this least squares optimum. Therefore, the profile likelihoods all point to a well-defined maximum likelihood, despite the posterior distributions in Figure 5.7 not being informative. For this reason, we ask why the posterior distribution fails to be informative of (some of) the model parameters, despite there being a well-defined maximum likelihood estimate of the parameters. The loss landscapes in Figure 5.8, show that the model parameters τ_a and τ_e are *sloppy*. For a dynamical system, parameters are called sloppy if their values of can be perturbed significantly without causing significant changes in the model output [139]. In the case of the squared distance profiles in Figure 5.8, it can be seen that large deviations in τ_e from the minimum of the bivariate plots still lead to similar values of the sum of squared errors.

We illustrate how the model is sloppy by fixing the values of α , τ_a, and τ_e at their maximum likelihood estimates from the the profiles in Figure 5.8 and simulating the model

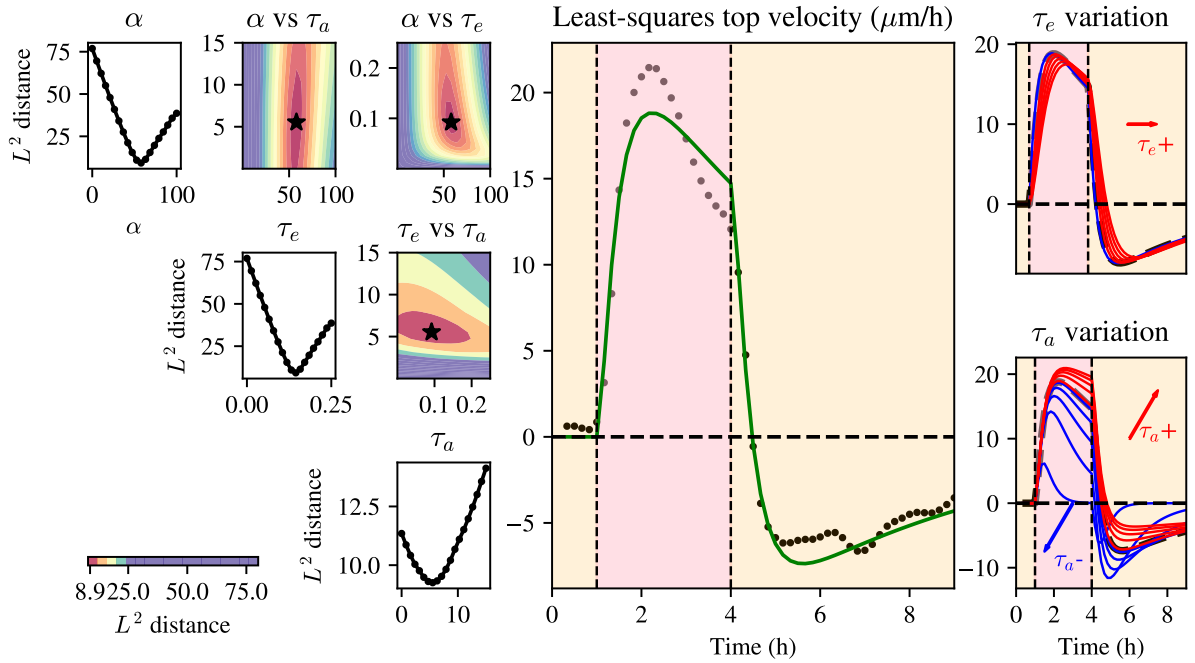


Figure 5.8: Analysis of model solutions on a parameter grid. Left panel: pairwise loss landscapes shown off-diagonal, with the colorbar denoting the L^2 distance between model solutions and parameter values (star denoting least squares parameter combination), marginal least squares profiles shown on diagonal. Middle panel: model solution corresponding to the least-squares parameter combination (green line) compared to experimental data (black points). Right panel: behaviour of model solutions as parameters τ_e and τ_a are varied.

according to changing values of τ_a and τ_e, as shown in the rightmost panel of Figure 5.8. While variations in τ_a result in pronounced changes in both the shape and magnitude of the model solutions, variations in τ_e result in much smaller overall changes in the solution. In particular, reducing the value of τ_e leads to very small changes in the solution, with the rise of the velocity to its peak velocity being nearly indistinguishable between different values of τ_e in the limit as $\tau_e \to 0$. We argue that, therefore, non-identifiability of the parameter τ_e in this case can be easily understood given the fact that these variations in model solutions are much smaller than the variance present in the experimental data. On a practical leve, these issues arise because data acquisition in the experiment is only in 10 minute intervals, the data are not temporally resolved enough to tell different values of τ_e apart when there is only one time series. The inference results in Section 5.2 demonstrate that this issue does not occur in the presence of more data, but in the case of a single time series the temporal resolution becomes problematic.

In summary, we have shown that relaxing the assumption that the effective signal in the monolayer is always positive can explain the finding that the velocity in the top edge of the monolayer turns negative for several hours after the electric field is turned off. By carefully investigating model predictions with varying parameter values, we found that the undershoot is well-described by the maximum likelihood value of the parameters, which was difficult to identify in practice owing to the sloppiness in the model. Interestingly, such an undershoot is not observed any of the other tissue regions studied in this work, which begs the question for future work: why does signal transduction during electrotaxis appear to be intrinsically different in different tissue regions?

Chapter 6

Distinguishing the impact of gene knockdowns on collective invasion using deep attention networks

The work in this chapter uses experimental data collected by Jen Kulesa-Kasemeier (Stowers Institute for Medical Research, Kansas City). My contributions to the work were processing and curating the experimental data, all computational implementations, and the writing of the results in this chapter.

6.1 Introduction

Neural crest (NC) cells are essential to vertebrate embryogenesis, where they travel long distances after exiting the dorsal neural tube and contribute to nearly every peripheral organ [140]. Although NC cells emerge all along the anteroposterior axis and are sculpted into discrete streams that invade the immature extracellular matrix (ECM) and loose mesoderm, the molecular mechanisms that underlie invasion and collective cell migration remain unclear. Recent advances in single cell transcriptome analysis have revealed that there is significant spatial molecular heterogeneity within a NC cell migratory stream [141–143], and that in the context of the chick head, leader cells at the front of NC invasion have a distinct and consistent transcriptional signature that is enriched for approximately 900 genes [141]. Further analyses, including RNA sequencing (RNA seq) of isolated cells, have added to the picture of invading cells having an altered genetic expression from the rest of the cell population (see Figure 6.1). These advances invite the question of understanding the role of each of these altered genes in facilitating and establishing

successful NC invasion into surrounding tissues during embryogenesis and the role of genetic alterations in cell invasion phenomena in other cell types, more generally.

An understanding of the genetic signature underlying NC invasion can be leveraged clinically. Mistakes in NC cell migration in the head are the cause of neurocristopathies, severe birth defects that include a shortened lower jaw, as well as cleft lip and palate, while a failure of NC cells to reach the end of the embryonic gut is implicated in Hirschsprung's disease [144, 145]. Additionally, owing to the ancestral relationship of the NC to melanoma, the NC is also a model for several invasive human mesenchymal cancers [146, 147]. For this reason, the identification and functional analysis of a critical set of cell invasion genes has the potential to give rise to new treatments for cancer metastases. Together, 900 genes that have an enriched transcriptional signature in the leader cells that trailblaze NC invasion provide a starting point for investigating the key genetic players involved in cell invasion phenomena.

Despite the previous identification of a list of candidate genes, the mechanistic role played by each of the candidate genes considered in the 900 gene list by Morisson *et al.* [141] is not well understood. For this reason, the bottom-up construction of a mathematical model that encodes fixed mechanisms will likely obscure a good understanding of the variety of different mechanisms mediated by the genes of interest. This difficulty relates to one of the challenges set out in the introductory chapter, which is the use of mathematical modelling to gain new mechanistic insights. For this reason, this chapter will focus on using computational techniques to discover the ways in which cell invasion is altered by expression of the different genes in this screen. This chapter is structured as follows. In Section 6.2, we describe the experimental data considered. In Section 6.3 we carefully quantify the migratory behaviour of individual cells during invasion to characterise the different modes of migration upon genetic perturbation. In Section 6.4 we describe a deep learning approach based on deep attention networks to quantify cell-cell interactions during invasion to find that there exists a large heterogeneity in the cell-cell interaction patterns. We provide a discussion for future work in both biological experiments and mathematical modelling in Section 6.5.

6.2 Data and methods

6.2.1 A set of 45 candidate genes for cell invasion

Systematic evaluation of invasion using a list of over 900 candidate genes is prohibitive, owing to the large human and experimental cost. For this reason, Morison *et al.* at the Kulesa laboratory at the Stowers Institute for Medical Research distilled the large list of over 900 genes from Morrison *et al.* [141] to a 45-gene panel that is common with

other highly invasive cell phenomena [142]. This was done by cross-referencing a list of previously published gene signatures with genes common to the chick leader cranial NC cell invasion signature (see Figure 6.1). In summary, a list of 34 published gene signatures from other invasive cell types, including those involved in cancer, wound healing, and the immune response, was selected and compared to the original candidate list from the study of Morrison *et al.* [141]. Of all the genes considered in the 34 published gene signatures, this systematic comparison produced a list of 45 candidate genes that were shared between chick NC cells and at least two other invasion signatures. Due to this overlap in origins, studying the function of these 45 genes might reveal a conserved function in invasion and cell migration, which would be apparent in other, non-neural crest-derived cell types. Although this list of 45 genes does not include all the genes necessary for cell migration and invasion, the reduced size of the list of candidate genes makes it amenable to a screening assay, where the role of each gene can be studied in isolation. By doing this, the 45 gene panel can be reduced further to a much smaller list of critical genes that have the most noticeable impact on cell migration and invasion *in vitro*, so that their role can be studied *in vivo*.

6.2.2 A rubber stopper high-content screening assay

To address the question of how best to isolate the role of different genes in establishing cell invasion, a high-content *in vitro* screening assay offers a rapid means of analysing large transciptional data. To test the role of each gene in establishing NC invasion, one might devise a high-content *in vitro* experiment to relate altered transcription patterns to observed differences in NC migration in cells dissected from a given model organism, with the transcriptional targets initially prioritised for cell migration and invasion. In the research of the Kulesa laboratory at the Stowers Institute for Medical Research, the model organism is chick with a specific focus on cranial neural crest cells. However, the dissection and isolation of large numbers of embryonic NC cells in a model organism with the purpose to seed multiple multiwell high-content assays is prohibitively labour-intensive. To overcome this obstacle, the Kulesa laboratory instead used the neural crest-derived human metastatic melanoma c8161 cell line, which is well known to display behaviours similar to embryonic NC cells [148–150]. Such behaviours include aggressive invasion of free space and ECM as well as collective cell migration. Knockdown of individual genes using siRNA transfection in cell culture offers a straightforward means to molecularly perturb gene function and, when combined with a high-content assay and time-lapse confocal microscopy, permits quantification and comparison of changes in cell migratory behaviours and invasion. For this reason, a high-content screening assay is a tractable experimental system to perform a quantitative analysis of gene function in NC invasion.

Experimentally, the 45 gene invasion signature was investigated using a high-content rubber stopper-based screening assay (see Figure 6.1). The cell line was transfected with siRNAs to knock down single genes in duplicate from the 45-gene panel prior to seeding cells in rubber stopper-plugged plates (along with water and negative siRNAs for controls). After seeding and adherence of cells in each well, plugs were removed from each well and cell trajectories were collected from 24 hour time-lapse confocal imaging sessions, where images were acquired at 10 minute intervals. This experimental setup yields data at a rich spatial and temporal resolution: the data are detailed enough that single cells can be followed during the course of the experiment (see Figure 6.1). We note that this experiment is different from typical scratch assays. The benefit of removing a rubber stopper from a cell population versus making a scratch in an already confluent monolayer is that removal of the rubber stopper does not disrupt cell integrity and provides a uniform circular wavefront.

Finally, to compare the role the 45 gene panel has in invasion phenomena specifically, compared to other effects that might be visible in cell types that do not natively display collective migration and invasion, the screening assay compared human c8161 melanoma versus human HT1080 fibrosarcoma cell lines. Both cell lines originate from highly aggressive cancers, but only c8161 melanoma cells are both ancestrally related to embryonic NC cells and have been observed to exhibit collective cell migratory behaviours. Insights into how the 45 gene knockdowns alter the collective movement in HT1080 cells can therefore act as a control experiment to isolate the altered invasiveness of the candidate genes from other generic cellular processes that might be disrupted during siRNA transfection.

6.3 Statistical assessment of cell invasion dynamics

Traditionally, cell migration and invasion studies involve a quantification of the fold changes in space occupied by the cells during the course of the experiment, as well as fold changes in the cell counts [23, 43, 63]. For the computation of free space areas, we used Fiji [98] to threshold and binarize the nuclear channel to obtain the individual cell locations, and then performed convolution with a Gaussian kernel with standard deviation of two cell widths to smooth the nuclear channel into a density field. This we then thresholded and binarized to obtain free space masks to measure the area of the free space. We did so by making a two-dimensional histogram of the cell densities using bins of size 5×5 cell lengths, and classifying spatial locations as being the wound if these contained one cell or fewer. We used the thresholded and binarized nuclear channel to count cells using Fiji and obtain cell count fold changes between the start and the end of the experiment. We then computed the area and cell count fold changes for each of the tissues recorded and grouped them together by knockdown, giving sets of 3-9 biological samples for each

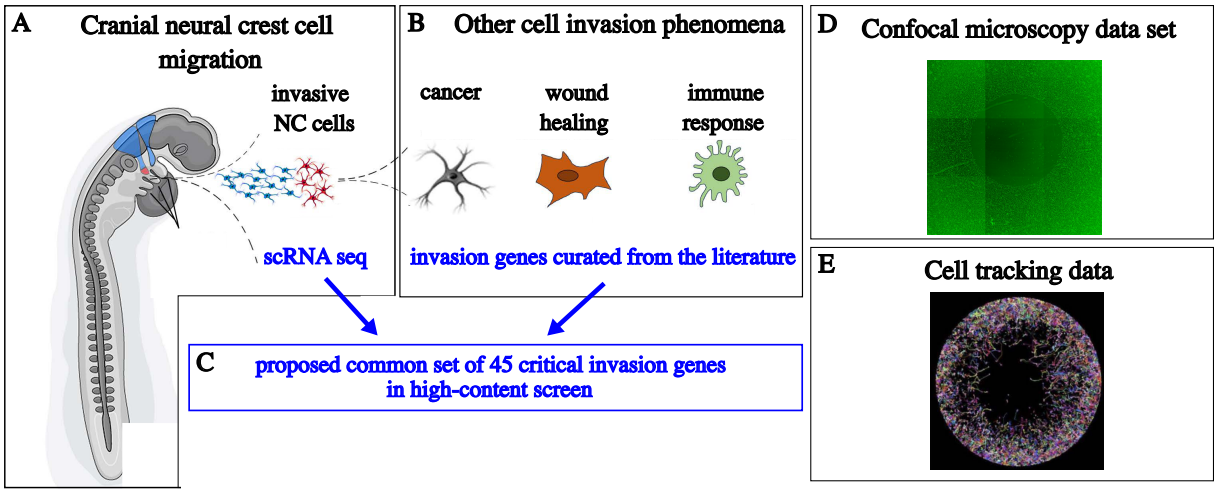


Figure 6.1: Schematic showing workflow of the experimental set up and the resulting experimental data. A: schematic of cranial neural crest migration, showing the BA1 branchial arch with its characteristic leader and follower cells. B: scRNA-seq analysis by Morrison *et al.* is combined with screening panels in the literature from other cell invasion phenoma to propose a common set of 45 critical cell invasion genes. C: each gene from the proposed set of 45 candidates is knocked down in a high-content siRNA screen using a rubber stopper experiment. D: confocal imaging is used to quantify cell locations, yielding a quantification of free space (top) and can be used to compute individual cell trajectories (bottom).

individual knockdown. We computed z-scores for each of the free space area and cell count fold changes and perform a standard Welch's t-test to compute the corresponding p-values for free space area and cell count fold change.

6.3.1 Invasion is affected primarily by cell motility rather than proliferation

As an initial screening examination to assess which siRNA knockdowns were associated with dynamic changes in human cancer cell behaviors during closure of the free space, we performed a statistical analysis of free space area and cell count fold changes (see Table 6.1). For each well, we computed z-statistics for the free space area and cell count fold change. We present gene knockdowns that significantly affected free space closure and find 14 out of 45 genes included in the screen resulted in significantly affected free space closure in the c8161 melanoma cells (see Table 6.1). Interestingly, only four out of 14 genes that significantly affected c8161 free space invasion also exhibited significantly reduced free space invasion in HT1080 fibrosarcoma cells (BMP4, ITGB1, KCNE3, and RASGRP1, see Table 6.1). For results in the HT1080 cell line and all other c8161 knockdowns that do not produce significantly altered area fold changes, we refer the reader to Appendix B.1.

To better understand the interplay between cell proliferation and cell motility during free

Gene knockdown	Area fold change	Cell fold change
BMP4	5.76	1.03
FN1	1.09	1.24
FOS	1.14	1.22
IFI30	4.34	1.36
IL13RA1	3.59	1.35
ITGB1	4.55	1.35
KCNE3	4.13	1.36
KYNU	3.91	1.33
NEXN	3.80	1.37
Neg	2.45	1.39
PRKCQ	4.09	1.35
RASGRP1	4.00	1.37
SH3TC1	3.71	1.34
UPP1	3.44	1.33
VIM	3.82	1.36

Table 6.1: Area and cell fold changes for the gene knockdowns that lead to a reduced free space closure in the c8161 melanoma cells. Green shading in the area fold column indicates improved free space closure, red shading in the cell fold column indicates lower cell counts. Knockdowns shaded in orange have statistically significant differences in free space closure in both c8161 and in the control HT1080 cell line with respect to Neg siRNA control (cyan).

space closure, we compared how these two metrics were jointly affected upon knockdown. In c8161 melanoma cells, only BMP4 knockdown inhibited the closure of free space in conjunction with significantly reduced cell proliferation (Table 6.1; first row). Other knockdowns of the 14 out of 45 genes that exhibited a reduced rate of free space closure (for example, ITGB1, KCNE3, and RASGRP1) did not show significantly altered cell proliferation in the c8161 melanoma cells. FN1 and FOS knockdowns in the c8161 melanoma cells showed significantly improved free space closure rates and both gene knockdowns also exhibited significantly reduced cell proliferation (see Table 6.1; rows 2-3). Likewise, no gene knockdowns in the c8161 melanoma cell line data set were associated with a reduced rate of free space closure and enhanced cell proliferation, or both an enhanced rate of free space closure and increased cell proliferation. In contrast, in the knockdowns where the invasion of HT1080 cells was affected, there was both reduced cell proliferation and changes in cell motility, with a reduction of cell proliferation being the more pronounced effect (see Appendix B.1) Together, these data suggest that the set of gene knockdowns in c8161 melanoma cells primarily affected free space closure through changes in cell motility, rather than through cell proliferation. Gene knockdowns had an effect on both cell motility and cell proliferation in the HT1080 fibrosarcoma cells and were more strongly associated with reduced cell proliferation.

6.3.2 Leader-follower dynamics are a hallmark of c8161 melanoma collective migration

While the metrics regarding cell counts and invasion of free space are useful in understanding the macroscale dynamics of wound closure, they cannot provide information on the migratory behaviour of individual cells. Thankfully, the spatial and temporal resolution of the confocal microscopy data makes it possible to track individual cells during collective migration and quantify interpretable statistics about their movement. We use TrackMate in Fiji [98] to obtain individual cell tracks from the binarized and thresholded images. We use the inbuilt Laplacian of Gaussian (LoG) detector with no track splitting and merging, as well as a gap closure of five frames and a maximum linking distance of 100μm. We set the cell diameter for c8161 cells to 20μm and that for ht1080 cells to 30 μm. From the TrackMate outputs, we filter for cell tracks that contain at least five consecutive frames. From each track, i, which consists of locations $(\mathbf{x}_i(t))_{t=t_1^i}^{t_n^i}$, we compute the average cell displacement, d, as

$$d = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{t_n^i - t_1^i}\sum_{k=2}^{n}\|\mathbf{x}_i(t_k) - \mathbf{x}_i(t_{k-1})\|_2, \tag{6.1}$$

and compute cell directionality with the formula

$$\theta = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{t_n^i - t_1^i}\sum_{k=2}^{n}\cos(\theta_i(t_k)), \tag{6.2}$$

where $\cos(\theta_i(t_k))$ is the angle between the unit vector perpendicular to the free space edge and the velocity at at time t_k. Here, velocities, $(\mathbf{v}_i(t))_{t=t_2^i}^{t_n^i}$ are defined as

$$\mathbf{v}_i(t) = \frac{\mathbf{x}_i(t_t^i) - \mathbf{x}_i(t_{t-1}^i)}{t_t^i - t_{t-1}^i}. \tag{6.3}$$

Cell displacements act as a measurement for how far cells move on average between two frames, with smaller values of cell displacement related to impaired cell motility. Cell directionality measures how directionally the cells move into free space, giving an indication of the strength of the free space cue in establishing the migration of leader cells. To understand the migratory differences between the cells at the edge of free space and in the bulk of the tissue, we classified cells as edge cells if their distance to the masked free space edge was less than five cell radii at their first detected time point.

The preceding statistical analysis of open free space areas and cell count fold changes suggested that invasion is driven by cell motility in the c8161 melanoma cells, but did

not precisely identify which changes in motility at the cell-level contribute to reduced or enhanced free space closure. To address this, we performed cell tracking to investigate the motility properties of individual cells more thoroughly. NC cells are well-known for their leader-follower dynamics, whereby leader cells must invade through immature ECM, loosely connected mesoderm and other cell types to pave the way for follower cells to move collectively behind the wavefront. To better understand whether the gene knockdowns in the screen perturb this pattern, we turned to the metrics obtained by applying Equations (6.1) and (6.2) to the cell trajectories in each of the knockdown experiments. By making a distinction between cells at the leading edge (edge) versus cells in the follower (bulk) subpopulation, we computed the displacement difference between edge and bulk cells for each of the different wells, defined as the difference between the mean edge and the mean bulk cell displacement.

The displacement difference acts as a measure for how motile on average the cells at the edge of the collective are relative to those in the bulk subpopulation. A large displacement difference would indicate a highly active edge subpopulation that rapidly invades towards the centre of the free space, with a less actively motile bulk population. Similarly, we computed the average directionality in the direction of the free space for each cell and determined the directionality difference as the difference between the average directionality for cells at the edge and that for cells in the bulk (see Figure 6.2). We find that c8161 melanoma cell edge subpopulations display significant differences in motility and directionality compared to the HT1080 fibrosarcoma cells across the different gene knockdowns (see Figure 6.2). This finding suggests that c8161 melanoma but not non-NC-derived HT1080 cells exhibit an edge subpopulation that rapidly migrates in a directed manner to fill up free space, regardless of gene knockdown.

We next compared the displacement difference directly to free space closure (see Figure 6.2). We find that for the c8161 melanoma cells, an enhanced motile edge subpopulation leads to enhanced free space closure (Spearman correlation coefficient of -0.452, p-value=0.00180); a trend that we did not observe in HT1080 fibrosarcoma cells (see Figure 6.2). Together, this suggests that for c8161 melanoma cells, the active leader cells within the edge subpopulation are most successful in closing the free space, in contrast to when all cells in the collective migrate according to the same rules. Lastly, the same analysis for the directionality difference showed no direct relationship with closure of the free space for the HT1080 fibrosarcoma cells.

6.3.3 Loss of leader cell behaviours upon genetic perturbation

Since our cell trajectory analysis suggests that highly motile leader cells are associated with faster free space closure, we examined whether the set of gene knockdowns that resulted in reduced free space closure also exhibited a reduction in relative edge displacements.

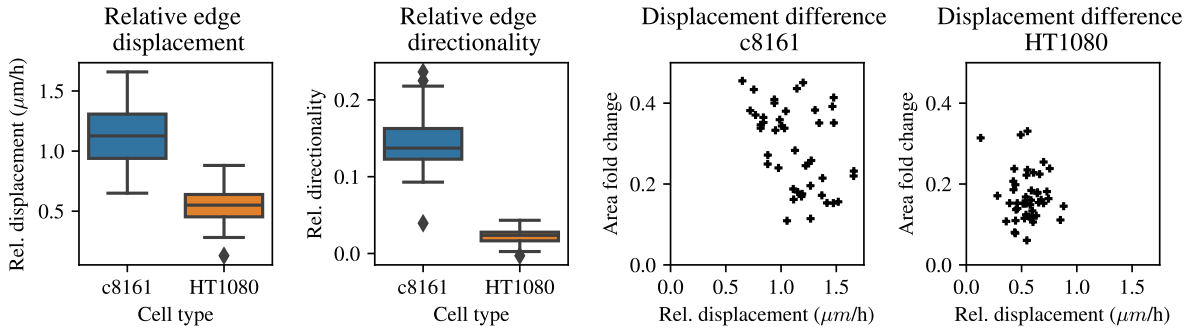


Figure 6.2: Differences in displacement lead to different invasion outcomes in c8161 cells. From left to right: relative edge displacement shows a pronounced leader cell behaviour in the c8161 cell line relative to the HT1080 cell line; the relative edge directionality is greater in c8161 cells than in HT1080 cells; displacement difference in c8161 cells plotted against the fold change in free space area; directionality difference in c8161 plotted against the fold change in free space area.

Indeed, leader cell motility is reduced by as much as 30% in the BMP4, ITGB1, KCNE3 and RASGRP1 knockdowns of c8161 cells, with a mean cell motility of 5.1, 4.4, 4.5, and 5.2 μm/hr, respectively, compared to 6.4 μm/hr in Neg siRNA controls (see Figure 6.3). Furthermore, by characterising the difference between displacements for leader versus bulk cells, we find that the knockdowns of BMP4, ITGB1, and RASGRP1 also exhibited a markedly reduced difference between leader and bulk cell motility. This is evident by the overlap in the interquartile range, contrary to the trend in control and KCNE3. Interestingly, only BMP4 exhibited a reduction of leader cell directionality (0.094 versus 0.23 for control), establishing that loss of BMP4 gives rise to less directed, more wandering c8161 melanoma cell behaviours at the leading edge.

Performing this analysis on the remaining 10 of the 14 gene knockdowns in c8161 melanoma cells that result in reduced free space closure (see Table 6.1; not significant in HT1080 fibrosarcoma cells), we find that all 10 genes except PRNP have significantly reduced edge cell displacement (see Appendix B.2). No gene knockdown shows altered edge directionality compared to Neg siRNA controls. These findings suggest that the previously identified group of critical gene knockdowns exhibit reduced free space closure due to inhibited leader cell motility. Furthermore, BMP4, ITGB1, and RASGRP1 also show a loss of leader cell-type behaviour by displaying a larger difference between leader and follower cell motility. Finally, we note that both the FN1 and FOS knockdowns (which have significantly higher rates of free space closure than the control) show the same trend, exhibiting significantly more motile edge cells (data not shown).

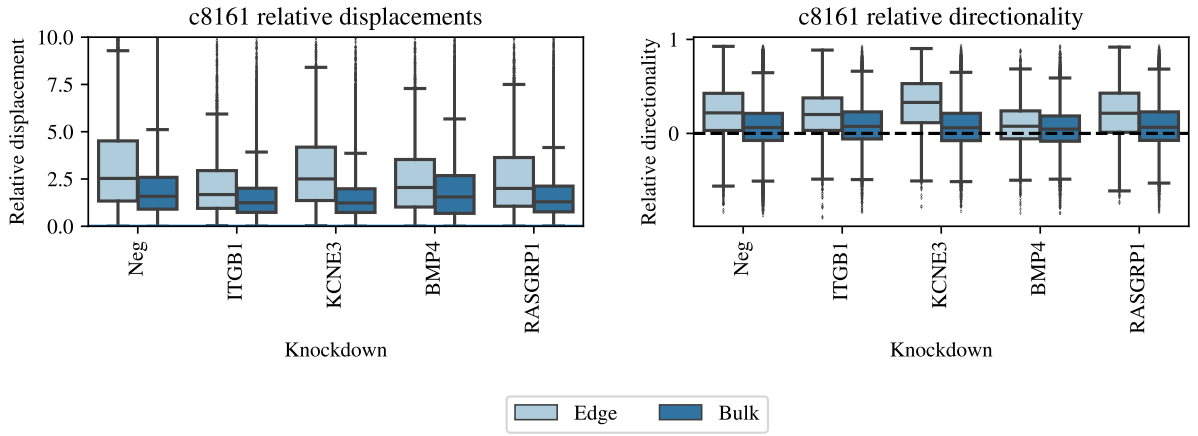


Figure 6.3: Loss of leader-cell type behaviour upon genetic perturbation. Left: relative displacements in bulk and edge populations in c8161 cells. Right: relative directionality in bulk and edge populations in c8161 cells.

6.4 Quantifying cell-cell interactions using deep attention networks

While individual cell behaviours at the leading edge have a significant impact on the rate of free space invasion, cells do not migrate in isolation as cell invasion is an inherently cooperative process. Therefore, we asked whether any of the 45 genes in the high-content screen play a role in cell-cell communication during the closure process. Deep attention networks are a mathematical framework that enable quantitative measurements of the spatial distribution of the cell-cell interactions with most influence over velocity alignment in a collective group of cells. Deep attention networks were previously introduced by Heras *et al.* [42] to quantify how zebrafish change their velocity and orientation based on the motion of other fish around them, but were later extended by LaChance *et al.* [41] to establish that different cell types (MDCK epithelial cells, HUVEC endothelial cells, and MDA-MB-231 metastatic breast cancer cells) have distinct characteristic rules for the extent to which cells pay attention to their neighbours during collective movement. This framework lends itself well to our data set of distinct genetic knockdowns, and we trained deep attention networks on each of the different siRNA gene knockdown data sets to quantify cell-cell attention patterns. For our deep learning implementation, we used the same hyperparameters as LaChance *et al.* [41] (the most important of which being a lower bound for prediction accuracy of at least 65%). Training was performed on the Oxford Advanced Research Computing cluster using five 48-core Cascade Lake (Intel Xeon Platinum 8268 CPU @ 2.90GHz) nodes.

Mathematically, deep attention networks take a collection of individual cell trajectories, and seek to forecast the logit probability, z, that a given cell – which we call the *focal* cell – will turn right, given the attributes of its neighbours. In the deep attention framework

this is done by learning a pairwise function, Π, that quantifies the influence that the focal cell's neighbours have, as a function of their velocity, distance, and acceleration. At the same time, the deep attention framework learns an *attention function* that places weights according to the apparent attention that cells pay their neighbours. Put together, the logit probability that focal cell j will turn right, z_j, is expressed as a function of the interaction between cell j and its neighbours, and weighted according to their relative velocities,

$$z_j = \sum_{i=1}^{n} \Pi(\alpha_i, \alpha_j) \frac{W(x_i, v_i, x_j)}{\sum_{\ell=1}^{n} W(x_i, v_i, x_\ell)}, \tag{6.4}$$

where $\alpha_\bullet$ is a vector containing the velocity, distance to the focal cell and acceleration of a given cell. Note that the attention function W, when normalised, acts as like a kernel that assigns different weights to the neighbour cells, depending on their location relative to the velocity of the focal cell. The mathematical description of cells aligning their velocities according to a pairwise interaction is reminiscent of Vicsek-type alignment models [151]. In this case, the attention function W can allow cells to pay attention only to a subset of their neighbours, rather than all neighbours within a certain distance. This makes the alignment *anisotropic*. We will discuss the implications of this choice for the mathematical modelling of interacting cells in Section 6.5.

6.4.1 Distinct patterns in collective cell migration after alteration after gene knockdowns

We first established typical characteristic patterns of attention for c8161 melanoma and HT1080 fibrosarcoma cells in the absence of gene knockdown (see Figure 6.4). In this view, c8161 melanoma cells are characterised by a highly localised attention pattern at the front of the cell. In stark contrast, we find a very narrow front-and-back pattern of attention for HT1080 fibrosarcoma cells. This pattern is consistent with the typical sliding of cells in between each other in a largely non-coordinated fashion. Having characterised the baseline cell-cell attention patterns, we next turned to the question of whether the knockdown of any of the 45 genes would reveal attention patterns distinct from the Neg siRNA knockdown scenario.

To better compare the patterns of cell neighbour relationships between the different gene knockdowns, we extracted quantitative metrics that describe the patterns of attention. We quantified both the extent to which cells pay attention to neighbours in front and behind, as well as the width of the attention field (see Figure 6.4 B and C). We noticed that c8161 melanoma cells have a much wider attention pattern than HT1080 fibrosarcoma cells. Furthermore, c8161 melanoma cells display a wide range of attention patterns depending

on gene knockdown (see Figure 6.4 D and E), while there is no qualitative change in the attention maps for any of the HT1080 knockdowns (data not shown). In particular, both the width of the cell-cell attention field and the front-to-back cell neighbour interactions were affected (see Figure 6.4 D). For example, we observed a wider attention range in the DAB2 knockdown that stretched from the front to along the sides of the cell, suggesting that the cell-cell interactions were broadened to side-by-side neighbours (Figure 6.4 D). On the contrary, the CA2 knockdown produced a pattern where the neighbours at the front and back of the cell were largely ignored. FN1 knockdown produced a pattern in the c8161 melanoma cells that resembled the HT1080 cells with a front-to-back pattern (compare Figure 6.4 C with D). In the CTNNB1 knockdown, cells appeared to pay attention to neighbours along the sides and behind rather than search for open free space. Strikingly, in the HT1080 fibrosarcoma cells we find that there is no qualitative variation in attention patterns (data not shown), consistent with the observation that migration is highly collective in the c8161 melanoma cell populations, but not in the HT1080 fibrosarcoma cell populations.

We next asked whether altered patterns of attention are associated with the observed free space closure changes. We compared the attention patterns for the four gene knockdowns leading to inhibited free space closure (BMP4, ITGB1, KCN3, RASGRP1) with the attention patterns for the knockdowns that resulted in improved free space closure (Figure 6.4 E). While loss of ITGB1 and KCNE3 retained the wild type pattern of attention, both BMP4 and RASGRP1 knockdowns displayed altered attention patterns (see Figure 6.4 E). Importantly, these respective attention patterns were qualitatively different; loss of BMP4 resulted in attention towards the front and back and loss of RASGRP1 resulted in attention only towards the front. Likewise, loss of FN1 and FOS promoted free space closure but showed no common qualitative features in the attention pattern. Lastly, inspection of the attention patterns of the gene knockdowns not associated with altered free space closure equally showed considerable variation in attention structure (*i.e.,* DAB2, CA2, and CTNNB1 knockdowns).

We analysed the relation between features of the attention maps and the rates of free space invasion by quantifying three features of the attention maps that we commented on in the previous paragraphs. First, we sought to quantify how much attention cells pay towards the neighbours to their front versus that to their back. We quantified this by computing the share of the total attention weight in the attention maps (for some examples see the maps in Figure 6.4) for each of the knockdowns. To assess how narrow or how wide the attention patterns looked in the front and back halves of the attention networks, we sought to quantify within which angle relative to the focal cell's velocity vector more than 50% of the attention weight was located. We summarise this schematically in Figure 6.5.

By comparing the three quantitative metrics of front attention, front angle, and back angle between the knockdowns that led to more closure, less closure, or no significant

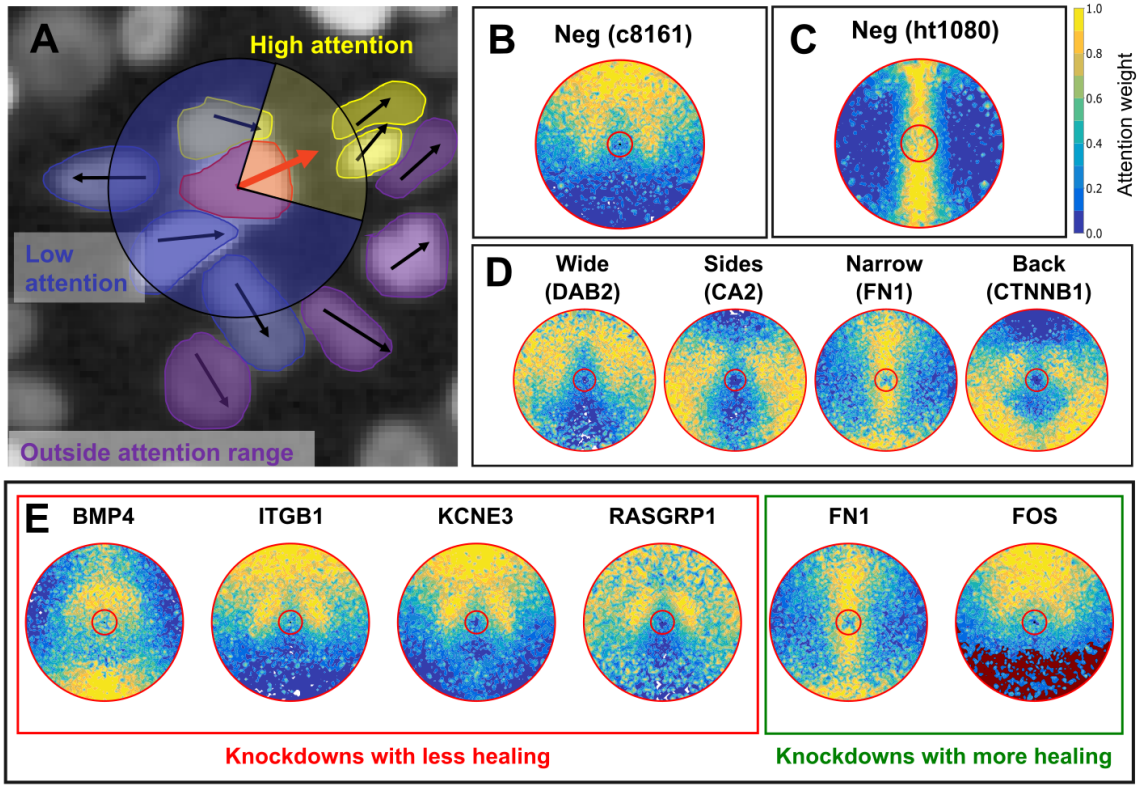


Figure 6.4: Deep learning reveals altered patterns of attention between cells. A: schematic showing focal cell (red) with a cartoon attention cone. Cells in yellow are cells in the high attention region, cells in blue in the low attention region and cells in green are outside the attention range. B: attention patterns for Neg knockdown in c8161 cells. C: attention patterns for Neg knockdown in ht1080 cells. D: effects of genetic perturbations on attention patterns, notably the width of the attention cone as well as the location to the front or back of the cell. E: attention networks for all perturbations with significantly affected free space closure rates, separated into knockdowns with less, and more, free space invasion, respectively.

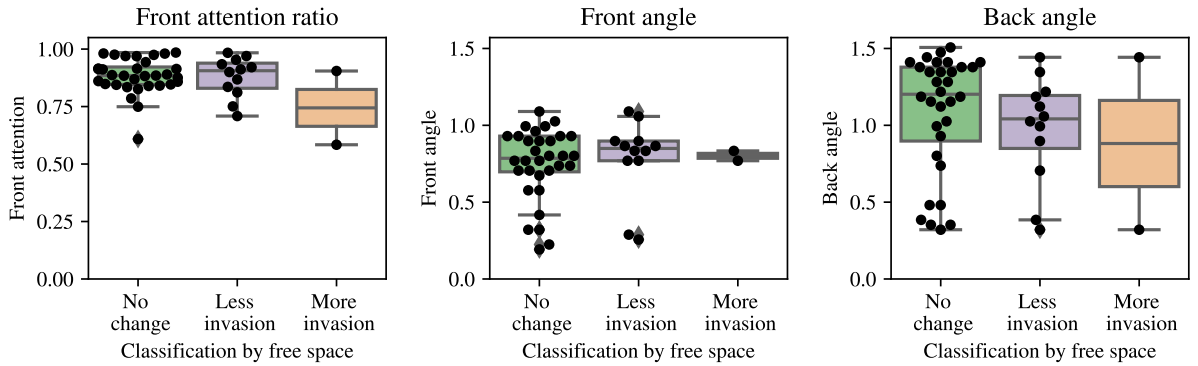


Figure 6.5: Clustering of knockdowns by their free space invasiveness shows no relationship with quantitative metrics of their respective attention networks. From left to right: front attention ratio, front angle, and back angle.

change, we observed that there was no clear difference between these features of the attention maps and the rates of free space closure. Furthermore, we observed that there is no relationship between mean cell displacement, mean cell directionality, or relative motility differences between leader cells and follower cells in the different knockdowns. This finding points to patterns of cell-cell attention being influenced separately by the different genes that were knocked down in this study.

6.4.2 Alterations in attention maps are not related to the biological function of individual siRNA knockdowns

The attention maps generated across the screen (see Figure 6.4) revealed several gene knockdowns that altered both the cell migratory properties and the nominal cell-cell interactions during *in vitro* migration of c8161 melanoma cells, consistent with the collective cell migratory characteristics of the c8161 melanoma cells and identifying genes for deeper analysis. This included changes in both the width of the cell-cell attention field and the front-to-back cell neighbour interactions. From this analysis, we found major alterations to the cell neighbour patterns in genes not necessarily correlated with significant reduction in free space closure, *e.g.* in DAB2, CA2, FN1, and CTNNB1 knockdowns (see Figure 6.4). For example, after loss of DAB2, cells appeared to expand communication with neighbours along the sides of the cell rather than simply in front of the cell in search of open free space (see Figure 6.4). After loss of FN1, the alteration of cell-cell interactions was more dramatic. Loss of FN1 led to a front-to-back cell-cell interaction pattern, suggesting a less coordinated sliding of cells between each other, similar to the control pattern of HT1080 cells (see Figure 6.4). DAB2, which encodes a mitogen-responsive phosphoprotein, has been shown to regulate the migration and invasion of prostate cancer cells [152]. Its knockdown by siRNA inhibits these characteristics in PC3 cells in vitro, but the mechanism of its action is unclear. In our data set, when DAB2 was knocked down by siRNA, c8161 melanoma cells altered their cell neighbour relationships to include a much wider attention field (see Figure 6.4) but did not show reduced migration. So, either DAB2 is not critical for melanoma migration, or its knockdown causes cells to seek alternative guidance signals.

The problem of relating the observed attention pattern in DAB2 knockdown to its known biological function begs the question of whether observed differences in DAB2 are similar to those in genes that serve a related biological function or, more generally, whether genes with a similar function will display similar patterns of attention upon knockdown. Motivated by this question, we used quantification of the most important features of the attention maps in Section 6.4.1, to explore if there is a relation between the biological function of each knockdown and the altered attention pattern. To do this, we received

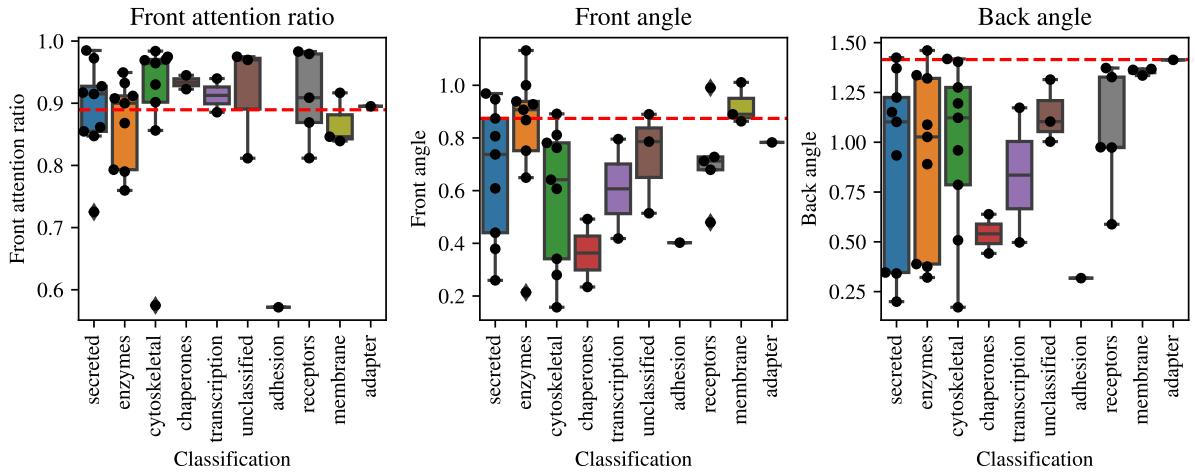


Figure 6.6: Genes clustered by knockdown show no systematic relationship in metrics from attention maps compared to that of Neg.

a classification of knockdowns according to the function of the transcribed protein[1]: enzymes, secreted proteins, cytoskeletal, adhesion, chaperones, membranes, receptors, adapters, transcription or unclassified (if there was not enough information to classify the knockdown). For each of these categories, we computed the three metrics of front attention, front angle, and back angle. We display the different groups in Figure 6.6, compared to Neg siRNA controls. Figure 6.6 shows large variation in the observed metrics, both within and between the different classifications. While some knockdown groups seemingly display shared trends compared to the Neg siRNA control, individual attention maps in the different groups display no qualitative agreement, with some knockdowns leading to attention patterns identical to those of Neg siRNA control (see, for example, ITBG1 and KCNE3 knockdowns in Figure 6.4). In summary, there is a large heterogeneity within each group, and a simple classification of the knockdown gene by biological function does not explain the characteristics of the attention maps. This invites further experimentation into the causes of the different attention patterns by probing in more detail the (mechanical) interactions between the cells.

6.5 Discussion and outlook

Previous single cell profiling produced a 45 gene panel implicated in a wide range of cell invasion phenomena [142]. While a panel of this size is impractical for single gene *in vivo* studies, it can be appropriately addressed in an siRNA high-content screening assay. By comparing changes in cell migration into the open free space and cell number after each of 45 siRNA gene knockdowns in neural crest-derived human c8161 metastatic melanoma

[1]Personal communication with Kulesa lab.

cells versus non-neural crest-derived HT1080 fibrosarcoma cells, we identified a subset of 14 out of 45 critical genes for cell invasion in the c8161 cell line. Confocal time-lapse imaging provided a very large set of cell trajectory data sufficient for statistical analysis and deep learning attention network analysis of cell neighbour relationships of both leading (edge) and follower (bulk) subpopulations. By using statistical assessment of traditional migratory metrics, we quickly converged on a set of four knockdowns that led to abnormal leader-follower dynamics in the c8161 cell line. Deep learning attention network analysis revealed several gene knockdowns that altered both the cell migratory properties as well as the nominal cell-cell interactions during *in vitro* migration of c8161 melanoma cells. While this quantification yielded insights into the altered modes of migration in c8161 melanoma cells after gene knockdown, relating these modes of migration to the biological role of the genes in the screen remains challenging. Therefore, while this study demonstrated the rapid and systematic analysis of a 45-gene panel to converge to a critical subset of genes using a combined experimental, statistical, and deep learning approach, it also opens future avenues of research about the origin, and the role, of cell attention patterns in collective migration. Answering these questions may have direct applications to emerging cell replacement therapies to repair neurocristopathies by providing a molecular blueprint to control transplanted cell behaviors and suggest therapeutic applications to inhibit NC cell-derived cancer metastasis.

6.5.1 Experimental follow-up to determine the biological origin of anisotropic attention between invading cells

In this chapter we discovered *that* cells interact with each other in a highly anisotropic way, yet it is not currently known *why* this might be the case. A natural approach to probe this question is to take a closer look at the mechanical and morphological implications of gene knockdowns on the main biological drivers of cell-cell interactions, which are the cytoskeletal machinery, the cell membrane, and cell-cell adhesion molecules. By staining each of these cellular components differently, one can perform a detailed quantification of the dynamics of cellular protrusions, cell-cell contacts and cell morphologies using multiplex imaging technologies. Since the confocal imaging method used in this study allows for image acquisition at the subcellular level, such an analysis will be able to provide precise measurements of the locations and movements of specific parts of the migrating cells. This might provide further insights into which processes at the subcellular level contribute to altered patterns of attention at the supracellular level.

6.5.2 Incorporating anisotropic attention in mathematical models of interacting particles

Mathematically, the deep attention network analysis in Section 6.4 revealed that statistical correlations between the velocities of different cells depend on interactions that are anisotropic. This runs counter to the assumption in many mathematical models of cell migration that represent cells as point particles and allows cells to interact with each other isotropically, a canonical example being the Vicsek model [151]. Recently, Gu *et al.* [153] proposed a model, based on data of human dermal fibroblasts, that incorporates cells aligning with other cells that are travelling in the same direction. However, they did not specify a functional form describing how this alignment can be represented mechanistically in an interacting particle model. Here, we propose that a minimal, general interaction model framework can be developed from our deep attention networks, in which cell-cell interactions are comprised of three mechanisms: attraction-repulsion, alignment, and drag. By making the canonical assumption that that motion is strongly over-damped, we can pose a simple force balance in Equation (6.5) for each cell, where the drag force, γv_i, where γ is a friction coefficient, is balanced by a force that results from attraction and repulsion, f_{ar}, and an alignment force, f_{align}. By making the assumption that the attraction force acts in the direction of the vector between the cell centres, while the alignment force acts in the direction of their relative velocities, we obtain

$$\gamma \mathbf{v}_i = \frac{1}{N}\sum_{j=1}^{N} f_{\text{ar}}(\|\mathbf{x}_i - \mathbf{x}_j\|, \theta_{ij})(\mathbf{x}_i - \mathbf{x}_j) + \sum_{j=1}^{N} f_{\text{align}}(\|\mathbf{x}_i - \mathbf{x}_j\|, \theta_{ij})(\mathbf{v}_i - \mathbf{v}_j) + \sigma\xi_i, \quad (6.5)$$

where θ_{ij} is the angle between the velocities v_i and v_j, ξ_i is a standard Gaussian random variable representing noise, and σ is the standard deviation of the noise. To abstract some of the features that we observed in our study of the attention maps, most notably that attention is typically concentrated within a cone in the front and in the back of the cell, we let the attention kernel be defined by a *critical angle* as follows. Given a critical angle, $\hat{\theta}$, we can define $\alpha = \cos(\hat{\theta})$. Letting

$$\alpha_{ij} = \frac{v_i \cdot (\mathbf{x}_i - \mathbf{x}_j)}{\|v_i\|\|\mathbf{x}_i - \mathbf{x}_j\|}, \quad (6.6)$$

we propose a simple phenomenological kernel defined as

$$W(\mathbf{x}_i, v_i, \mathbf{x}_j) = \left[a_f \mathbb{I}\left(\alpha_{ij} > \alpha\right) + (1 - a_f)\mathbb{I}\left(\alpha_{ij} < -\alpha\right)\right] \cdot \mathbb{I}\left(\|\mathbf{x}_i - \mathbf{x}_j\| < r_c\right), \quad (6.7)$$

where a_f is the share of attention which is located in the front, α is a critical angle that modulates how wide the attention kernel is, and r_c is a critical distance beyond which cells do not pay attention to any neighbours. Note in particular here that there is no dependence on the distance between cells, which matches qualitatively with the findings from the attention network experiment, in which the weights do not gradually decrease as the distance from the focal cell is increased.

For the attraction-repulsion force, which can be assumed to be isotropic as it relates to volume exclusion and cell-cell adhesion, one can use the canonical Morse potential, $\mathcal{M}(r)$, in the form

$$\mathcal{M}(r) = C_R e^{-r/L_R} - C_A e^{-r/L_A}, \tag{6.8}$$

and obtain that the attraction-repulsion force, f_{ar}, is proportional to this potential. We will assume that alignment is always anistropic, and that the degree of alignment does not further depend on the distance to the focal cell, which yields an alignment potential in the form of Equation (6.7). This allows us to write:

$$\gamma v_i = k_r \sum_{j=1}^{N} \mathcal{M}(\|\mathbf{x}_i - \mathbf{x}_j\|)(\mathbf{x}_i - \mathbf{x}_j) - \eta \sum_{j=1}^{N} \frac{W(\mathbf{x}_i, v_i, \mathbf{x}_j)}{\sum_{k=1}^{N} W(\mathbf{x}_i, v_i, \mathbf{x}_k)}(\mathbf{v}_i - \mathbf{v}_j) + \sigma\xi_i, \tag{6.9}$$

where the second term is taken to be zero if the denominator is zero, k_r is the strength of the attraction-repulsion force, η represents the strength of alignment to the neighours, and σ is the standard deviation of the noise. This yields a minimal interacting particle model that can be used to probe the roles of anisotropic attention in interacting cell models. We propose that it can be used to probe the role of attention in promoting cellular invasion. Finally, this model can be used to probe the mechanistic origins of different attention patterns.

Although we can prescribe the functional form of Equation (6.7) as a starting point to incorporate cellular anisotropy in interacting particle models, such patterns might well arise from other sources, including altered patterns in cell-cell interactions. For this reason, a first starting point is to understand whether a mechanism in the form of Equation (6.9) might be responsible for creating anisotropic patterns of attention could be to first simulate synthetic data with the model in Equation (6.9), and see if the attention patterns generated for this data yield attention maps similar to the corresponding attention kernel in Equation 6.7.

In summary, we propose that future mathematical models of interacting cells should take into account the highly anisotropic patterns of attention that we have observed in the data set studied in this chapter, and we have provided a starting point for its exploration.

Chapter 7

Conclusions and outlook

The contributions of this thesis have been towards data-driven mathematical modelling of collective cell migration. This thesis was organised around three grand challenges in modern mathematical biology: using mathematical modelling to control biological systems; connecting detailed mathematical models with high-dimensional data; and learning new models from data. With each chapter contributing to several of the grand challenges, we conclude this thesis with a discussion of the different grand challenges addressed in this work, and discuss the contributions of this thesis towards each. We shine light on future work in each of the areas, with an overview of opportunities for both experimental and theoretical follow-ups to the main findings of this thesis. We hope that the portfolio of case studies in this thesis can help guide future research in similar areas of mathematical biology and the life sciences, more generally.

7.1 Controlling biological systems

The grand challenge of controlling biological systems with mathematical modelling was described in the introduction as: given a suitable model that can predict a biological system's response to some experimental or clinical intervention, can it also be used to derive a strategy that will lead to a desired outcome?

In Chapter 4, we developed a detailed continuum model for the evolution of the velocity of the centre of mass of an epithelial monolayer during electrotaxis. This model could explain the temporal dynamics of the velocity well under several experimental conditions, which prompted the use of this model to formulate an optimal control problem for several different desired movement patterns, including maximising the total distrance travelled by the monolayer, maximising the velocity of the monolayer, and enforcing constant velocity of the monolayer. This approach can be readily extended to other target velocity patterns and provides a template for the optimal control of monolayer velocities in a wide range of

settings. The model predictions presented in Chapter 4 can be tested experimentally by implementing these bespoke EFs *in vitro*. By comparing the resulting experimental data with the theoretical predictions, one can further refine and calibrate the models presented.

While controlling the velocity dynamics of the centre of mass of an epithelial monolayer is a necessary first step to control the collective migration of an epithelial monolayer, we identified the need to consider whole-tissue models of collective electrotaxis owing to the significant spatial heterogeneity that occurs during collective electrotaxis. In Chapter 5, we showed that the interplay between endogenous migratory cues at the tissue edges and the exogenous cue provided by the EF is responsible for creating a complicated spatial patterning of the velocity field. By carefully comparing experimental data of collective electrotaxis, we developed a detailed continuum modelling framework to describe the velocity field in the monolayer, by taking into account the different field responses to the EF in the different regions of the tissue. We used this framework to propose a heuristic approach to designing a spatially varying EF that can be used to override the endogenous migratory cues and force the tissue to migrate in a desired movement pattern. There are several ways in which our modelling approach can be enhanced for applications in bio-engineering on the one hand, and for a better understanding of the cell signalling and mechanics involved in establishing collective electrotaxis, on the other. On the mathematical side, one could develop an optimal control framework, like that of Chapter 4, to describe the desired velocity patterns inside the monolayer as the minimisers of a given functional. By doing this, one would overcome the need to work with the heuristic EF introduced in Chapter 5. The coupling of the PDE for the evolution of the cell density with the ODE model for the effective signal and inhibitor that dictate the strength of the response to the EF is likely mathematically challenging in the optimal control context both from a conceptual and a numerical point of view.

On the experimental side, our approach can be used to provide detailed spatially and temporally varying EFs to control tissues so that they move in potentially complicated spatial patterns. While some studies have attempted to move tissues along EFs that are not uniaxial [40, 125], the goal of implementing EFs is still restricted to moving the centre of mass, or moving the tissue as a whole. The approach proposed in Chapter 5 provides a means to control each region of the tissue separately, which is promising for the control of monolayer migrations in situations where one wants to control specific parts of the tissue. One such example was already encountered in Chapter 5, where we showed how the EF could be used to override the direction of migration at the tissue edges. Other applications that can arise immediately are increasing or decreasing local cell density by inducing a convergent or divergent EF, respectively, in the bulk of the tissue. This would provide a non-invasive way to manipulate cell densities, which is useful in studying cell-density dependent mechanics and proliferation dynamics.

Finally, we hope that our framework can be used to develop closed-loop approaches to controlling collective cell migration during electrotaxis. Realistic applications of electrotaxis, which include wound healing and advanced tissue engineering, rely on applying an EF in settings where migration has already started, or where there are other, possibly unknown migratory cues (such as chemokines in wound healing). As a result, when controlling such biological systems, one needs to respond in real-time to the observed migratory velocity, and adapt the EF so that the tissue migrates in the desired direction and velocity of migration. The modelling approaches in this work then need to be carefully studied through the lens of uncertainty quantification, so that one obtains not just a prediction, but also a measure of certainty of whether the applied EF can control the migration of the monolayer in the desired way.

7.2 Connecting mathematical models with high-dimensional data

In this thesis, we used Bayesian inference in a wide range of case studies to relate experimental data to mathematical models of collective cell migration. In Chapter 4 and Chapter 5, we used an MCMC approach to Bayesian inference to obtain posterior distributions of model parameters. The resulting distributions were used to calibrate the models in the respective chapters, which allowed us to generate physically realistic predictions of the monolayer velocities in response to the EF. While most parameters could be confidently identified from the data, we encountered a case in Chapter 5 where some of the parameters were not readily identified using MCMC inference, which prompted the use of a careful profile likelihood analysis to probe sources of non-identifiability. By doing this analysis, we found that changes in one of the non-identifiable parameters resulted in changes of the model predictions that were very small compared to the uncertainty expressed in the experimental evidence. This gave rise to a practical recommendation to increase the temporal sampling rate and average over a larger number of biological replicates to reduce the noise. Together, our approach here can be used more generally to understand under which circumstances the experimental protocol needs to be changed or refined in order to account for the inference procedure used.

Our most significant contribution to the challenge of connecting mathematical models with high-dimensional data lies in the work presented in Chapter 2. In Chapter 2, we proposed a novel approach to approximate Bayesian computation in the context of using detailed agent-based modelling to describe a high-throughput scratch assay experiment of human endothelial cells. The motivation for developing this approach arose from the fact that the data used were high-dimensional, and consisted of many experimental

replicates. This fact, combined with the detailed agent-based model that incorporated density-dependent interactions describing the behaviours of the individual cells in the experiment, meant that traditional inference methods were too slow. For this reason, we proposed an extension to ABC based on stochastic gradient descent, which was a methodological advance that accommodated for such vast amounts of data.

There are few areas in which the contributions of this work are significant, and have not yet been discussed in the discussion section of Chapter 2. In Chapter 2 we carried out a careful uncertainty quantification that involved exploring how the results of the inference procedure would change when one changed either the available experimental data, or the hyperparameters used in the inference algorithm. On the one hand, the procedure we used to establish the validity of conclusions with varying experimental data available provides a blueprint for future studies. By carefully considering how outputs of the inference procedure are affected under different amounts of data, and different quality of data, we provide a template for the quality checks that can be carried out when one is attempting to convey the uncertainty underlying experimental results. At this point, it is worth mentioning that our capacity of doing such uncertainty quantification relied on our capacity to generate synthetic data from the mathematical model. This shows another practical use of mathematical modelling in designing optimal experimental setups. Finally, with regards to carrying out uncertainty quantification, we remark that while the model and data presented in Chapter 2 are specific to the application of a scratch assay involving wound healing of human endothelial cells, the ethos of the approach presented in Chapter 2 is sufficiently general so as to apply to other large datasets and other computationally costly likelihood estimation methodologies.

7.3 Learning new models from data

The final grand challenge considered in this work is that of learning new models from experimental data. This thesis presented two case studies in which we have demonstrated, using a data-driven approach, that new mathematical models can be learnt from data.

In Chapter 3, we used a continuum model to infer the relationship between the average amount of power expended by single cells and their active forces within an MDCK epithelial monolayer. We found evidence for a previously undiscussed mechanism that relates the average rate of mechanical energy expenditure to future changes in velocity: regions of the epithelial monolayer that have expended more power than the average across the tissue will reduce their future force production, and regions of the epithelial monolayer that have expended less power than the average across the tissue will increase their future force production. This chapter demonstrated a case study in using data-driven modelling to identify and characterise new mechanisms using experimental data. We used the novel

mechanism that links mechanical power to future force production in a model extension in Chapter 4. In doing so, we found that incorporating cellular energetics provided a way to describe the data from electrotaxis experiments running over multiple stimulation windows. These data show that there is a signficant slowdown between multiple windows of EF stimulation, which cannot be explained by adaptation to the field alone. In this way, by using a model learned from data, we have extended and refined an model for electrotaxis that can capture the velocity dynamics of electrotaxis in a single window of stimulation to give a model that can capture the dyamics of the velocity observed in more complicated experiments. This, again, paved the way to designing an optimal control framework for stimulating MDCK epithelial monolayers with an EF to guide collective electrotaxis.

In the context of electrotaxis, data-driven equation learning can be used to answer some of the unknown phenomena identified in this thesis. For example, we identified in Chapter 5, that a straight-forward model that expresses the velocity in different regions of an epithelial monolayer using linear superposition of competing migratory cues overestimated the velocity in the leading edge, where the endogenous cue to migrate outward is in the same direction as the EF. To understand the precise interplay of polarity and the EF, one would be greatly aided by understanding the detailed dynamics of polarity and cytoskeletal organisation of the constituent cells in the different regions of the monolayer. For that reason, a data-driven approach, wherein changes in polarity and cytoskeletal organisation are related to the EF and the resulting velocity, can be used to develop more refined models that describe how the cells integrate the different cues, which could overcome the issue of overestimating the velocity at the leading edge.

In our final case study, in Chapter 6, we used data-driven approached to learn the dynamics of different populations of neural-crest derived c8161 melanoma cells when they were subjected to genetic perturbations using RNAi. By using detailed cell trajectory data obtained using confocal time-lapse imaging, we characterised the relationship between spatial position in a migrating population of c8161 cells and the migratory behaviour of single cells. In particular, we learned that, regardless of genetic perturbations, all c8161 cell populations display a characteristic behaviour, in which cells at the tissue edge behave much more directionally and persistently than cells in the tissue bulk. By using statistical assessment of traditional migratory metrics, we quickly converged on a set of four knockdowns that led to abnormal leader-follower dynamics in the c8161 cell line. Our body of work in the characterisation of leaders and followers in the c8161 cell line results in useful findings that can help inform realistic agent-based models of collective migration in neural-crest derived cell populations, through the provision a range of observed values for the measured migratory metrics.

In addition to characterising the migratory behaviours of different cell subpopulations within an invading collective of c8161 melanoma cells, the work in Chapter 6 also elucidated the different patterns of cell-cell interactions during collective migration. Deep learning

attention network analysis revealed several gene knockdowns that altered both the cell migratory properties as well as the nominal cell-cell interactions during *in vitro* migration of c8161 melanoma cells. In the context of collective cell migration, it has very recently been reported that the spatial extent to which cells 'pay attention' to each other during migration is significant, and that different cell lines display different nominal attention patterns [41]. The contributions of this work lie in establishing that cell-cell attention patterns are altered when different genes are knocked down, and that the resulting changes in attention patterns can be drastic. While these results challenge our understanding of collective cell migration by indicating that cells have a characteristic spatial distribution with which they interact with their neighbours, more attention is needed to understand precisely the mechanical interpretation of the attention networks inferred using the deep attention networks analysis proposed by LaChance *et al.* [41] and deployed in this work.

The consequence of having identified distinct spatial patterns of cell-cell attention on future work in the field is two-fold. Within the area of mathematical modelling, the results in Chapter 6 point toward an important role for the spatial distribution of cell-cell attention in collective migration. One question of interest, for example, is: do changes in cell-cell attention translate to different migration patterns, such as the spatial distribution of velocities, or the average speed of migration? A final question, on the experimental side, is to investigate the bio-molecular causes for altered cell-cell attention. Since the analysis in Chapter 6 did not provide a direct relationship between the function of the knocked down genes and the resulting characteristics of the altered cell-cell attention patterns, more research is needed to understand why cells display different patterns upon genetic perturbation with RNAi. Such research should involve detailed imaging of sub-cellular components to reveal which changes in cell organisation and morphology underlie the observed changes in migratory behaviour.

7.4 Concluding remarks

Modern quantitative biology offers a wealth of experimental data that needs to be analysed, interpreted, and embedded into applications using sophisticated mathematical techniques. To do so, practitioners must overcome several challenges, of which three played centre stage. In this thesis, we have proposed several computational approaches to controlling biological systems, combining detailed mathematical modelling with experimental data, and learning new models from experimental data. We hope that the work presented in this provides a foundation and a starting point for future studies in quantitative biology and tissue engineering.

Appendix A

Uncertainty quantification for Chapter 2

A.1 Data processing: removing data artifacts

A.1.1 Presence of both wound edges and final condition

Owing to the large number of samples in the high-throughput experiment, there exists large variations in the wound locations recorded in Williams *et al.*'s data set [43]. While the algorithm described in Section 3 is well-placed to incorporate a variable wound location, measurements where the image of the initial condition fails to capture one of the wound edges are problematic for the implementation. Given the zero-flux boundary conditions, when one (or both) wound edges are absent, model simulations will result in there being no cells present on that side of the wound, meaning that for the simulated final condition there will be no population of cells on that side of the wound. However, experimental data can potentially show outgrowth into the field of view, leading to a large discrepancy between observed and simulated data – even when the model would be able to capture the dynamics on one of the edges of the wound well. To mitigate this effect, we use the subset of the data in which both edges of the wound are present. We also screen so that only observations with both an initial and final condition are considered. From the total of 332 genetic knockdowns present in our data of Williams *et al.'s* high-throughput screen, applying both filters yields 118 candidates for investigation.

A.1.2 Presence of extruded cells in the wound area

After having selected for the samples in which there is an initial and final condition as well as containing both wound edges, we use a wound mask to remove extruded cells in the wound area. During the experiment, some very isolated cells appear in the centre of the wound. These cell locations are recorded by the DeepScratch protocol [63], but the cells identified are dead and have been extruded by the tissue. Hence, we apply a mask to filter out the locations of extruded cells. For each sample, we compute a mask with the outline of the wound, and remove all cells that are on the inside of the wound.

A.2 ABC-SMC parameter sampling

A.2.1 Minibatch-ABC

In this section we describe in detail the ABC algorithms used for inference in Chapter 2. As noted in Chapter 2, the idea behind ABC-SMC is to propagate a sample from the prior distribution, $\boldsymbol{\pi}(\theta)$, through a sequence of intermediate distributions towards the target distribution, $P(\boldsymbol{\theta} \,|\, \mathcal{D}_{\text{obs}})$. The intermediate distributions are typically a sequence of ABC approximations, $P_t(\boldsymbol{\theta} \,|\, \mathcal{D}_{\text{obs}})$ for $t \in \{1, \ldots, T\}$. We use a sequential importance sampling approach to ABC-SMC [66, 68, 69], also called population Monte Carlo. Each Monte Carlo sample $\{\boldsymbol{\theta}_n^{(t)}, w_n^{(t)}\}$ built at generation t is used to construct an importance distribution by perturbing the Monte Carlo samples with a kernel $K(\bullet \,|\, t)$. This importance distribution is then used to generate the next generation's Monte Carlo sample using importance sampling, detailed in Algorithm 1. The final sample, $\{\boldsymbol{\theta}_n^{(T)}, w_n^{(T)}\}$ forms the output of the ABC-SMC algorithm. The complete ABC-SMC algorithm is summarised in Algorithm 3.

Algorithm 2: Minibatch-importance sampling

Input: Data, $\mathcal{D}_{\text{obs}} = \{y^i_{\text{obs}}\}_{i=1}^{N}$; number of parameters to sample, N_{gen}; number of parameters to accept, N_{acc}; batch size, N_{bs}; prior distribution, $\pi(\boldsymbol{\theta})$; importance distribution, $\hat{q}(\boldsymbol{\theta})$; model, $f(\bullet \,|\, \boldsymbol{\theta})$; distance function, d.

Output: N_{acc} weighted samples $\{\boldsymbol{\theta}_n, w_n\}_{n=1}^{N_{\text{acc}}}$.

for $n = 0, \ldots, N_{gen}$ **do**
    Generate parameter vector $\boldsymbol{\theta}_n$ according to given importance distribution $\hat{q}(\boldsymbol{\theta})$;
    Set $\epsilon_n = 0$;
    for $i = 1, \ldots, N_{bs}$ **do**
        Draw data index ξ uniformly at random from $\{1, \ldots, N\}$;
        Simulate $y_{\text{sim}} \sim f\left(\bullet \,|\, \boldsymbol{\theta}_n, y^{\xi}_{\text{obs}}(0)\right)$;
        Update $\epsilon_n = \epsilon_n + d(y_{\text{sim}}, y^{\xi}_{\text{obs}})$;
    end
end
From $\{\boldsymbol{\theta}_n, w_n\}_{n=1}^{N_{\text{gen}}}$, select the N_{acc} parameters with lowest value of ϵ_n;
Set $w_n = \pi(\boldsymbol{\theta}_n)/\hat{q}(\boldsymbol{\theta}_n)$ for all accepted parameters $\boldsymbol{\theta}_n$.

Algorithm 3: Sequential Monte Carlo ABC

Input: Data, $\mathcal{D}_{\text{obs}} = \{y^i_{\text{obs}}\}_{i=1}^{N}$; number of parameters to generate, N_{gen}; number of parameters to accept, N_{acc}; batch size, N_{bs}; prior distribution, $\pi(\boldsymbol{\theta})$; importance distribution, $\hat{q}(\boldsymbol{\theta})$, proportional to $q(\boldsymbol{\theta})$; model, $f(\bullet \,|\, \boldsymbol{\theta})$; distance function, d; number of generations, T.

Output: Weighted samples $\{\theta_n^{(T)}, w_n^{(T)}\}_{n=1}^{N_T}$.

Set $q_1 = \pi$;

for $t = 1, \ldots, T-1$ **do**

  Produce weighted samples $\{\theta_n^{(t)}, w_n^{(t)}\}_{n=1}^{N_t}$ using Algorithm 1 (ABC-IS) with importance distribution $\hat{q}_t$, number of parameters to generate N_{gen}, number of parameters to accept N_{acc}, batch size N_{bs}, model f and distance d;

  Define the importance distribution, $\hat{q}_{t+1}$, proportional to

$$q_{t+1}(\theta) = \begin{cases} \sum_{n=1}^{N_t} w_n^{(t)} K_t(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}_n^{(t)}) / \sum_{m=1}^{N_t} w_m^{(t)} & \text{if } \pi(\boldsymbol{\theta}) > 0 \\ 0 & \text{else;} \end{cases}$$

end

A.2.2 Choice of knockdown kernel

In this work, we use the multivariate normal kernel proposed by Filippi *et al.* [52], which is designed to approximate the target density well, while also maximising the acceptance rate at each step [52]. In effect, this kernel perturbs each sampled particle according to a multivariate distribution with covariance matrix $\Sigma^{(t)}$, which depends on the covariance of the previous population. Filippi *et al.* [52] prove that the optimal choice of kernel is then approximately

$$\Sigma^{(t)} \approx \sum_{k=1}^{N_{\text{acc}}} \sum_{\ell=1}^{N_{\text{acc}}} w_k^{(t-1)} w_\ell^{(t-1)} (\boldsymbol{\theta}_k - \boldsymbol{\theta}_\ell)(\boldsymbol{\theta}_k - \boldsymbol{\theta}_\ell)^T. \tag{A.1}$$

This allows us to define the proposal mechanism $K(\bullet|\boldsymbol{\theta}_n^{(t)})$ defined in Algorithm 3 as

$$K(\bullet|\boldsymbol{\theta}_n^{(t)}) \sim \mathcal{N}(\boldsymbol{\theta}_n^{(t)}, \Sigma^{(t+1)}), \tag{A.2}$$

where $\Sigma^{(t+1)}$ is defined in Equation (A.1).

A.2.3 Model calibration: number of ABC-SMC generations

For each ABC-SMC implementation of this specific model, we must define the number of generations, T. This is an additional hyperparameter whose choice requires further justification. Choosing a higher number of generations will be computationally more costly, while choosing a lower number of generations risks not approximating the true posterior accurately. To balance between these two requirements, we perform test runs of minibatch ABC-SMC with our chosen batch size of $N_{\text{bs}} = 10$ with a varying number of generations to quantify the gain of information after each subsequent generation. One way to measure this gain is the Kullback-Leibler (KL) divergence [70] between each generation. The KL divergence for two probability distributions, p and q, defined on a common space Ω, is defined as

$$\text{KLDiv}(p\|q) = \int_{\Omega} p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x. \tag{A.3}$$

The KL divergence as a measure of change of the distribution between one generation and the next has been used previously by Filippi *et al.* [52] to derive kernels that optimally increase the difference between two generations. Note that we also use the resulting kernel proposed by Filippi *et al.* [52]. In this work, we similarly consider how the quantity

$$d_t = \text{KLDiv}(p_{t+1}\|p_t), \tag{A.4}$$

evolves over time, where p_t is the ABC-SMC posterior after t generations. We plot d_t for the three genetic knockdowns for which most data is available (Mock, CDC42, and CDH5 knockouts) in Figure A.1 to find that including additional generations after $T = 4$ generations have virtually no impact on the posteriors, hence we choose this cut-off for the number of generations.

A.3 Justification of model reduction

We explore the impact of adding a density-dependent term to contribute to the rate of cell death, analogous to those used to model cell movement and proliferation. This gives rise to two competing models. In the simpler model, death occurs at a constant rate, d, independent of local crowding. In the more complex model, the density-dependent death

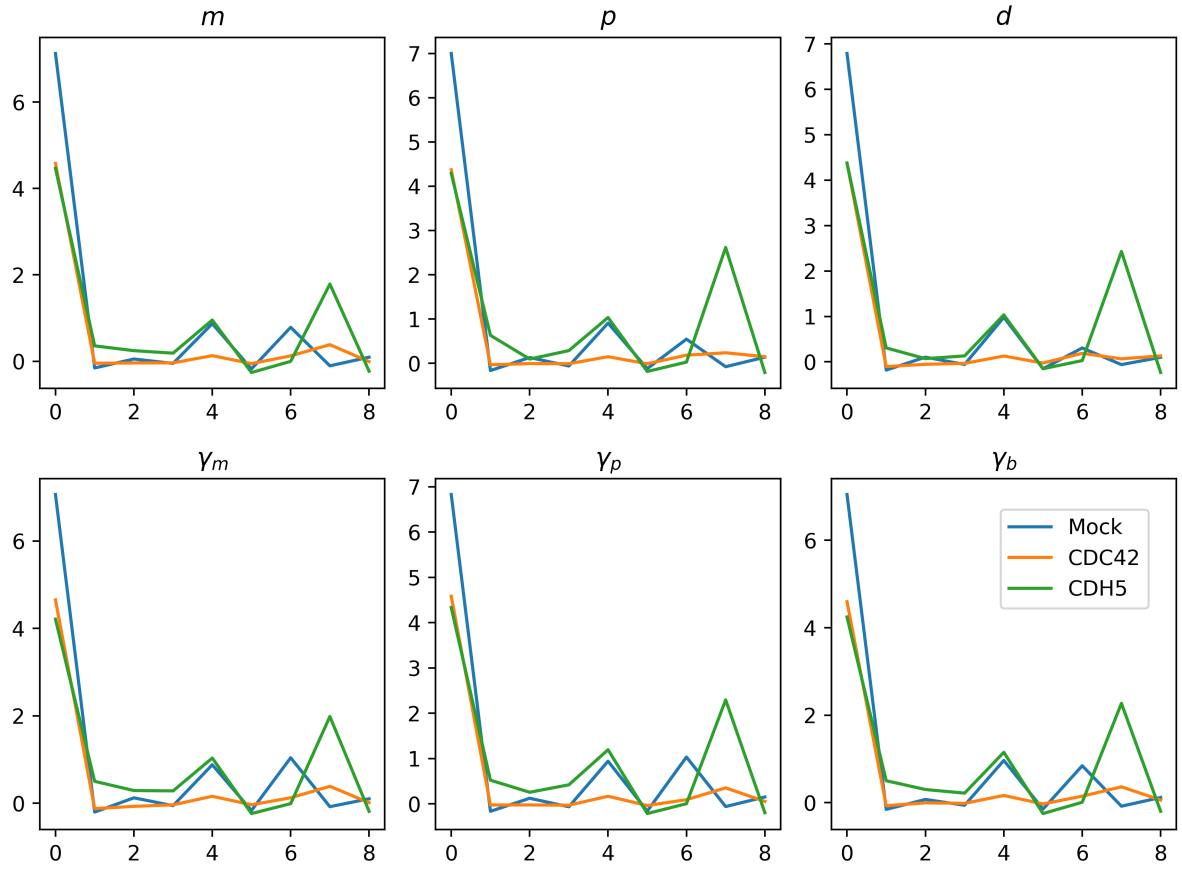


Figure A.1: KL divergence between the posterior distributions at one generation and the next, plotted for a total of 10 generations. The KL divergence has decreased significantly after just two generations as compared to the prior and subsequent generations do not diminish the KL divergence substantially.

rate for cell n, D_n, is defined by

$$D_n = \max\left\{0, d - \gamma_d \sum_{i=1, i\neq n}^{N(t)} \exp\left(\frac{-\|\boldsymbol{x}n - \boldsymbol{x}_i\|^2}{2\sigma^2}\right)\right\}. \tag{A.5}$$

To quantify whether the data contain evidence in favour of one model as opposed the other, we compute the Bayes factor between the simple (density-independent) model, M_{simple}, and the complex (density-dependent) model, M_{complex}. The Bayes factor is then defined as

$$\mathcal{B} = \frac{\mathbb{P}(\mathcal{D}_{\text{obs}}|M_{\text{simple}})}{\mathbb{P}(\mathcal{D}_{\text{obs}}|M_{\text{complex}})}, \tag{A.6}$$

and the value of $\mathcal{B}$ quantifies the strength of the evidence in favour of M_{simple} [66]. A value of $\mathcal{B}$ between 1 and 3 represents very weak evidence in favour of M_{simple}, a value between 3 and 20 represents positive evidence, between 20 and 150 strong evidence and above 150 very strong evidence [67]. From the ABC-SMC posterior distributions the Bayes factor for each model can be computed straightforwardly [65]. For a given set of population weights $\{w_i^t\}_{i=1,\ldots,n,t=1,\ldots,T}$, where n is the number of samples in each generation, and T is the number of generations,

$$\mathbb{P}(\mathcal{D}_{\text{obs}}|M) = \prod_{t=1}^{T} \frac{1}{n} \sum_{i=1}^{N} w_t^i. \tag{A.7}$$

For the Mock, CDC42 and CDH5 knockout cases, we compute the Bayes factor, $\mathcal{B}$, using the simple and complex models. This gives $\mathcal{B} = 10.7, 4.69, 4.34$ for Mock, CDC42 and CDH5 knockouts, respectively, meaning that the data provides positive evidence in favour of the simpler model.

A.4 Reliable inference using small batch sizes

The introduction of minibatches in ABC-SMC comes with the task of selecting the batch size as an algorithm hyperparameter. A good choice for the batch size, N_{bs}, needs to take into account the balance between computational complexity on the one hand and a reliable estimate of the discrepancy between model and data on the other hand. In general, the capacity of any ABC implementation to infer parameter values from data will depend on the specific computational model and the data available (for instance, through its spatial and temporal resolution, or its signal to noise ratio). To investigate the performance of minibatch ABC-SMC in identifying parameters from data in our

specific context, we define two different parameter regimes and assess the performance of minibatch-ABC-SMC in identifying the parameter values on a synthetic dataset created using the parameter values in the two regimes. We let parameter regime I correspond to a high movement rate with a positive impact of density on movement and a low proliferation rate, and choose $m = 1.5\text{h}^{-1}$, $p - d = 0.01\text{h}^{-1}$, $\gamma_m = -1\text{h}^{-1}$, $\gamma_p = 0.01\text{h}^{-1}$ and $\gamma_b = 20\mu\text{m}$ (to sample $p - d = 0.01\text{h}^{-1}$, we sample $p \in (0, 0.05)$ and then choose $d = 0.01 - p$). We let parameter regime II correspond to a low movement rate with contact inhibition for movement and a high proliferation rate, and choose $m = 0.5\text{h}^{-1}$, $p - d = 0.025\text{h}^{-1}$, $\gamma_m = 1.5\text{h}^{-1}$, $\gamma_p = 0.01\text{h}^{-1}$ and $\gamma_b = 20\mu\text{m}$ (p and d are sampled analogously as in parameter regime I).

A.4.1 Variability of the posterior mean

For each parameter regime, an *in silico* data set is created by simulating one model output at $t = 24\text{h}$ for every initial condition corresponding to Mock. That is, each of the two *in silico* datasets contains 117 data points, each corresponding to one data point in the *in vitro* dataset from the high-throughput experiment. Then, we perform minibatch-ABC-SMC with each of the two synthetic data sets with varying batch size, such that $N_{\text{bs}} = 1, 5, 10, 50, 117$. For each parameter regime and batch size, we repeat the inference process five times to quantify the variability of the estimated posterior distributions. Figures A.2 and A.3 show boxplots for the resulting posterior means for each of the parameter regimes as the batch size is increased. It is clear from Figures A.2 and A.3 that the performance of the algorithm in identifying the true parameter value (horizontal lines in the figure) is affected by variations in the batch size, N_{bs}. However, there is a critical value – in Figures A.2 and A.3 this is typically around $N_{\text{bs}} = 10$ – after which increasing the batch size, N_{bs}, does not reduce the variability of the estimated posterior means. Similarly, after this critical value, the capacity of minibatch-ABC-SMC to identify the true posterior mean does not improve, even when the batch size is increased to the full data size. At any value of the batch size, N_{bs}, the minibatch-ABC-SMC implementation is capable of correctly discerning structural differences between the two parameter regimes. For instance, for the data generated using parameter regime I, the estimated intrinsic motility rate, m, is consistently estimated as much higher than that of the data from parameter regime II, as is consistent with the definition of the parameter regimes. Conversely, the net intrinsic proliferation rate, $p - d$, is estimated as much higher in parameter regime II than in parameter regime I. For the density-dependent interaction term, γ_m, the minibatch ABC-SMC implementation can discern at any batch size that the effect of crowding on motility is positive in regime I, and negative in regime II.

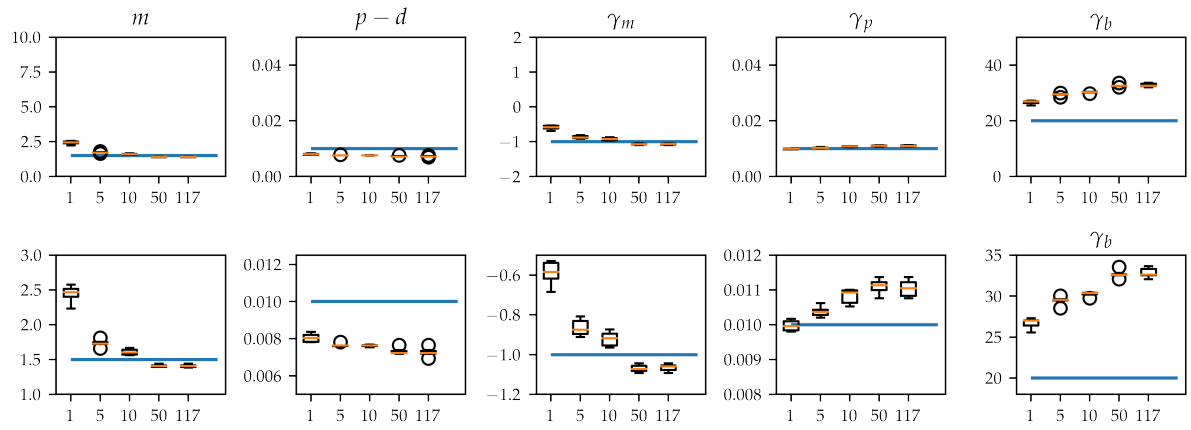


Figure A.2: Boxplots for the means of the posterior distributions obtained by performing minibatch ABC-SMC on data generated in parameter regime I when the batch size is varied. The blue horizontal line indicates the true parameter value. Top row: boxplots over the full prior range. Bottom row: zoomed range to more accurately display the discrepancy between estimates and the true parameter value.

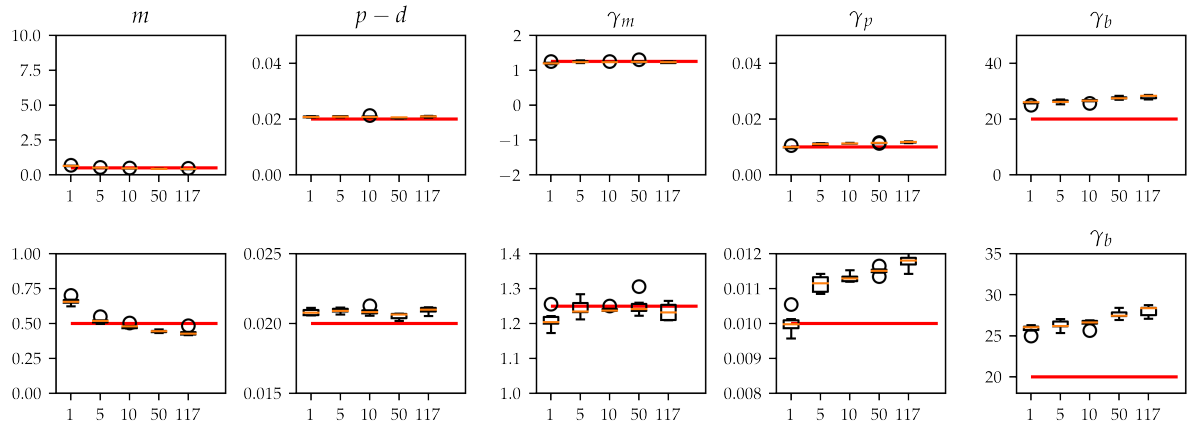


Figure A.3: Boxplots for the means of the posterior distributions obtained by performing minibatch-ABC-SMC on data generated in parameter regime II when the batch size is varied. The red horizontal line indicates the true parameter value. Top row: boxplots over the full prior range. Bottom row: zoomed range to more accurately display the discrepancy between estimates and the true parameter value.

A.4.2 Variability of the posterior variance

To further assess the similarity of the learned posterior distributions when the batch size, N_{bs}, is varied, we choose one of the posterior distributions for each of the experiments using each of $N_{\text{bs}} = 1, 5, 10$ and show the distributions in Figure A.4. The behaviour of the posterior distributions in Figure A.4 is in agreement with the idea that the qualitative behaviours of the model can be estimated well even at small batch size, as even the smallest batch sizes provide posterior distributions with posterior means close to the true value used to generate the synthetic data. It is clear from Figure A.4 that increasing the batch size, N_{bs}, helps to decrease the spread of the posterior distribution.

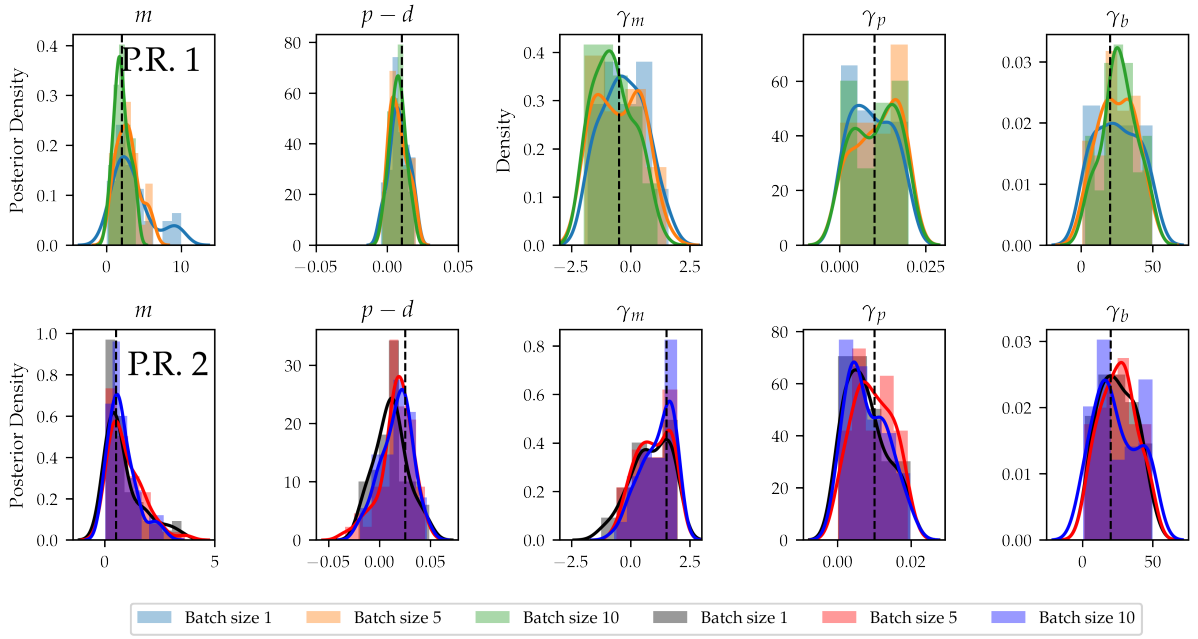


Figure A.4: Posterior distributions obtained using the two synthetic data sets corresponding to parameter regimes I and II. In all plots, the dashed black line indicates the true parameter value. Top row: posterior distributions for the data set generated using parameter regime I. Bottom row: posterior distributions for the data set generated using parameter regime II.

Figure A.4 suggests that increasing the batch size, N_{bs}, reduces the variance of the posterior distribution. To quantify this effect, we analyse the standard deviation of the learned posterior distributions at each different batch size. The standard deviation of the posterior density serves as a useful measure to quantify the uncertainty associated with the posterior distribution. Figures A.5 and A.6 reveal that for three key parameters in the model, the standard deviation decreases with an increasing batch size. Strikingly, the standard deviation drops rapidly when the batch size is increased initially, but then stays relatively constant across larger batch sizes. In both parameter regimes, this happens for three of the parameters around $N_{\text{bs}} = 10$, while for the remaining parameters the posterior standard deviation remains roughly constant when the batch size is increased further. Taken together, the analysis of posterior means and posterior standard deviation suggests that a minibatch implementation of ABC-SMC with small batch sizes can have a similar performance to conventional ABC-SMC (that uses very large batch sizes), while the number of simulations needed is much lower (in this case, 10 simulations per sampled parameter vector instead of 117).

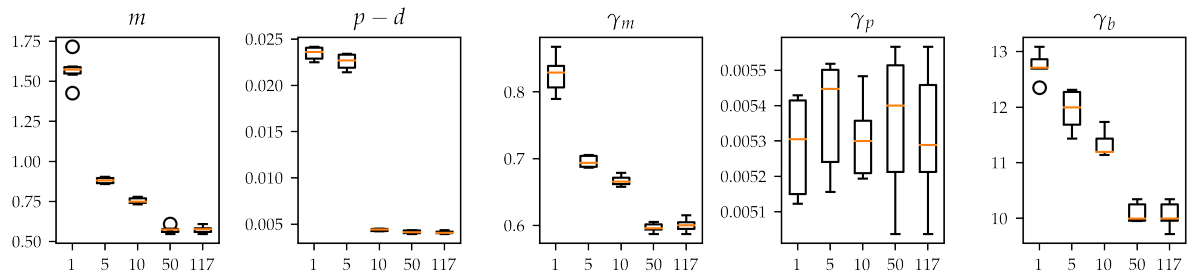


Figure A.5: Boxplots for the standard deviations of the posterior distributions obtained by performing minibatch-ABC-SMC on parameter regime I when the batch size is varied.

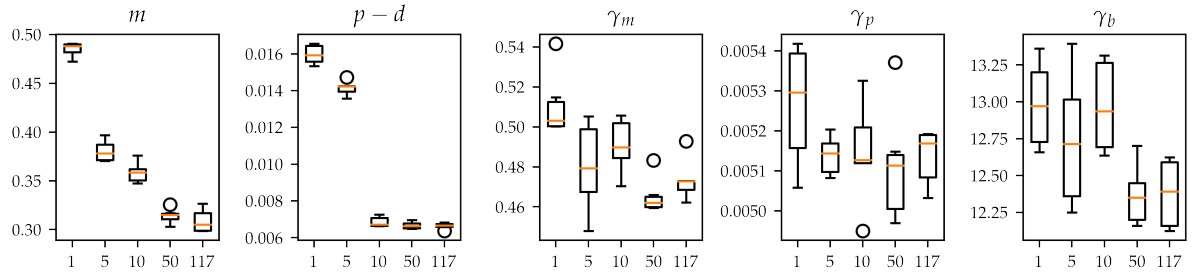


Figure A.6: Boxplots for the standard deviations of the posterior distributions obtained by performing minibatch-ABC-SMC on parameter regime II when the batch size is varied.

A.4.3 Performance of different batch sizes on experimental data

Taken together, our analysis shows that on synthetic data, a minibatch-ABC-SMC implementation has comparable performance to a traditional ABC-SMC implementation in retrieving a known model parameter value and is capable of inferring meaningful mechanistic insights using the parametrised model. To explore the effect that a small batch size has on the qualitative properties of the posterior distribution when applied to real data, we perform inference on the Mock data set, where we vary the batch size from 1 to 117. In Figure A.7, we compare the posterior distributions. Figure A.7 shows that the learned posterior distributions are qualitatively very similar, meaning that the intrinsic variability of the data is captured well by batch sizes as small as $N_{\text{bs}} = 10$. Taken together with the uncertainty quantification carried out for the two parameter regimes that are used to create the synthetic data, we propose to use a batch size $N_{\text{bs}} = 10$ for the analysis of the knockdowns in Chapter 2 for which a large number of observations are recorded in Williams *et al.*'s mRNAi screen [43].

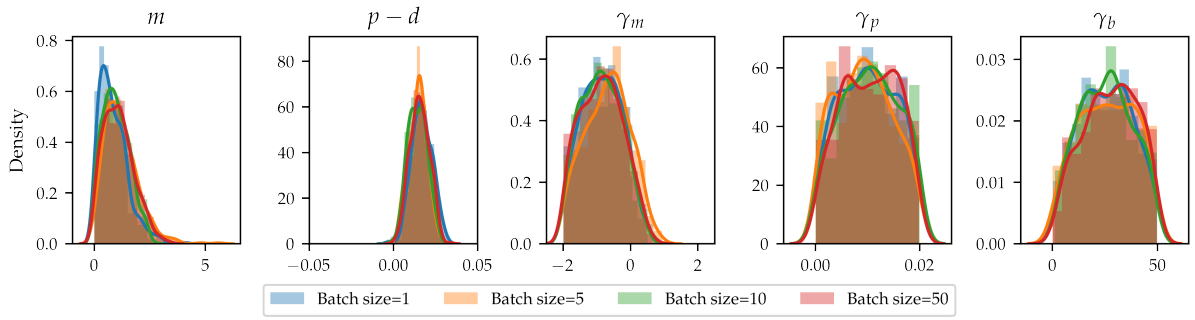


Figure A.7: Posterior densities of model parameters for the Mock data set. When the batch size, N_{bs}, is increased, the posteriors remain qualitatively very similar and consistently identify the same maximum a posteriori parameter value.

A.4.4 Capacity of inference with very small batch sizes on experimental data

Our previous experiments suggest that for this specific model and data, the choice of $N_{\text{bs}} = 10$ offers a good balance between computational complexity, robust inference of the posterior mean and low variance of the posterior distribution. Decreasing the batch size to $N_{\text{bs}} = 2$ still offers reliable identification of the posterior mean, albeit at the cost of a higher variance in the posterior distribution. Since many of the knockdowns in the high-throughput data set contain only two datapoints, we wish to understand if the posterior distributions obtained by applying minibatch ABC-SMC when only two datapoints are available can still lead to the identification of mechanistically relevant differences between the knockdowns. To do this, we select 40 datapoints from each of the Mock and CDH5 knockdown experiments at random and without replacement, and divide them into 20 pairs of two observations for each of the datasets. We perform minibatch ABC-SMC on each of the resulting *miniature data sets* and record their posterior distributions. Figure A.8 shows boxplots for the posterior distributions for each of the Mock and CDH5 knockdown experiments.

From Figure A.8 it can be seen that while the spread of the parameter values is indeed larger for each of the experimental conditions than when the batch size is larger, the boxplots of the posterior means suggest that meaningful differences between the two different genetic knockdowns can be identified. This suggests that even a very small number of observations can be used to identify the posterior distributions of model parameters given data in the high-throughput experiment.

To compare the performance of minibatch ABC-SMC with naive subsampling from the data, we fix $N_{\text{bs}} = 10$ as in the numerical experiments in Chapter 2 and investigate the quality of the posterior distributions obtained by performing ABC-SMC on subsamples of size N_{bs} of the data. From the Mock dataset, we select 100 datapoints at random and without replacement, and divide them into 10 subsets of $N_{\text{bs}} = 10$ datapoints.

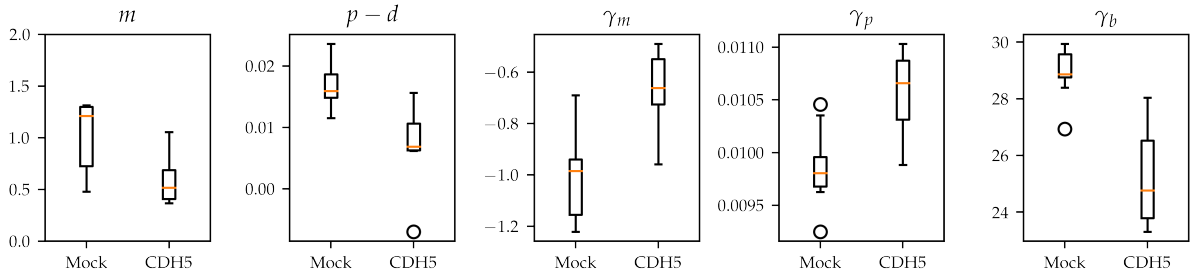


Figure A.8: Boxplots showing the means of the posterior distributions found by performing minibatch-ABC-SMC on *miniature data sets* created by taking subsets of the data. The boxplots show that there are meaningful separations of the posterior means between the Mock and CDH5 knockdowns, even when each *miniature data set* consists of only two datapoints.

As in Appendix A.4.3, we perform ABC-SMC on each of the resulting *miniature data sets* and record their posterior distributions. We independently run 10 instances of of minibatch ABC-SMC, again using $N_{\text{bs}} = 10$, and record the posterior distributions. Figure S9 shows boxplots of the means of the posterior distributions found by performing ABC-SMC on subsets of the data as well as using minibatch ABC-SMC. The boxplots show that the subsample approach yields more variable posterior means than minibatching.

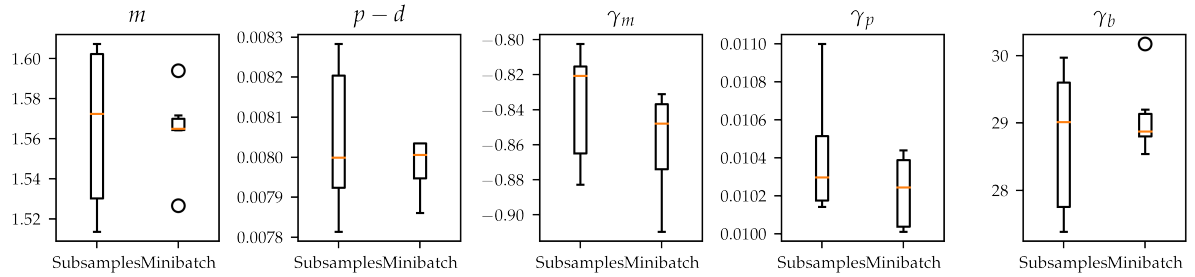


Figure A.9: Boxplots of the means of the posterior distributions found by performing ABC-SMC on subsets of the data as well as using minibatch-ABC-SMC. Subsampling yields more variable posterior means than minibatching.

Figure S10 shows boxplots of the variance of the posterior distributions generated by performing ABC-SMC on subsets of the data as well as using minibatch-ABC-SMC. The boxplots show that the subsample approach yields on average larger posterior variance than minibatching, indicating that the posterior obtained using subsampling contains more uncertainty than that obtained using minibatch ABC-SMC.

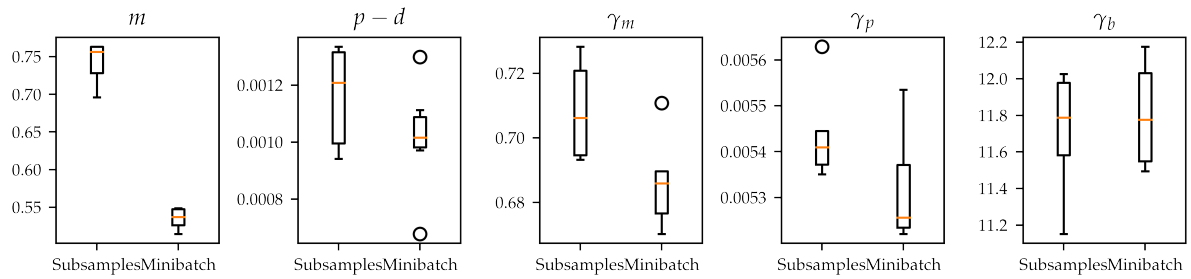


Figure A.10: Boxplots of the variances of the posterior distributions found by performing ABC-SMC on subsets of the data as well as using minibatch-ABC-SMC. Subsampling yields larger posterior variances than minibatching.

A.5 Robustness of model predictions to cell size esimates

An important model parameter in the IBM is the cell size, which is used to define the Gaussian kernel that governs cell-cell interactions. Within each genetic knockdown, the cell area is estimated using the protocol of Javer *et al.* [63] for each each individual observed cell. In Figure A.11 we show the distribution of cell areas across all different genetic knockdowns (for a total of approximately $5 \cdot 10^5$ cells. We see that the area distribution is sharply peaked, with some very large outliers corresponding to errors in the Voronoi tesselation of Javer *et al.* [63]. For this reason, we consider only cells whose areas are measured as smaller than 500 square pixels. For each knockdown, we estimate the cell area

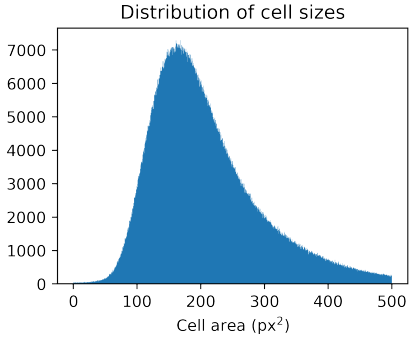


Figure A.11: Distribution of cell sizes in the dataset.

parameter for simulations as the median of the cell area measurements for that knockdown. To address the robustness of our model predictions to errors in cell area measurement, we carry out uncertainty quantification analysis by varying the cell area input parameter for both Mock and CDH5 knockdowns. For each of Mock and CDH5 knockdown, we generate *in silico* data by simulating from the model using a range of cell area measurements defined by variations of up to 50% of the cell area median. This corresponds roughly to simulating within five standard deviations of the input parameter. For each value in the cell area

range, we simulate the model ten times. In Figure A.12, we plot how much the resulting posterior mean changes relative to the posterior mean obtained when using the measured cell area parameter, *i.e.* for μ taken as the posterior mean obtained by using the cell area input parameter used throughout Chapter 2, and $\mu^i_{A_j}$ the i-th posterior mean obtained by using cell area A_j, we estimate the relative error for cell size A_j, ϵ_{rel}, as

$$\epsilon_{\text{rel}} = \frac{\frac{1}{10}\sum_{i=1}^{10} |\mu^i_{A_j} - \mu|}{|\mu|}. \tag{A.8}$$

We find that for cell areas within 25% of the median cell area, which roughly corresponds

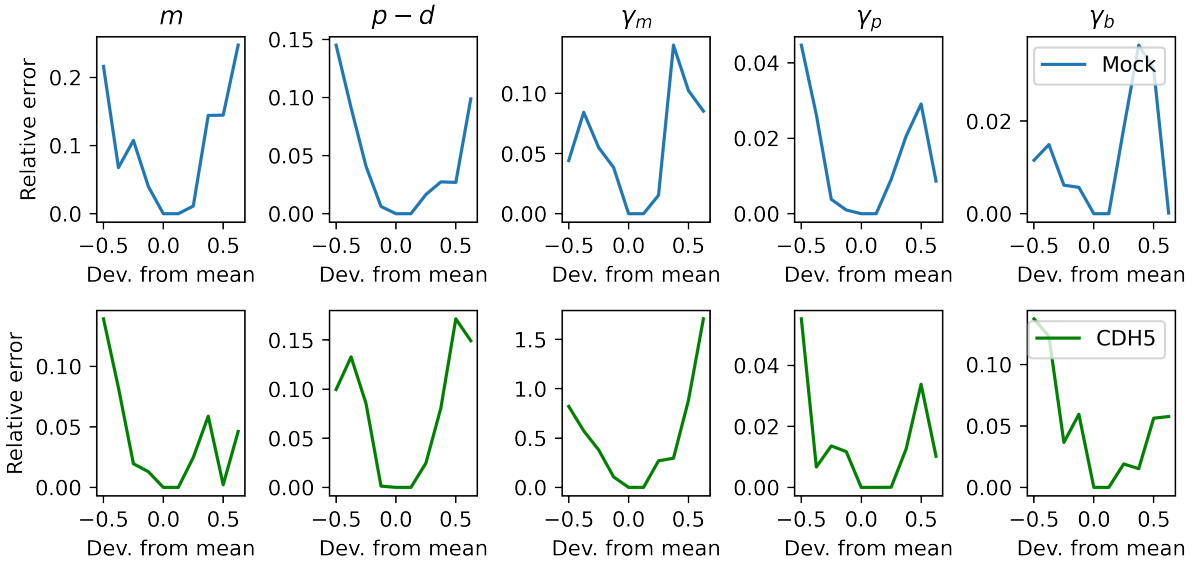


Figure A.12: Relative error plotted against deviations of the cell area as input parameters for minibatch-ABC-SMC. Deviations within roughly three standard deviations from the median result in the relative error in the posterior mean below 10%.

to cell areas within three standard deviations from the median, the relative error in the posterior mean is below 10% for both the Mock and CDH5 knockdown, from which we conclude that our predictions are robust to small errors in the estimation of cell areas.

A.6 Changing the number of clusters for K means clustering

In Section 2.3.3.1, K-means clustering is performed on the posterior means of the different knockdowns for $K = 3$. In this section, we show the sum of squared distances when K is varied in the range $1, \ldots, 10$ to find that marginal improvement in sum of squares rapidly

diminishes above $K = 3$, suggesting that $K = 3$ is a good choice for the number of clusters in the analysis.

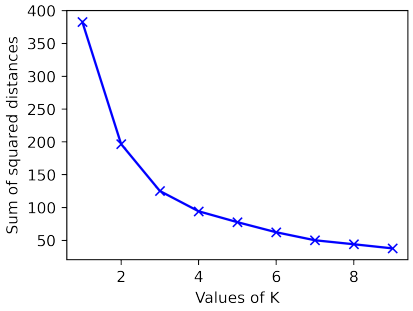


Figure A.13: Sum of squared distances for K-means clustering as K is increased. The marginal improvement of the sum of squared distances markedly decreases after $K = 3$, suggesting that this is the optimal choice for the number of clusters.

A.7 Confidence intervals for model parameters

In this section, we report the confidence intervals computed for the experiments in Section 2.3.2. For each experiment, we compute the 99% confidence interval, which for a sample of size n of a random variable x, with sample mean $\bar{x}$ and sample standard variance σ under the normality assumption, is given by

$$\mathrm{CI} = \left(\bar{x} - z\frac{\sigma}{\sqrt{n}}, \bar{x} + z\frac{\sigma}{\sqrt{n}} \right), \tag{A.9}$$

where z equals the z-score, which for a 95% confidence interval is $z = 1.96$.

A.7.1 Confidence intervals for posterior means of Mock, CDC42, CDH5 knockdowns

Experiment	Posterior mean	Confidence interval
Mock, m	0.7194	(0.6789,0.7599)
Mock, $p-d$	0.01493	(0.01448, 0.01538)
Mock, γ_m	-1.06206	(-1.1092, -1.01493)
Mock, γ_p	0.01014	(0.009652, 0.01062)
Mock, γ_b	29.1372	(28.08166, 30.1927)
CDC42, m	0.3278	(0.3054, 0.3501)
CDC42, $p-d$	0.01089	(0.01030, 0.01147)
CDC42, γ_m	-0.2915	(-0.3121, -0.2709)
CDC42, γ_p	0.009133	(0.008648, 0.009618)
CDC42, γ_b	25.4509	(24.2694, 26.6324)
CDH5, m	0.43168	(0.40364, 0.4597)
CDH5, $p-d$	0.005825	(0.005353, 0.0062966)
CDH5, γ_m	-0.6170	(-0.6524, -0.5816)
CDH5, γ_p	0.01069	(0.01022, 0.01116)
CDH5, γ_b	27.02407	(25.8691, 28.1790)

Table A.1: Posterior means and confidence intervals of posterior means for each of the experiments.

A.7.2 Confidence intervals for posterior means of different initial wound sizes

Experiment	Posterior mean	Confidence interval
Mock, small wound, m	1.4164	(1.3403, 1.4926)
Mock, middle wound, m	0.6873	(0.6506, 0.7242)
Mock, large wound, m	0.8639	(0.8168, 0.9111)
Mock, small wound, $p-d$	0.01869	(0.01825, 0.01913)
Mock, middle wound, $p-d$	0.01479	(0.01436, 0.01522)
Mock, large wound, $p-d$	0.01462	(0.01415, 0.01509)
Mock, small wound, γ_m	-1.1068	(-1.1620, -1.0513)
Mock, middle wound, γ_m	-1.1668	(-1.2102, -1.1234)
Mock, large wound, γ_m	-1.0026	(-1.05234, -0.95288)
Mock, small wound, γ_p	0.010966	(0.01048, 0.01145)
Mock, middle wound, γ_p	0.01136	(0.01089, 0.011835)
Mock, large wound, γ_p	0.009991	(0.009506, 0.01047)
Mock, small wound, γ_b	30.9606	(29.9807, 31.9406)
Mock, middle wound, γ_b	30.8423	(29.8193, 31.8654)
Mock, large wound, γ_b	29.336	(28.2649, 30.4070)

Table A.2: Posterior means and confidence intervals of posterior means of different initial wound sizes for Mock.

Experiment	Posterior mean	Confidence interval
CDH5, small wound, m	0.4034	(0.3738, 0.4330)
CDH5, large wound, m	0.3584	(0.3352, 0.3815)
CDH5, small wound, $p-d$	0.006947	(0.0064504, 0.007443)
CDH5, large wound, $p-d$	0.0068595	(0.006376, 0.0073425)
CDH5, small wound, γ_m	-0.7420	(-0.7818, -0.7022)
CDH5, large wound, γ_m	-0.8277	(-0.8653, -0.7901)
CDH5, small wound, γ_p	0.01090	(0.01042, 0.01139)
CDH5, large wound, γ_p	0.01088	(0.01040, 0.01136)
CDH5, small wound, γ_b	27.366	(26.2392, 28.4930)
CDH5, large wound, γ_b	29.1108	(27.9956, 30.2259)

Table A.3: Posterior means and confidence intervals of posterior means of different initial wound sizes for CDH5 knockdown.

Appendix B

Statistics for Chapter 6

B.1 Full free space and cell fold counts

B.1.1 C8161 cell line

	Area change	Cell change	Area p value	Cells p value
ADM	0.195798	1.228939	0.480699	0.005730
ANXA1	0.156112	1.205149	0.074701	0.000457
ANXA6	0.153308	1.204590	0.056274	0.000818
ARPC3	0.346326	1.236376	0.270917	0.002488
ATP1B1	0.152806	1.212426	0.112657	0.000677
BMP4	0.575718	1.027039	0.035106	0.004357
CA2	0.219851	1.232568	0.561967	0.001305
COL4A4	0.176958	1.237034	0.183920	0.000162
CREB3L2	0.249358	1.237598	0.935144	0.001885
CTNNB1	0.258187	1.188671	0.832608	0.000940
DAB2	0.231965	1.264097	0.758293	0.002436
DKK3	0.283147	1.373354	0.389225	0.484604
DPYD	0.161853	1.243875	0.163762	0.004428
ESM1	0.175589	1.204809	0.271576	0.000255
FLT1	0.172017	1.163745	0.179422	0.001288
FN1	0.109174	1.237674	0.023363	0.005803
FOS	0.114384	1.221061	0.017640	0.001814
HSPB1	0.180172	1.270795	0.159804	0.027378
HSPB8	0.187703	1.244425	0.350790	0.002856
IFI30	0.433646	1.363879	0.004275	0.426149
IGFBP2	0.169157	1.233952	0.174888	0.000174
IL13RA1	0.359084	1.355416	0.026246	0.236051
INHBA	0.332892	1.369781	0.071780	0.470495
ITGB1	0.455130	1.356478	0.002525	0.388973
JUN	0.350960	1.344330	0.070708	0.160492
KCNE3	0.413838	1.355903	0.003068	0.240005

KITLG	0.271513	1.391821	0.520139	0.952590
KYNU	0.391756	1.329141	0.015750	0.072351
LY86	0.214558	1.239344	0.400510	0.000023
MMP11	0.338043	1.318175	0.323939	0.075644
NEXN	0.380099	1.372642	0.006325	0.344845
NRP1	0.351177	1.372356	0.077623	0.565380
Neg	0.245456	1.390080	0.000000	0.000000
PRKCQ	0.409094	1.347402	0.004245	0.210850
PRNP	0.436088	1.157084	0.527842	0.039764
RAC2	0.450711	1.282138	0.062461	0.036473
RASGRP1	0.400240	1.371004	0.017678	0.550462
RASGRP3	0.352173	1.369590	0.092609	0.552177
SERPINI1	0.381621	1.336063	0.131955	0.177954
SH3TC1	0.370675	1.338483	0.034355	0.055398
TUBB6	0.251051	1.379618	0.950177	0.753255
UNC5B	0.338073	1.471418	0.229963	0.139784
UPP1	0.343596	1.331398	0.037685	0.061448
VEGFC	0.239766	1.325796	0.935917	0.062131
VIM	0.382822	1.364478	0.039939	0.403961
Water	0.364654	1.365936	0.064213	0.469137

B.1.2 HT1080 cell line

	Area change	Cell change	Area p value	Cells p value
ADM	0.060671	2.360882	0.100075	0.454369
ANXA1	0.111119	2.310983	0.829313	0.972958
ANXA6	0.109515	2.275246	0.756113	0.568778
ARPC3	0.170999	2.315639	0.256375	0.905142
ATP1B1	0.078286	2.344303	0.146900	0.548516
BMP4	0.228129	2.238961	0.023015	0.249295
CA2	0.198282	2.302874	0.040118	0.925501
COL4A4	0.151819	2.317743	0.303948	0.793022
CREB3L2	0.321569	2.256347	0.011786	0.394986
CTNNB1	0.161098	2.333009	0.189644	0.685487
DAB2	0.152574	2.307474	0.385637	0.986537
DKK3	0.160044	2.212050	0.269088	0.024381
DPYD	0.206739	2.290188	0.029418	0.752397
ESM1	0.136717	2.250454	0.658075	0.324794
FLT1	0.139458	2.312727	0.458627	0.943273
FN1	0.107640	2.360531	0.653361	0.381311
FOS	0.080495	2.336010	0.241023	0.645965
HSPB1	0.154287	2.321480	0.442457	0.835308
HSPB8	0.184105	2.302492	0.195812	0.915435
IFI30	0.163818	2.186225	0.183722	0.052009
IGFBP2	0.152717	2.317739	0.395108	0.814243
IL13RA1	0.167971	2.252620	0.170088	0.330354
INHBA	0.181355	2.270976	0.070265	0.510632

ITGB1	0.237767	2.092749	0.037206	0.001807
JUN	0.115148	2.186637	0.874458	0.052887
KCNE3	0.238361	2.224366	0.025737	0.228415
KITLG	0.121661	2.295384	0.929384	0.818458
KYNU	0.177149	2.213171	0.191290	0.118696
LY86	0.148752	2.368724	0.293066	0.127556
MMP11	0.151985	2.176011	0.319110	0.044410
NEXN	0.179379	2.241230	0.106181	0.086917
NRP1	0.224441	2.240666	0.067912	0.253695
Neg	0.119361	2.308428	0.000000	0.000000
PRKCQ	0.133575	2.085036	0.643451	0.003936
PRNP	0.106043	2.325978	0.556138	0.688312
RAC2	0.254465	2.051176	0.026583	0.002027
RASGRP1	0.330709	2.106775	0.003025	0.003915
RASGRP3	0.221897	2.172930	0.032968	0.031101
SERPINI1	0.145195	2.207296	0.418429	0.097006
SH3TC1	0.185942	2.163782	0.124020	0.001381
TUBB6	0.152696	2.195290	0.550592	0.059595
UNC5B	0.314067	2.157342	0.031434	0.017828
UPP1	0.123905	2.124277	0.865388	0.007320
VEGFC	0.115821	2.118857	0.893674	0.005428
VIM	0.151533	2.231283	0.391336	0.270793
Water	0.235001	2.213517	0.002211	0.050700

B.2 Full edge and bulk cell displacement data

B.2.1 C8161 cell line

	Bulk displacements	Edge disp.	Bulk directionality	Edge dir.
ADM	2.213160	3.479522	0.053536	0.188612
ANXA1	2.237449	3.755544	0.066522	0.229271
ANXA6	2.151219	3.567129	0.062300	0.206856
ARPC3	2.066974	2.882849	0.055793	0.148703
ATP1B1	2.092258	3.565953	0.063256	0.232771
BMP4	1.914626	2.570568	0.054439	0.093814
CA2	2.277274	3.934356	0.064924	0.234497
CREB3L2	2.269566	3.148630	0.057502	0.153740
CTNNB1	2.230157	3.503308	0.050209	0.173016
DAB2	2.239577	3.898306	0.063295	0.220543
DKK3	1.806836	2.933276	0.090384	0.239571
DPYD	2.301159	3.415315	0.062463	0.177170
ESM1	2.317044	3.508237	0.059195	0.178569
FLT1	2.252542	3.621923	0.060720	0.200898
FN1	2.263700	3.317877	0.052999	0.161297
FOS	2.224881	3.492062	0.066497	0.202983
HSPB1	2.284621	3.435008	0.064839	0.169949
HSPB8	2.272366	3.382905	0.060725	0.186787

IFI30	1.536626	2.290209	0.084632	0.198326
IGFBP2	2.211297	3.396694	0.061496	0.191689
IL13RA1	1.752064	2.740147	0.104980	0.247370
INHBA	1.754219	2.703615	0.081153	0.218387
ITGB1	1.568666	2.218469	0.091399	0.200614
JUN	1.652961	2.998326	0.103846	0.304856
KCNE3	1.545553	3.022180	0.076553	0.313016
KITLG	1.791576	2.671687	0.099161	0.212507
KYNU	1.633417	3.098280	0.082156	0.292595
LY86	2.159272	3.535097	0.060593	0.224255
MMP11	1.747122	2.778403	0.093377	0.240755
NEXN	1.560756	2.604402	0.081627	0.242972
NRP1	1.640866	3.118727	0.082169	0.300113
Neg	1.985498	3.208801	0.076569	0.225969
PRKCQ	1.524028	2.463256	0.073847	0.239499
PRNP	2.052514	3.196175	0.060315	0.191083
RAC2	1.520675	2.720272	0.084558	0.309600
RASGRP1	1.672461	2.612561	0.086271	0.217729
RASGRP3	1.722242	2.563269	0.093178	0.206566
SERPINI1	1.656211	2.377090	0.093785	0.217711
SH3TC1	1.671004	2.440149	0.087264	0.216894
TUBB6	1.855111	3.104653	0.104222	0.270439
UNC5B	1.663960	2.479510	0.077427	0.220611
UPP1	1.723207	2.730316	0.089385	0.237897
VEGFC	1.857010	2.833884	0.107997	0.240654
VIM	1.608489	2.916338	0.084678	0.276951
Water	1.955226	2.795318	0.067862	0.192853

B.2.2 HT1080 cell line

	Bulk displacements	Edge disp.	Bulk directionality	Edge dir.
ADM	3.077594	3.625212	0.052131	0.075719
ANXA1	2.991515	3.842929	0.049224	0.088416
ANXA6	3.215406	3.660972	0.049790	0.063543
ARPC3	3.165426	3.446829	0.057253	0.059790
ATP1B1	3.111107	3.554215	0.052307	0.082830
BMP4	3.072032	3.679884	0.049727	0.080510
CA2	3.132244	3.576175	0.054925	0.078574
COL4A4	3.127955	3.648445	0.056670	0.075493
CREB3L2	3.087058	3.579496	0.048549	0.067244
CTNNB1	3.110491	3.787300	0.046975	0.085800
DAB2	3.071208	3.769821	0.051319	0.071348
DKK3	3.333250	3.918330	0.054459	0.076715
DPYD	3.161241	3.586303	0.051736	0.061839
ESM1	3.184269	3.634473	0.049206	0.067691
FLT1	3.151249	3.613359	0.048276	0.072298
FN1	3.111713	3.472922	0.051557	0.063979

FOS	3.149730	3.587505	0.054174	0.068700
HSPB1	3.119123	3.773605	0.052543	0.076259
HSPB8	3.148571	3.736789	0.055092	0.070926
IFI30	3.307869	4.051353	0.044213	0.069464
IGFBP2	3.133281	3.594164	0.053990	0.070457
IL13RA1	3.316263	3.854580	0.058512	0.074989
INHBA	3.324967	4.054065	0.050346	0.078083
ITGB1	3.382761	3.816472	0.047661	0.060337
JUN	3.380446	3.906758	0.056787	0.082177
KCNE3	3.245116	3.999454	0.052198	0.095414
KITLG	3.296447	3.928556	0.051991	0.077898
KYNU	3.285835	3.927088	0.055895	0.079770
LY86	3.093963	3.671370	0.055592	0.076653
MMP11	3.319815	3.833382	0.052248	0.084427
NEXN	3.297336	3.941213	0.054113	0.080176
NRP1	3.270955	3.933900	0.048340	0.078768
Neg	3.198601	3.766283	0.051651	0.079405
PRKCQ	3.248695	3.846235	0.051361	0.087511
PRNP	3.103201	3.706341	0.055567	0.075689
RAC2	3.268873	3.968348	0.056137	0.086945
RASGRP1	3.384100	3.937529	0.053101	0.066721
RASGRP3	3.342151	3.886945	0.046353	0.071334
SERPINI1	3.193579	4.074218	0.056616	0.089246
SH3TC1	3.370471	3.797619	0.051829	0.065760
TUBB6	3.391689	3.783172	0.059456	0.066107
UNC5B	3.367556	3.497217	0.049962	0.047063
UPP1	3.410255	3.950039	0.057334	0.096509
VEGFC	3.395784	3.982384	0.058004	0.080005
VIM	3.309995	3.851563	0.054315	0.073615
Water	3.175879	3.729353	0.053979	0.078377

Bibliography

[1] Abraham E Wolf, Matthew A Heinrich, Isaac B Breinyn, Tom J Zajdel, and Daniel J Cohen. Short-term bioelectric stimulation of collective cell migration in tissues reprograms long-term supracellular dynamics. *PNAS Nexus*, 1(1), 3 2022.

[2] Juan Carlos Duhart, Travis T. Parsons, and Laurel A. Raftery. The repertoire of epithelial morphogenesis on display: Progressive elaboration of Drosophila egg structure. *Mechanisms of Development*, 148:18–39, 12 2017.

[3] Li Li, Yong He, Min Zhao, and Jianxin Jiang. Collective cell migration: Implications for wound healing and cancer invasion. *Burns and Trauma*, 1(1):21–26, 6 2013.

[4] Shreyansh Jain, Victoire M.L. Cachoux, Gautham H.N.S. Narayana, Simon de Beco, Joseph D'Alessandro, Victor Cellerin, Tianchi Chen, Mélina L. Heuzé, Philippe Marcq, René Marc Mège, Alexandre J. Kabla, Chwee Teck Lim, and Benoit Ladoux. The role of single-cell mechanical behaviour and polarity in driving collective cell migration. *Nature Physics*, 16(7):802–809, 7 2020.

[5] Michelangelo Von Dassow and Lance A. Davidson. Variation and robustness of the mechanics of gastrulation: The role of tissue mechanical properties during morphogenesis. *Birth Defects Research Part C - Embryo Today: Reviews*, 81(4):253–269, 12 2007.

[6] Mahmood S Choudhery. *Regenerative medicine*. IntechOpen, London, United Kingdom, 2020.

[7] Evanthia T. Roussos, John S. Condeelis, and Antonia Patsialou. Chemotaxis in cancer. *Nature Reviews Cancer*, 11(8):573–587, 8 2011.

[8] Alex Mogilner, Roy Wollman, and Wallace F. Marshall. Quantitative Modeling in Cell Biology: What Is It Good for? *Developmental Cell*, 11(3):279–287, 9 2006.

[9] M Sunnåker, A Giovanni Busetto, E Numminen, J Corander, M Foll, and C Dessimoz. Approximate Bayesian Computation. *PLoS Comp. Biol.*, 9(1):1–10, 2013.

[10] T Toni and M P H Stumpf. Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics*, 26(1):104–110, 2010.

[11] J Liepe, P Kirk, S Filippi, T Toni, C P Barnes, and M P H Stumpf. A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. *Nat. Protocols*, 9(2):439–456, 2014.

[12] Suzanne Lenhart and John T. Workman. Optimal Control Applied to Biological Models. Technical report, Taylor & Francis Group, LLC, Boca Raton, 2007.

[13] Matthew J. Simpson, Alexander P. Browning, David J. Warne, Oliver J. Maclaren, and Ruth E. Baker. Parameter identifiability and model selection for sigmoid population growth models. *Journal of Theoretical Biology*, 535, 2 2022.

[14] Andreas Raue, Clemens Kreutz, Fabian Joachim Theis, and Jens Timmer. Joining forces of Bayesian and frequentist methodology: A study for inference in the presence of non-identifiability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984), 2 2013.

[15] Oana Teodora Chis, Julio R. Banga, and Eva Balsa-Canto. Structural identifiability of systems biology models: A critical comparison of methods. *PLoS ONE*, 6(11), 11 2011.

[16] Marisa C. Eisenberg and Michael A.L. Hayashi. Determining identifiable parameter combinations using subset profiling. *Mathematical Biosciences*, 256:116–126, 10 2014.

[17] Nicolette Meshkat, Seth Sullivant, and Marisa Eisenberg. Identifiability Results for Several Classes of Linear Compartment Models. *Bulletin of Mathematical Biology*, 77(8):1620–1651, 9 2015.

[18] Yudi. Pawitan. *In all likelihood : statistical modelling and inference using likelihood*. Clarendon Press, Oxford ;, 2001.

[19] D R (David Roxbee) Cox. *Principles of statistical inference*. Cambridge core. Cambridge University Press, Cambridge ;, 2006.

[20] D A (David Arthur) Sprott. *Statistical inference in science*. Springer series in statistics. Springer, New York, 2000.

[21] Yunzhang Zhu, Xiaotong Shen, and Wei Pan. On High-Dimensional Constrained Maximum Likelihood Inference. *Journal of the American Statistical Association*, 115(529):217–230, 1 2020.

[22] Simon Martina-Perez, Matthew J. Simpson, and Ruth E. Baker. Bayesian uncertainty quantification for data-driven equation learning. *Proceedings of the Royal Society B: Biological Sciences*, 477(2254), 10 2021.

[23] Simon Martina-Perez, Heba Sailem, and Ruth E. Baker. Efficient Bayesian inference for mechanistic modelling with high-throughput data. *PLoS Computational Biology*, 18(6), 6 2022.

[24] Matthew A. Nunes and David J. Balding. On optimal selection of summary statistics for approximate bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010.

[25] Paul Fearnhead and Dennis Prangle. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *J. R. Statist. Soc. B*, 74:419–474, 2012.

[26] Simon Aeschbacher, Mark A. Beaumont, and Andreas Futschik. A novel approach for choosing summary statistics in approximate Bayesian computation. *Genetics*, 192(3):1027–1047, 11 2012.

[27] Aki Vehtarh, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul Christian Burkner. Rank-Normalization, Folding, and Localization: An Improved (Formula presented) for Assessing Convergence of MCMC (with Discussion)*†. *Bayesian Analysis*, 16(2):667–718, 2021.

[28] B M de Silva, D M Higdon, S L Brunton, and J N Kutz. Discovery of Physics From Data: Universal Laws and Discrepancies. *Front. Artificial Intelligence*, 3:25, 2020.

[29] John H. Lagergren, John T. Nardini, Ruth E. Baker, Matthew J. Simpson, and Kevin B. Flores. Biologically-informed neural networks guide mechanistic modeling from sparse experimental data. *PLoS Computational Biology*, 16(11), 12 2020.

[30] M Raissi, P Perdikaris, and G E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comp. Phys.*, 378:686–707, 2019.

[31] J H Lagergren, J T Nardini, G M. Lavigne, E M Rutter, and K B Flores. Learning partial differential equations for biological transport models from noisy spatio-temporal data. *Proc. Roy. Soc. A*, 476(2234):20190800, 2020.

[32] S H Rudy, S L Brunton, J L Proctor, and J N Kutz. Data-driven discovery of partial differential equations. *Science Adv.*, 3:4, 2017.

[33] O Eriksson, A Jauhiainen, S Maad Sasane, A Kramer, A G Nair, C Sartorius, and J Hellgren Kotaleski. Uncertainty quantification, propagation and characterization by Bayesian analysis combined with global sensitivity analysis applied to dynamical intracellular pathway models. *Bioinformatics*, 35(2):284–292, 2019.

[34] L Arriola and J M Hyman. Sensitivity Analysis for Uncertainty Quantification in Mathematical Models. In *Mathematical and Statistical Estimation Approaches in Epidemiology*, pages 195–247. Springer Netherlands, Dordrecht, 2009.

[35] P D W Kirk, A C Babtie, and M P H Stumpf. Systems biology (un)certainties. *Science*, 350(6259):386, 2015.

[36] M Komorowski, M J Costa, D A Rand, and M P H Stumpf. Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. *Proc. Natl. Acad. Sci. USA*, 2011.

[37] B Thijssen, T M H Dijkstra, T Heskes, and L F A Wessels. Bayesian data integration for quantifying the contribution of diverse measurements to parameter estimates. *Bioinformatics*, 34(5):803–811, 2018.

[38] Min Zhao, Bing Song, Jin Pu, Teiji Wada, Brian Reid, Guangping Tai, Fei Wang, Aihua Guo, Petr Walczysko, Yu Gu, and others. Electrical signals control wound healing through phosphatidylinositol-3-OH kinase-γ and PTEN. *Nature*, 442(7101):457–460, 2006.

[39] Daniel J. Cohen, W. James Nelson, and Michel M. Maharbiz. Galvanotactic control of collective cell migration in epithelial monolayers. *Nature Materials*, 13(4):409–417, 2014.

[40] Tom J. Zajdel, Gawoon Shim, Linus Wang, Alejandro Rossello-Martinez, and Daniel J. Cohen. SCHEEPDOG: Programming Electric Cues to Dynamically Herd Large-Scale Cell Migration. *Cell Systems*, 10(6):506–514, 6 2020.

[41] Julienne LaChance, Kevin Suh, Jens Clausen, and Daniel J. Cohen. Learning the rules of collective cell migration using deep attention networks. *PLoS Computational Biology*, 18(4), 4 2022.

[42] Francisco J.H. Heras, Francisco Romero-Ferrero, Robert C. Hinz, and Gonzalo G. De Polavieja. Deep attention networks reveal the rules of collective motion in zebrafish. *PLoS Computational Biology*, 15(9), 2019.

[43] Steven P Williams, Cathryn M Gould, Cameron J Nowell, Tara Karnezis, Marc G Achen, Kaylene J Simpson, and Steven A Stacker. Systematic high-content genome-wide RNAi screens of endothelial cell migration and morphology. *Scientific Data*, 4(1), 2017.

[44] Suarez-Arnedo Alejandra and Felipe A N D Clavijo Camila A N D Arbeláez Pablo A N D Cruz Juan C A N D Muñoz-Camargo Carolina Torres Figueroa. An image J plugin for the high throughput image analysis of in vitro scratch wound healing assays. *PLOS ONE*, 15(7):1–14, 10 2020.

[45] Justin C Yarrow, Zachary E Perlman, Nicholas J Westwood, and Timothy J Mitchison. A high-throughput cell migration assay using scratch wound healing, a comparison of image-based readout methods. *BMC Biotechnology*, 4(1):21, 2004.

[46] Jan Hasenauer and Julio R Banga. Editorial overview: 'Mathematical modelling of high-throughput and high-content data'. *Current Opinion in Systems Biology*, 29:100405, 2022.

[47] Alejandro F Villaverde, Dilan Pathirana, Fabian Fröhlich, Jan Hasenauer, and Julio R Banga. A protocol for dynamic model calibration. *Briefings in Bioinformatics*, 23(1), 10 2021.

[48] Marko Järvenpää, Michael U Gutmann, Arijus Pleska, Aki Vehtari, and Pekka Marttinen. Efficient acquisition rules for Model-Based Approximate Bayesian computation. *Bayesian Analysis*, 14(2):595 – 622, 2019.

[49] Stuart Barber, Jochen Voss, and Mark Webster. The rate of convergence for approximate Bayesian computation. *Electronic Journal of Statistics*, 9(1):80 – 105, 2015.

[50] Scott A Sisson, Yanan Fan, and Mark A Beaumont. *Handbook of Approximate Bayesian Computation (1st ed.)*. Chapman and Hall/CRC, 2018.

[51] Thomas P Prescott and Ruth E Baker. Multifidelity Approximate Bayesian Computation. *SIAM/ASA J. Uncertainty Quantification*, 8(1):114–138, 2020.

[52] Sarah Filippi, Chris P Barnes, Julien Cornebise, and Michael P H Stumpf. On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo. *Statistical Applications in Genetics and Molecular Biology*, 12(1):87–107, 2013.

[53] Aden Forrow and Ruth E Baker. Measuring the accuracy of likelihood-free inference, 2021.

[54] Richard G. Everitt and Paulina A. Rowińska. Delayed Acceptance ABC-SMC. *Journal of Computational and Graphical Statistics*, 30(1):55–66, 2021.

[55] Praneeth Netrapalli. Stochastic Gradient Descent and Its Variants in Machine Learning. *Journal of the Indian Institute of Science*, 99(2):201–213, 2019.

[56] Léon Bottou. Large-Scale Machine Learning with Stochastic Gradient Descent. *Proc. of COMPSTAT*, 10 2010.

[57] Dominic Masters and Carlo Luschi. Revisiting Small Batch Training for Deep Neural Networks. *CoRR*, abs/1804.07612, 2018.

[58] Paul Stapor, Leonard Schmiester, Christoph Wierling, Simon Merkt, Dilan Pathirana, Bodo M H Lange, Daniel Weindl, and Jan Hasenauer. Mini-batch optimization enables training of ODE models on large-scale datasets. *Nature Communications*, 13(1):34, 2022.

[59] Daniel Seita, Xinlei Pan, Haoyu Chen, and John Canny. An Efficient Minibatch Acceptance Test for Metropolis-Hastings. In *International Joint Conferences on Artificial Intelligence Organization*, pages 5359–5363, 10 2018.

[60] Alexander P Browning, Wang Jin, Michael J Plank, and Matthew J Simpson. Identifying density-dependent interactions in collective cell behaviour. *Journal of The Royal Society Interface*, 17(165):20200143, 2020.

[61] Rachelle Binny, Alex James, and Michael Plank. Collective Cell Behaviour with Neighbour-Dependent Proliferation, Death and Directional Bias. *Bulletin of Mathematical Biology*, 78, 10 2016.

[62] Matthew J Simpson, Ruth E Baker, Sean T Vittadello, and Oliver J Maclaren. Practical parameter identifiability for spatio-temporal models of cell invasion. *Journal of The Royal Society Interface*, 17(164):20200055, 2020.

[63] Avelino Javer, Jens Rittscher, and Heba Sailem. DeepScratch: Single-Cell Based Topological Metrics of Scratch Wound Assay. *Computational and Structural Biotechnology Journal*, 18, 10 2020.

[64] Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[65] Xavier Didelot, Richard G Everitt, Adam M Johansen, and Daniel J Lawson. Likelihood-free estimation of model evidence. *Bayesian Analysis*, 6(1):49 – 76, 2011.

[66] T Toni, D Welch, N Strelkowa, A Ipsen, and M P H Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. Roy. Soc. Interface*, 6(31):187–202, 2009.

[67] Robert E Kass and Adrian E Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795, 6 1995.

[68] P Del Moral, A Doucet, and A Jasra. Sequential Monte Carlo samplers. *J. Roy. Stat. Soc. B*, 68(3):411–436, 2006.

[69] S A Sisson, Y Fan, and Mark M Tanaka. Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA*, 104(6):1760–1765, 2007.

[70] S Kullback and R A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86, 1951.

[71] T F Pak, R E Baker, and J M Pitt-Francis. Pakman: a modular, efficient and portable tool for approximate Bayesian inference. *J. Open Source Software*, 5(47):1716, 2020.

[72] Jaime Melendez, Matthew Grogg, and Yi Zheng. Signaling Role of Cdc42 in Regulating Mammalian Physiology. *Journal of Biological Chemistry*, 286(4):2375–2381, 2011.

[73] Y Shimoyama, G Tsujimoto, M Kitajima, and M Natori. Identification of three human type-II classic cadherins and frequent heterophilic interactions between different subclasses of type-II classic cadherins. *Biochemical Journal*, 349(1):159–167, 2000.

[74] Wang Jin, Kai-Yin Lo, Shih–En Chou, Scott W McCue, and Matthew J Simpson. The role of initial geometry in experimental models of wound closing. *Chemical Engineering Science*, 179:221–226, 2018.

[75] Rena Nainggolan, Resianta Perangin-angin, Emma Simarmata, and Astuti Feriani Tarigan. Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method. *Journal of Physics: Conference Series*, 1361(1):12015, 11 2019.

[76] Robert L Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.

[77] C Xiang, Lv Y, Wei Y, Wei J, Miao S, Mao X, Gu X, Song K, and Jia S. Effect of EphA7 Silencing on Proliferation, Invasion and Apoptosis in Human Laryngeal Cancer Cell Lines Hep-2 and AMC-HN-8. *Cell Physiol Biochem*, 36(2):435–445, 2015.

[78] J M Vicencio, C Ortiz, A Criollo, A W E Jones, O Kepp, L Galluzzi, N Joza, I Vitale, E Morselli, M Tailler, M Castedo, M C Maiuri, J Molgó, G Szabadkai, S Lavandero, and G Kroemer. The inositol 1,4,5-trisphosphate receptor regulates autophagy through its interaction with Beclin 1. *Cell Death & Differentiation*, 16(7):1006–1017, 2009.

[79] Rudolf Virchow. *La pathologie cellulaire basée sur l'etude physiologique et pathologique des tissus*. J.-B. Baillière et fils, Paris, 4e éd. / revue, c.. edition, 1874.

[80] Peter Friedl and Darren Gilmour. Collective cell migration in morphogenesis, regeneration and cancer. *Nature reviews Molecular cell biology*, 10(7):445–457, 2009.

[81] Ricard Alert and Xavier Trepat. Physical Models of Collective Cell Migration. *The Annual Review of Condensed Matter Physics is Annu. Rev. Condens. Matter Phys. 2020*, 11:77–101, 2020.

[82] Tambe Dhananjay T. and Ugo A N D Trepat Xavier A N D Park Chan Young A N D Kim Jae Hun A N D Millet Emil A N D Butler James P A N D Fredberg Jeffrey J Croutelle. Monolayer Stress Microscopy: Limitations, Artifacts, and Accuracy of Recovered Intercellular Stresses. *PLOS ONE*, 8(2):1–12, 10 2013.

[83] C Blanch-Mercader, R Vincent, E Bazellières, X Serra-Picamal, X Trepat, and J Casademunt. Effective viscosity and dynamics of spreading epithelia: a solvable model. *Soft Matter*, 13(6):1235–1243, 2017.

[84] Xavier Serra-Picamal, Vito Conte, Romaric Vincent, Ester Anon, Dhananjay T. Tambe, Elsa Bazellieres, James P. Butler, Jeffrey J. Fredberg, and Xavier Trepat. Mechanical waves during tissue expansion. *Nature Physics*, 8(8):628–634, 2012.

[85] Romaric Vincent, Elsa Bazellières, Carlos Pérez-González, Marina Uroz, Xavier Serra-Picamal, and Xavier Trepat. Active Tensile Modulus of an Epithelial Monolayer. *Physical Review Letters*, 115(24), 12 2015.

[86] Yougan Cheng, Bryan Felix, and Hans G. Othmer. The Roles of Signaling in Cytoskeletal Changes, Random Movement, Direction-Sensing and Polarization of Eukaryotic Cells, 6 2020.

[87] Jonathan Fouchard, Tom P J Wyatt, Amsha Proag, Ana Lisica, Nargess Khalil-gharibi, Pierre Recho, Magali Suzanne, Alexandre Kabla, and Guillaume Charras. Curling of epithelial monolayers reveals coupling between active bending and tissue tension. *Proceedings of the National Academy of Sciences*, 117(17):9377–9383, 4 2020.

[88] Elias H Barriga, Kristian Franze, Guillaume Charras, and Roberto Mayor. Tissue stiffening coordinates morphogenesis by triggering collective cell migration in vivo. *Nature*, 554(7693):523–527, 2018.

[89] Sebastian J Streichan, Christian R Hoerner, Tatjana Schneidt, Daniela Holzer, and Lars Hufnagel. Spatial constraints control cell proliferation in tissues. *Proceedings of the National Academy of Sciences*, 111(15):5586–5591, 2014.

[90] Gawoon Shim, Danelle Devenport, and Daniel J Cohen. Overriding native cell coordination enhances external programming of collective cell migration. *Proceedings of the National Academy of Sciences*, 118(29):e2101352118, 2021.

[91] Shao-Zhen Lin, Peng-Cheng Chen, Liu-Yuan Guan, Yue Shao, Yu-Kun Hao, Qunyang Li, Bo Li, David A Weitz, and Xi-Qiao Feng. Universal Statistical Laws for the Velocities of Collective Migrating Cells. *Advanced Biosystems*, 4(8):2000065, 2020.

[92] Aashrith Saraswathibhatla, Emmett E Galles, and Jacob Notbohm. Spatiotemporal force and motion in collective cell migration. *Scientific Data*, 7(1):197, 2020.

[93] Andreas Bauer, Magdalena Prechová, Lena Fischer, Ingo Thievessen, Martin Gregor, and Ben Fabry. pyTFM: A tool for traction force and monolayer stress microscopy. *PLoS Computational Biology*, 17(6), 6 2021.

[94] Stephen J DeCamp, Victor M K Tsuda, Jacopo Ferruzzi, Stephan A Koehler, John T Giblin, Darren Roblyer, Muhammad H Zaman, Scott T Weiss, Ayşe Klç, Margherita De Marzio, and others. Epithelial layer unjamming shifts energy metabolism toward glycolysis. *Scientific Reports*, 10(1):1–15, 2020.

[95] Matthew R Zanotelli, Aniqua Rahman-Zaman, Jacob A VanderBurgh, Paul V Taufalele, Aadhar Jain, David Erickson, Francois Bordeleau, and Cynthia A Reinhart-King. Energetic costs regulated by cell mechanics and confinement are predictive of migration path during decision-making. *Nature communications*, 10(1):4185, 2019.

[96] Lin Shao-Zhen, Wu-Yang Zhang, Dapeng Bi, Bo Li, and Xi-Qiao Feng. Energetics of mesoscale cell turbulence in two-dimensional monolayers. *Communications Physics*, 4(1):21, 2 2021.

[97] Xingbo Yang, Matthias Heinemann, Jonathon Howard, Greg Huber, Srividya Iyer-Biswas, Guillaume Le Treut, Michael Lynch, Kristi L. Montooth, Daniel J. Needleman, Simone Pigolotti, Jonathan Rodenfels, Pierre Ronceray, Sadasivan Shankar, Iman Tavassoly, Shashi Thutupalli, Denis V. Titov, Jin Wang, and Peter J. Foster. Physical bioenergetics: Energy fluxes, budgets, and constraints in cells, 6 2021.

[98] Johannes Schindelin, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, Curtis Rueden, Stephan Saalfeld, Benjamin Schmid, Jean-Yves Tinevez, Daniel James White, Volker Hartenstein, Kevin Eliceiri, Pavel Tomancak, and Albert Cardona. Fiji: an open-source platform for biological-image analysis. *Nature Methods*, 9(7):676–682, 2012.

[99] Gawoon Shim, Isaac B Breinyn, Alejandro Martinez-Calvo, Sameeksha, and Daniel J Cohen. Pump it up: bioelectric stimulation controls tissue shape and size. *bioRxiv*, pages 2011–2022, 2022.

[100] Tom J Zajdel, Gawoon Shim, and Daniel J Cohen. Come together: On-chip bioelectric wound closure. *Biosensors and Bioelectronics*, 192:113479, 2021.

[101] Julienne LaChance and Daniel J Cohen. Practical fluorescence reconstruction microscopy for large samples and low-magnification imaging. *PLoS computational biology*, 16(12):e1008443, 2020.

[102] William Thielicke and Eize Stamhuis. PIVlab–towards user-friendly, affordable and accurate digital particle image velocimetry in MATLAB. *Journal of open research software*, 2(1), 2014.

[103] Jean-Yves Tinevez, Nick Perry, Johannes Schindelin, Genevieve M Hoopes, Gregory D Reynolds, Emmanuel Laplantine, Sebastian Y Bednarek, Spencer L Shorte, and Kevin W Eliceiri. TrackMate: An open and extensible platform for single-particle tracking. *Methods*, 115:80–90, 2017.

[104] Ricard Alert, Carles Blanch-Mercader, and Jaume Casademunt. Active Fingering Instability in Tissue Spreading. *Phys. Rev. Lett.*, 122(8):88104, 3 2019.

[105] Kelly Vazquez, Aashrith Saraswathibhatla, and Jacob Notbohm. Effect of substrate stiffness on friction in collective cell migration. *Scientific Reports*, 12(1), 2022.

[106] Sham Tlili, Estelle Gauquelin, Brigitte Li, Olivier Cardoso, Benoît Ladoux, Hélène Delanoë Ayari, and François Graner. Collective cell migration without proliferation: Density determines cell velocity and wave velocity. *Royal Society Open Science*, 5(5), 5 2018.

[107] S. Tlili, M. Durande, C. Gay, B. Ladoux, F. Graner, and H. Delanoë-Ayari. Migrating Epithelial Monolayer Flows like a Maxwell Viscoelastic Liquid. *Physical Review Letters*, 125(8), 8 2020.

[108] Amalie Christensen, Ann-Katrine Vrans West, Lena Wullkopf, Janine Terra Erler, Lene Broeng Oddershede, and Joachim Mathiesen. Friction-limited cell motility in confluent monolayer tissue. *Physical Biology*, 15(6):66004, 7 2018.

[109] Chao Fang, Jiaxing Yao, Yuanjun Zhang, and Yuan Lin. Active chemo-mechanical feedbacks dictate the collective migration of cells on patterned surfaces. *Biophysical Journal*, 121(7):1266–1275, 2022.

[110] Julia C. Arciero, Qi Mi, Maria F. Branca, David J. Hackam, and David Swigon. Continuum model of collective cell migration in wound healing and colony expansion. *Biophysical Journal*, 100(3):535–543, 2 2011.

[111] C Pérez-González, R Alert, C Blanch-Mercader, M Gómez-González, T Kolodziej, E Bazellieres, J Casademunt, and X Trepat. Active wetting of epithelial tissues. *Nat. Phys.*, 15(1):79–88, 2019.

[112] Grégoire Peyret, Romain Mueller, Joseph d'Alessandro, Simon Begnaud, Philippe Marcq, René-Marc Mège, Julia M Yeomans, Amin Doostmohammadi, and Benoît Ladoux. Sustained Oscillations of Epithelial Cell Sheets. *Biophysical Journal*, 117(3):464–478, 2019.

[113] Claudia G Vasquez, Vipul T Vachharajani, Carlos Garzon-Coral, and Alexander R Dunn. Physical basis for the determination of lumen shape in a simple epithelium. *Nature Communications*, 12(1):1–12, 2021.

[114] Matthew A Heinrich, Ricard Alert, Julienne M LaChance, Tom J Zajdel, Andrej Košmrlj, and Daniel J Cohen. Size-dependent patterns of cell proliferation and migration in freely-expanding epithelia. *Elife*, 9:e58945, 2020.

[115] Adam Shellard, Peter Hampshire, Namid Stillman, Christina Dix, Richard Thorogate, Albane Imbert, Guillaume Charras, Ricard Alert, and Roberto Mayor. Frictiotaxis underlies adhesion-independent durotaxis. *bioRxiv*, pages 2023–2026, 2023.

[116] Adam Shellard and Roberto Mayor. Collective durotaxis along a self-generated stiffness gradient in vivo. *Nature*, 600(7890):690–694, 2021.

[117] Raimon Sunyer, Vito Conte, Jorge Escribano, Alberto Elosegui-Artola, Anna Labernadie, Léo Valon, Daniel Navajas, José Manuel Garc\'\ia-Aznar, José J Muñoz, Pere Roca-Cusachs, and others. Collective cell durotaxis emerges from long-range intercellular force transmission. *Science*, 353(6304):1157–1161, 2016.

[118] Barbara Cortese, Ilaria Elena Palamà, Stefania D'Amone, and Giuseppe Gigli. Influence of electrotaxis on cell behaviour. *Integrative Biology*, 6(9):817–830, 2014.

[119] Emil Du Bois-Reymond. *Untersuchungen über thierische Elektricität*, volume 2. G. reimer, 1849.

[120] Brian Reid and Min Zhao. The electrical response to injury: molecular mechanisms and wound healing. *Advances in wound care*, 3(2):184–201, 2014.

[121] Colin D McCaig, Ann M Rajnicek, Bing Song, and Min Zhao. Controlling cell behavior electrically: current views and future potential. *Physiological reviews*, 2005.

[122] The MathWorks Inc. MATLAB version: 9.13.0 (R2022b), 2022.

[123] Radek Erban and Hans G. Othmer. From signal transduction to spatial pattern formation in E. coli: A paradigm for multiscale modeling in biology. *Multiscale Modeling and Simulation*, 3(2):362–394, 2005.

[124] Greg M. Allen, Alex Mogilner, and Julie A. Theriot. Electrophoresis of cellular membrane components creates the directional cue guiding keratocyte galvanotaxis. *Current Biology*, 23(7):560–568, 4 2013.

[125] José Leal, Sebastian Shaner, Nicole Jedrusik, Anna Savelyeva, and Maria Asplund. Electrotaxis evokes directional separation of co-cultured keratinocytes and fibroblasts. *Scientific Reports*, 13(1):11444, 7 2023.

[126] Bing Song, Yu Gu, Wenkai Jiang, Ying Li, Wayne Nishio Ayre, Zhipeng Liu, Tao Yin, Christopher Janetopoulos, Miho Iijima, Peter Devreotes, and Min Zhao. Electric signals counterbalanced posterior vs anterior PTEN signaling in directed migration of Dictyostelium. *Cell and Bioscience*, 11(1), 12 2021.

[127] Runchi Gao, Siwei Zhao, Xupin Jiang, Yaohui Sun, Sanjun Zhao, Jing Gao, Jane Borleis, Stacey Willard, Ming Tang, Huaqing Cai, Yoichiro Kamimura, Yuesheng Huang, Jianxin Jiang, Zunxi Huang, Alex Mogilner, Tingrui Pan, Peter N. Devreotes, and Min Zhao. A large-scale screen reveals genes that mediate electrotaxis in Dictyostelium discoideum. *Science Signaling*, 8(378), 5 2015.

[128] Ifunanya Nwogbaga, A Hyun Kim, and Brian A. Camley. Physical limits on galvanotaxis, 2022.

[129] Michael Clerx, Martin Robinson, Ben Lambert, Chon Lok Lei, Sanmitra Ghosh, Gary R. Mirams, and David J. Gavaghan. Probabilistic Inference on Noisy Time Series (PINTS). *Journal of Open Research Software*, 7:1–6, 2019.

[130] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R.J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T. Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, and Yoshiki Vázquez-Baeza. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 3 2020.

[131] Yan Zhang, Guoqing Xu, Jiandong Wu, Rachel M Lee, Zijie Zhu, Yaohui Sun, Kan Zhu, Wolfgang Losert, Simon Liao, Gong Zhang, Tingrui Pan, Zhengping Xu, Francis Lin, and Min Zhao. Propagation dynamics of electrotactic motility in large epithelial cell sheets. *iScience*, 25(10):105136, 2022.

[132] Satomi Matsuoka and Masahiro Ueda. Mutual inhibition between PTEN and PIP3 generates bistability for polarity in motile cells. *Nature Communications*, 9(1), 12 2018.

[133] Changji Shi, Chuan Hsiang Huang, Peter N. Devreotes, and Pablo A. Iglesias. Interaction of Motility, Directional Sensing, and Polarity Modules Recreates the Behaviors of Chemotaxing Cells. *PLoS Computational Biology*, 9(7), 7 2013.

[134] Gillian DeWane, Alicia M. Salvi, and Kris A. DeMali. Fueling the cytoskeleton-links between cell metabolism and actin remodeling, 2 2021.

[135] J D (James Dickson) Murray. *Mathematical biology I : an introduction*. Interdisciplinary applied mathematics ; 17. Springer, New York, N.Y, 3rd ed. edition, 2002.

[136] C Falcó, D J Cohen, J A Carrillo, and R E Baker. Quantifying tissue growth, shape and collision via continuum models and Bayesian inference. *Journal of the Royal Society Interface*, 2023.

[137] Li Li, Robert Hartley, Bjoern Reiss, Yaohui Sun, Jin Pu, Dan Wu, Francis Lin, Trung Hoang, Soichiro Yamada, Jianxin Jiang, and others. E-cadherin plays an essential role in collective directional migration of large epithelial sheets. *Cellular and Molecular Life Sciences*, 69(16):2779–2789, 2012.

[138] Rafael Bailo, José A. Carrillo, Hideki Murakawa, and Markus Schmidtchen. Convergence of a fully discrete and energy-dissipating finite-volume scheme for aggregation-diffusion equations. *Mathematical Models and Methods in Applied Sciences*, 30(13):2487–2522, 12 2020.

[139] Oana-Teodora Chis, Alejandro F Villaverde, Julio R Banga, and Eva Balsa-Canto. On the relationship between sloppiness and identifiability. *Mathematical Biosciences*, 282:147–161, 2016.

[140] Nicole Le Douarin and Chaya Kalcheim. Source of the Pigment Cells. In *The Neural Crest*, pages 252–303. Cambridge University Press, United States, 1999.

[141] Jason A Morrison, Rebecca McLennan, Lauren A Wolfe, Madelaine M Gogol, Samuel Meier, Mary C McKinney, Jessica M Teddy, Laura Holmes, Craig L Semerad, Andrew C Box, Hua Li, Kathryn E Hall, Anoja G Perera, and Paul M Kulesa. Single-cell transcriptome analysis of avian neural crest migration reveals signatures of invasion and molecular transitions. *eLife*, 6:e28415, 2017.

[142] Jason A Morrison, Rebecca McLennan, Jessica M Teddy, Allison R Scott, Jennifer C Kasemeier-Kulesa, Madelaine M Gogol, and Paul M Kulesa. Single-cell reconstruction with spatial context of migrating neural crest cells and their microenvironments during vertebrate head and neck formation. *Development*, 148(22):dev199468, 11 2021.

[143] David Tatarakis, Zixuan Cang, Xiaojun Wu, Praveer P. Sharma, Matthew Karikomi, Adam L. MacLean, Qing Nie, and Thomas F. Schilling. Single-cell transcriptomic analysis of zebrafish cranial neural crest reveals spatiotemporal regulation of lineage decisions during development. *Cell Reports*, 37(12), 12 2021.

[144] Guillermo A. Vega-Lopez, Santiago Cerrizuela, Celeste Tribulo, and Manuel J. Aybar. Neurocristopathies: New insights 150 years after the neural crest discovery, 12 2018.

[145] Sharien Fitriasari and Paul A Trainor. Gene–environment interactions in the pathogenesis of common craniofacial anomalies. In Robert J Lipinski and Robert S Krauss, editors, *Current Topics in Developmental Biology*, volume 152, pages 139–168. Academic Press, 2023.

[146] Peter Friedl. Prespecification and plasticity: shifting mechanisms of cell migration. *Current Opinion in Cell Biology*, 16(1):14–23, 2004.

[147] Peter Friedl and Stephanie Alexander. Cancer invasion and the microenvironment: Plasticity and reciprocity, 11 2011.

[148] Paul M Kulesa, Jennifer C Kasemeier-Kulesa, Jessica M Teddy, Naira V Margaryan, Elisabeth A Seftor, Richard E B Seftor, and Mary J C Hendrix. Reprogramming metastatic melanoma cells to assume a neural crest cell-like phenotype in an embryonic microenvironment. *Proceedings of the National Academy of Sciences*, 103(10):3752–3757, 2006.

[149] Caleb M. Bailey, Jason A. Morrison, and Paul M. Kulesa. Melanoma revives an embryonic migration program to promote plasticity and invasion. *Pigment Cell and Melanoma Research*, 25(5):573–583, 9 2012.

[150] Caleb M. Bailey and Paul M. Kulesa. Dynamic interactions between cancer cells and the embryonic microenvironment regulate cell invasion and reveal EphB6 as a metastasis suppressor. *Molecular Cancer Research*, 12(9):1303–1313, 9 2014.

[151] Francesco Ginelli. The Physics of the Vicsek model, 11 2016.

[152] Stella C Ogbu, Phillip R Musich, Jinyu Zhang, Zhi Q Yao, Philip H Howe, and Yong Jiang. The role of disabled-2 (Dab2) in diseases. *Gene*, 769:145202, 2021.

[153] Mengyang Gu, Xinyi Fang, and Yimin Luo. Data-Driven Model Construction for Anisotropic Dynamics of Active Matter. *PRX Life*, 1(1), 8 2023.