

Parameterizing models of collective cell spreading



Andrew Parker

Wolfson College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity Term 2019

The highs and lows that have contributed to the work herein have been the most extreme I have experienced in my life. Through the highs, there has been immense gratification, and through the lows there has been difficult times that I could not have escaped without the care and kindness of my wife, Sarah. Without her, this work could not have been completed, and it is to her I predominantly dedicate this work.

These past five years have guided me out of the university way-of-life and into an immensely rewarding lifestyle, surrounded by friends and family. Without friends to help me forward, and without the reassurance of my parents, and my brother and his family, the lows would have been nearly unbearable. I am incredibly thankful that they are a part of my life.

My D.Phil. years are coming to an end far differently to how they began; I have a new niece, a new sister-in-law, a new place to call home, new friendships, new experiences, and most importantly, a stronger connection with my closest friend, whom I have relished living with, getting married to, and buying a house with. Thank you.

Acknowledgements

I would like to acknowledge the hard work and excellent support of my supervisors, Matthew J. Simpson and Ruth E. Baker, without whom this work would not have reached its conclusion. I would also like to thank the Systems Biology Doctoral Training Centre for an enjoyable first year experience and introduction to the D.Phil., through which I received a studentship from the UK's Engineering and Physical Sciences Research Council (EPSRC, EP/G03706X/1).

Statement of Originality

I hereby certify that this is entirely my own work unless otherwise stated.

Abstract

The study of how cell populations grow and spread is integral to understanding and predicting the invasion of cancer, the speed of wound repair and the robustness of embryonic development [20, 81, 96]. However, the extent to which cell populations grow and spread is governed by multiple processes, including motility, proliferation, adhesion and death, making it difficult to elucidate the relative contributions of these processes to the growth and invasion of a cell colony [106]. As such, *in vitro* cell biology assays are routinely used to probe the mechanisms by which cells interact, and the key processes involved in the growth and expansion of cell colonies. These *in vitro* assays generally involve seeding a population of cells on a two-dimensional substrate, and observing the population as the individual cells move and proliferate and the density of the monolayer increases towards confluence. A useful approach to interpret the results of these assays involves using a mathematical model that incorporates mechanistic descriptions of processes such as cell motility and proliferation. By parameterizing and validating the models using quantitative data from *in vitro* assays it is possible to provide quantitative insights into the mechanisms driving the growth and spreading of a cell population, and make experimentally testable predictions. However, it is not always clear how best to choose the experimental design, nor which summary statistics of the data to collect, in order to accurately and efficiently parameterize and validate models. The importance of various experimental design choices in arriving at testable predictions and outcomes are evaluated in this thesis, for a variety of models, cell types and cur-

rently employed experimental protocols. The use of inference is demonstrated for the first time in a previously developed force-based mathematical model, extended to account for proliferation.

Contents

1	Introduction	1
1.1	Experimental methods	3
1.2	Computational methods	5
1.3	Inference methods	9
1.4	Aims	11
1.5	Chronology of Thesis	13
2	The impact of experimental design on parameter inference	16
2.1	Introduction	16
2.1.1	Aims and outline	19
2.2	Methods	20
2.2.1	Mechanistic model	20
2.2.2	<i>In silico</i> data	21
2.2.3	Inference of model parameters	22
2.3	Results	33
2.3.1	Can we infer model parameters using a single summary statistic?	33
2.3.2	What is the impact of the initial number of cells on the posteriors?	37
2.3.3	Are trajectory statistics necessary for maximal information gain?	38
2.3.4	Can data-cloning ABC provides an efficient means to estimate model parameters?	41

2.4	Discussion	41
2.4.1	Predictions for experimental data	44
2.4.2	Conclusions	46
3	The influence of motility bias on parameter inference	48
3.1	Introduction	48
3.2	Methods	50
3.2.1	<i>In silico</i> data	50
3.2.2	Mechanistic model	52
3.2.3	Inference of model parameters	54
3.3	Results	60
3.3.1	Is it possible to infer the bias parameter?	60
3.3.2	What bias direction leads to more informative statistics? . . .	61
3.3.3	What initial configuration maximises information gain?	63
3.3.4	Inference using experimental data	68
3.4	Discussion	68
4	Bayesian inference applied to a force-based off-lattice model	72
4.1	Introduction	72
4.2	Methods	75
4.2.1	Model	75
4.2.2	Experimental data	78
4.2.3	Statistics	79
4.2.4	Inference - ABC-SMC	82
4.3	Results	88
4.3.1	Inference using synthetic data: experimental design	88
4.3.2	Inference using experimental data	100
4.4	Discussion	105

5	Coarse graining of an off-lattice model with proliferation	107
5.1	Introduction	107
5.2	Linear force law model	108
5.2.1	Numerical solution of node positions model	110
5.2.2	Continuum approximation	110
5.2.3	Derivation of cell density model	112
5.2.4	Comparing the accuracy of the continuum model with different boundary conditions.	116
5.3	General force law model	118
5.3.1	Continuum approximation	120
5.3.2	Derivation of cell density model	122
5.3.3	Comparing the accuracy of the continuum model with different force laws.	124
5.4	Introducing a proliferation mechanism	124
5.4.1	Proliferation mechanism	126
5.4.2	Derivation of cell density model with proliferation	128
5.4.3	Continuum approximation	132
5.4.4	Derivation of the corresponding cell density model	135
5.4.5	Evolution of cell number	136
5.4.6	Numerical solution	137
5.4.7	Comparing the cell density model with the discrete model . . .	139
5.5	Discussion	142
6	Conclusions and discussion	144
6.1	Conclusions	145
6.2	Extensions	146
6.2.1	Final remarks	147

Appendix A Consistency of results from Chapter 2	149
A.1 Results incorporating a larger spatial domain	149
A.2 Results for longer experiments	150
A.3 Results for different parameter regimes	150
A.4 Extracting statistics from experimental images	153
A.4.1 Manual extraction	153
A.4.2 Automated extraction	158
Appendix B Additional methods and results for Chapter 3	163
B.1 Replication of initial domains	163
B.2 Regression adjustment	165
B.2.1 Inference without regression adjustment on synthetic data . .	165
B.2.2 Inference with regression adjustment on experimental data . .	165
B.3 Different experimental domains or parameter regimes	169
B.3.1 Are parameters inferable with less or no bias?	169
B.3.2 If proliferation is included how do results change?	169
Appendix C Extraction of experimental data and further results from Chapter 4	173
C.1 Extracting data from experiments or simulations	173
C.1.1 Conversion from positions to density	174
C.1.2 Conversion from density to positions	174
C.1.3 Extracting statistics from experimental data	175
C.2 Other inference experiments	176
C.2.1 Combining statistics	176
C.2.2 Results using a hyperbolic proliferation law	176
Appendix D Numerical solution of the model in Chapter 5	180
D.1 Investigating the higher order term	180

D.2 Numerical solution of the full system	182
Bibliography	185

List of Figures

1.1	Examples of cell spreading experimental data.	4
2.1	Experimental images from the two experimental designs considered and their replication in the lattice model.	18
2.2	Two examples of the binning variance summary statistic.	27
2.3	Several examples of the gyration tensor summary statistic.	28
2.4	A schematic showing how regression adjustment works. Left: Exclude parameters that lead to statistics outside a tolerance (grey circles), $\delta = 0.25$, then (optionally) weight those within this tolerance according to some weighting function (e.g. an Epachnikov kernel [59]), and find the line of best fit (red-dashed line). Right, a larger version of the box on the left: Adjust parameter sets falling within the tolerance so it were as if those parameter sets produced a statistic exactly matching the experimental statistic (black circles are adjusted to red dots). . .	31
2.5	A selection of single summary statistic posteriors distributions the ex- perimental designs considered, and their information gains.	36
2.6	Posterior distributions and information gains demonstrating the im- pact of varying the initial cell number.	39
2.7	Posterior distributions and information gains demonstrating the im- pact of using multiple summary statistics.	40
2.8	Posterior distributions and parameter trajectories when using ABC-DC.	42

2.9	Posterior distributions and results of posterior predictive checks for the two cell types considered.	45
3.1	Experimental data at various times and its replication in the lattice model.	51
3.2	The impact of varying the bias parameter or initial cell number on the final positions of cells for a triangular shape.	53
3.3	The various initial tissue shapes that are considered of varying size and initial degree of confluence.	55
3.4	Posterior distributions and information gains for synthetically created data initialised in a large square.	62
3.5	Posterior distributions and information gains for synthetically created data initialised in a triangle with upwards or downwards bias.	63
3.6	The two-dimensional information gains for each initial density and initial shape, varying the shape radius.	65
3.7	The two-dimensional information gains for each initial shape and initial shape radius, varying the initial density.	66
3.8	The two-dimensional information gains for each initial density and initial shape radius, varying the initial shape.	67
3.9	Posterior distributions and information gains for experimental data initialised in a large square.	69
3.10	Delay in onset of bias in experimental data.	71
4.1	The method used by Jin et al. [36] to extract cell densities from images.	73
4.2	Schematic of the off-lattice model of Murray et al. [57].	74
4.3	Schematic detailing proliferation in the off-lattice model.	74
4.4	Schematic detailing relabelling after a proliferation event in the off-lattice model.	77

4.5	The resulting cell density data obtained by Jin et al. [36].	79
4.6	The method used to convert experimental cell densities into node positions.	81
4.7	Posterior distributions and information gains from ABC-SMC investigations with synthetic data computed using $Z = 2$, $T = 24$ hr.	90
4.8	Posterior distributions and information gains from ABC-SMC investigations with synthetic data computed using $Z = 4$, $T = 24$ hr.	91
4.9	Posterior distributions and information gains from ABC-SMC investigations with synthetic data computed using $Z = 2$, $T = 48$ hr.	93
4.10	Information gains from ABC-SMC investigations with synthetic data computed using $Z = 2$, $T = 24$ hr, varying the initial number of cells.	94
4.11	Information gains from ABC-SMC investigations with synthetic data computed using $Z = 2$, $T = 24$ hr, varying the initial density of cells.	96
4.12	Posterior distributions and information gains from ABC-SMC investigations with synthetic data, using a combination of experiments.	97
4.13	Information gains of the various ABC-SMC investigations, compared with the synthetic experiment with $Z = 2$, $T = 24$ hr.	99
4.14	One-dimensional posterior distributions and information gains from ABC-SMC investigations with experimental data.	102
4.15	Two-dimensional posterior distributions and information gains from ABC-SMC investigations with experimental data.	103
4.16	Posterior predictive checks for each experiment, using the posterior distribution computed from the cell density statistic.	106
5.1	Comparison of leading edge positions and cell densities between the ODE and the PDE solutions with different boundary conditions.	117
5.2	The force laws considered.	120

5.3	Comparisons of the leading edge positions and cell densities between ODE and PDE solutions when varying the force law.	125
5.4	A schematic of the proliferation mechanism introduced into the off-lattice model.	126
5.5	The proliferation laws considered.	127
5.6	Comparisons of the leading edge positions between stochastic simulations and PDE solutions for different force and proliferation laws. . .	140
5.7	Comparisons of the number of cells between stochastic simulations and PDE solutions for different force and proliferation laws.	141
A.1	Information gains for each summary statistic and initial scratch density in a larger domain, varying the size of scratch.	151
A.2	Information gains for each summary statistic and initial scratch density, initialising with 24 cells, varying the length of the simulation. . .	152
A.3	Information gains for each summary statistic and each initial number of cells, when initialising over six rows, varying the parameter ratio. .	154
A.4	Information gains for each summary statistic and each initial number of cells, when initialising over 24 rows, varying the parameter ratio. .	155
A.5	Information gains for each summary statistic and each initial scratch density, when initialising with 24 cells, varying the parameter ratio. .	156
A.6	Information gains for each summary statistic and each initial scratch density, when initialising with 48 cells, varying the parameter ratio. .	157
A.7	Method for manual segmentation of cell positions.	158
A.8	Method for automated segmentation of cell positions.	161
A.9	Method for automated tracking of cell positions.	162
A.10	Determining the ideal search radius for tracking cells through time. .	162

B.1	Posterior distributions and information gains for synthetically created data initialised in a large square.	166
B.2	Posterior distributions and information gains for experimental data initialised in a large square, when using regression adjustment.	167
B.3	Distances between experimental and synthetic experimental data and simulations from the model.	168
B.4	Posterior distributions resulting from synthetic experiments with varying bias parameters.	170
B.5	Posterior distributions and information gains from synthetic experiments when including both bias and proliferation in the model.	172
C.1	Posterior distributions and information gains when using either a one- or two- statistics, for synthetic data generated with $Z = 2$, $T = 24$	177
C.2	Posterior distributions and information gains from ABC-SMC investigations with synthetic data computed with hyperbolic proliferation.	179
D.1	The difference between leading edges, number of cells and cell densities when using the higher order proliferation terms in the PDE model.	181

Chapter 1

Introduction

Understanding cell spreading is fundamental to understanding cell biology. The invasion of cancer, the development of the embryo, and the healing of wounds are all controlled by the spread of cells, and gaining insight into these processes can only be achieved through an understanding of the mechanisms by which cells spread, namely motility, proliferation and interaction [20, 81, 96].

Cell spreading at its lowest level comprises of a complex series of regulations of proteins in the cell: signalling kinases, cytoskeleton adaptations and membrane receptors to name a few. In a simplified view of cell migration, a cell will probe its local environment using protrusions, and the cues it experiences will guide its movement toward a particular direction. Cell polarisation ensures the cell aligns itself along its chosen direction, before a complex rearrangement and realignment of the actin cytoskeleton ensures the front extends forward, and the rear retracts inwards, according to the adhesions the cell makes with its extracellular matrix [23, 49, 67, 94].

These intricate details of the motion of a single cell are often simplified and applied to a population of cells, as the phenomena we are interested in generally involve a collection of cells spreading together. Multi-cell spreading is further simplified by first considering the biological processes in a more controlled environment, for example

through *in vitro* experiments rather than *in vivo* experiments. By considering cells outside of an *in vivo* environment, it is generally possible to extract more details from experiments through the capturing of more detailed images, for example, or by the placement of cells into a plethora of situations that would be difficult or impossible in an *in vivo* setting, however, even in an *in vitro* setting, mechanisms driving the cell biology are still complex [24, 60, 106].

Teasing apart the behaviour that gives rise to observations collected from *in vitro* experiments is generally achieved by making assumptions about the underlying biology to allow further investigation. One such method to aid with this investigation is to establish computational or mathematical models that capture the phenomena under investigation. The development of these models is usually an iterative task, where complexity is slowly added, and existing models are adapted, rather than establishing a completely novel model, so that existing analysis can be utilised in the scope of the experimental data. Notably the models used vary greatly depending on the experiments being conducted and the processes of interest. In multicellular tissues, the focus of the model may be to understand the cell shape, the interactions between cells, or the rearrangements of touching cells. For a collection of cells that are not in direct contact, models instead may focus on the individual cell processes, such as movement and proliferation, or hows cells interact with each other [75, 103, 104, 109].

The choice of a particular computational or mathematical model should be approached through reference to both the underlying biology and the data that can be obtained from experiments, so that a model is chosen that is appropriate to the phenomena under consideration. Interdisciplinary discussions play a large role in the development of a model, as connecting models with experimental data is benefited from different perspectives; both in order to obtain relevant data, and in ensuring the model can capture the required effects [18]. It is this interdisciplinary approach that is perhaps strongest in systems biology, and particularly in the area of cell spreading

processes.

From the seminal works of Turing and Fisher through to the modern age, the advancements in understanding cell spreading have been a direct result of close collaborations between mathematicians, computational scientists and experimental biologists [28, 105]. Further, with the progression of technology, the availability of data from experiments has increased; in quality, quantity or through new measurement techniques [41]. Advances in imaging methods and in computational power mean we are gifted with more quantitative biological data, however it is still not well understood how best to utilise this additional data for beneficial usage.

In this thesis I establish methods connecting experimental data with computational models. The methods developed and the processes followed herein will help guide future experimental designs, demonstrate the mathematical approaches that allow us to gain insight into cell behaviours exhibited in experiments, and hence allow us all to gain a deeper understanding of cell spreading processes.

1.1 Experimental methods

In order to understand cell spreading processes, experiments are generally devised to observe cells moving either individually or as part of a larger population. Observations at a population-level consider a group of cells moving collectively or expanding into an empty region, and reveal the average behaviours of cells [22, 52]. Observations of individual cells, however, allow us to observe the interaction between these processes, for example the effect of crowding on proliferation and motility events [71, 110]. Only through linking individual-level processes with population-level behaviour can one account for effects that occur on these different spatial scales and arrive at predictions that are a good representation of cell behaviour [102].

Typically there are three types of *in vitro* experiments that are considered, and

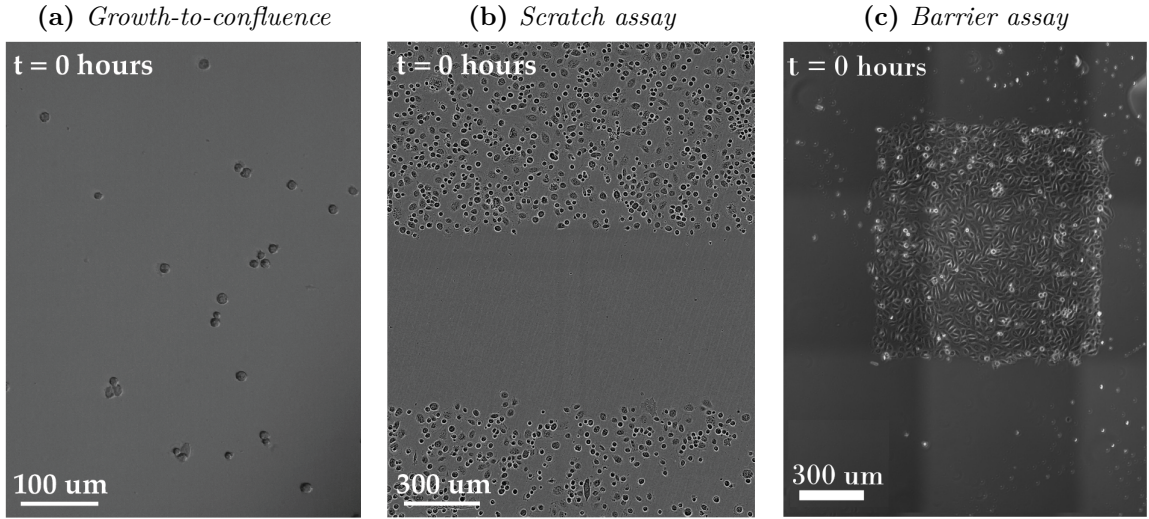


Figure 1.1: Examples of cell spreading experiments. Images shown are from the start of experiments. (a) A *growth-to-confluence* assay, (b) A *scratch assay*, (c) A *barrier assay*.

these are demonstrated in Figure 1.1. The first type is a *growth-to-confluence* assay, where a number of cells are initially seeded uniformly, and their spread is observed over a period of time, until they grow in number to occupy the two-dimensional domain under observation. These experiments allow quantification of how cell numbers increase over time, and how long the initially low density culture takes to reach confluence [10, 17].

A second experiment, a *scratch assay*, involves perturbing an assay of cells at confluence by “scratching off”, or removing, a region of the cell population, and observing the resulting spread of the remaining cells into this, now empty, region. These experiments reveal how quickly a population of cells can reach confluence from a new configuration, and how they adapt to a dramatic change in their local environment [55, 112]. The third type of experiment is a *barrier assay*, which involves keeping cells confined behind a barrier, before removing the barrier and capturing images of their resulting spread into the external region [91, 101, 107].

Unfortunately, each of these experiments have limitations as to the reproducibility of the initial domain - in a *growth-to-confluence* assay two assays are likely to be seeded differently, in a *scratch assay* two assays may be scratched differently by

different people, and in a *barrier assay* cells may cling to the barrier as it is extracted. Fortunately these problems can be alleviated to a large extent by using automated procedures, such as those involved in IncuCyte ZOOM[®] assays [41]. Consistency in these experiments is important so as to ensure that the predictions made from them are robust to differences in the initial conditions.

Though images of cell spreading experiments provide a means of quantifying cell spreading, they cannot be used alone to make predictions of how a group of cells will spread in unfamiliar circumstances, such as in a different experimental geometry. In experiments, there are many non-linear interacting processes being exhibited by cells, that could manifest at different temporal or spatial scales, depending on their local environments. Verbal models are therefore insufficient at describing the behaviours seen in experiments. To make accurate predictions of cell movements, the behaviour of cells in the cell spreading experiments above are used to inform our development of a computational or mathematical model, which can then be used to reliably predict cell behaviour.

1.2 Computational methods

Various modelling approaches have been used to capture the behaviour of cells in experiments. By systematically recreating complex biological processes in a tractable mathematical or computational framework, one can simplify the driving mechanisms into interpretable functional forms or model parameters, allowing us to gain insight into mechanisms driving cell spreading. In general, the processes that contribute most to cell spreading are motility and proliferation, so being able to accurately quantify these processes from experiments is vitally important in order to make meaningful predictions [5, 9, 90].

The behaviour of cells can be considered either individually or collectively. Many

experiments focus on observations of collections of cells, and so one can easily obtain information about how cell numbers change through time, or about the average motile behaviour of a cell population [17]. However, models are often formulated by considering cells at the level of the individual, where individual rules for processes can be prescribed, and observations are made for how these individual rules lead to population-level behaviours [86, 87]. Understanding cell processes in individual cells is therefore integral to formulating realistic population-level representations [66, 97, 110].

Individual-level models of motile cell populations generally speaking fall into two types, on- and off-lattice. On-lattice models consider cell positions located on a pre-defined lattice, whereas cell positions in off-lattice models are unrestricted. There are differences in these two approaches when it comes to the computational complexity of model simulations and how well they are perceived to represent biological experiments.

In on-lattice models, the cell diameter typically defines the lattice spacing, and volume exclusion is accounted for by allowing no two cells to occupy the same site [86]. In many such models, movement events and proliferation events into nearby lattice sites are specified to occur with certain probabilities at each discrete time-step, and adhesion can be accounted for by supposing, for example, a cell is less likely to move out of a site with more occupied neighbour sites [44, 76, 89]. In other models, cell movement events and proliferation events can be attempted into neighbour sites that are occupied through pushing mechanisms, or further away by using non-local jumping mechanisms [95, 110].

Although simple on-lattice models can display similar wave-front speeds to off-lattice models, crowding due to volume exclusion is often not well captured [65]. Cellular Potts models mitigate these issues by increasing the resolution of the model so a cell occupies multiple lattice sites simultaneously, allowing crowding and cell

deformations to be more realistically captured, though this approach can still lead to artefacts from the use of a lattice [25, 68]. To escape lattice-based artefacts, off-lattice models with volume exclusion can be implemented by allowing cells to occupy any space provided it does not impede on the space occupied by any other cell [26]. Spread of cells in these models are sometimes accounted for by allowing each cell to move or proliferate in discrete time-intervals in directions chosen randomly, and aggregation may be taken into account similarly to the on-lattice case [38, 44]. In off-lattice models, long-range interactions between cells are more difficult to account for, however long-range interactions are natural to introduce in force-based models.

Force-based models allow each cell to move according to the gradient of a force potential, representing repulsion (to prevent overlap) or attraction (for aggregation) between cells. These force laws can be chosen so that, for example, repulsion becomes stronger as cells move very close together, in order to maintain separation of cells [57]. There are many examples of potentials that can be used, for example the Morse potential or the Lennard-Jones potential [35, 56, 58]. However, choice of potential is usually made in an ad-hoc fashion and models are not rigorously validated against experimental data. A further issue with all models discussed is that the impact on model predictions of the particular model choices (rules for movement, proliferation, interactions e.t.c.) are not well understood, and choices are rarely validated against quantitative data. In order to overcome problems regarding validity, models at both the individual- and population-level should be tested, validated and refined using a range of quantitative experimental data that encompasses behaviours from both the individual and the population.

On- and off-lattice models help us gain understanding of cell-level processes but, in general, both suffer from a lack of analytical tractability. To mitigate this issue, coarse graining methods have been developed to elucidate population-level representations of these individual-level models. Comparison of population- and individual-level

behaviours is achieved by deriving population-level equations from individual-based models (IBMs), by considering changes in average cell densities (on-lattice) or probability density functions for the locations of each individual cell (off-lattice) [26, 88]. Comparisons between simulation and experiments can typically then be achieved at lower computational cost, as it is no longer necessary to run a computationally expensive simulation to determine the behaviour of a population of cells.

Establishing a population-level description of cell densities from an IBM is typically done through a mean-field approximation (MFA), which assumes that the positions of any two cells are independent of one another. This independence is known to fail in many cases, for example when significant proliferation takes place, or cellular movements are strongly biased in a particular direction [88]. To overcome issues with the MFA, it is possible to derive a series of moment dynamics equations describing the probabilities of positions of pairs, triplets or larger multiples of cells, accounting for more of the spatial information available. This system of equations must then be closed through a closure assumption, for example, the Kirkwood superposition approximation (KSA), which asserts that the positions of a triplet of cells can be written in terms of their pairwise positions [54]. The KSA has been used in coarse-graining both on- and off-lattice models, and it results in a marked increase in ability to accurately replicate the average behaviour of the IBM [56, 90].

Although these coarse graining methods provide an analytical representation of a model, in order to link models to experimental data it is necessary to be able replicate experimental data in the model, and this is achieved through parameter inference. Inference of model parameters allows predictions from derived models to resemble experiments, and further allows quantitative determination of the change in model parameters under different experimental conditions. Parameter inference also allows measurements of how the relative contributions of mechanisms accounted for in a model are impacted when there is a change in the experimental design.

1.3 Inference methods

Linking experiments and models has previously been achieved by minimising the distance between the data and an observable across parameter space. This method allows one to find a single likely estimate for each model parameter, for example, one can consider the wave-speed in a model and compare this to a recorded wave-speed in an experiment [14, 52]. However, when statistics are averaged in this way and a point estimate is obtained, information is lost about the variability of these observations, and what other parameter estimates may have led to similar observations. Further, understanding how parameters of the system depend on one-another, and hence some insight about the underlying mechanisms, is lost.

Integrating a model and experimental data is an effective means of achieving good model predictions; the inference of model parameters, therefore, is crucial to much of systems biology. Inference is approached generally in one of two ways, either through a frequentist approach or a Bayesian approach [2, 12]. In frequentist inference, one often seeks a point estimate of a parameter through maximum likelihood estimation, and captures uncertainty through confidence intervals for that estimate. A Bayesian approach instead derives a predictive posterior distribution based on the observed data. The predictive posterior distribution describes the probability distribution of model parameters, θ , giving rise to observed data, D . Obtaining the posterior distribution can be tackled by considering the model parameters themselves as random variables, and using Bayes rule to rewrite the posterior distribution, $p(\theta|D)$, as $p(\theta|D) \propto p(D|\theta)p(\theta)$, where $p(D|\theta)$ is the likelihood of obtaining the data from given model parameters, and $p(\theta)$ captures the prior knowledge of the parameters. Arriving at a posterior distribution therefore relies on obtaining the likelihood function. A posterior distribution can be obtained in a simple closed form if the likelihood function and prior are conjugate distributions, resulting in a posterior whose distribution can be easily calculated from data [32].

In general, however, posterior distributions will depend on many parameters, and a simple closed form for the joint posterior distribution will not exist. In these cases, it is sometimes possible to consider univariate conditional posterior distributions, which are the joint posterior distribution conditioned on all but one variable [31]. Bayes theorem allows sampling from these univariate conditional posterior distributions, as they are proportional to the joint posterior distribution: for example, let $p(\theta_1, \dots, \theta_K | x)$ be a K -dimensional posterior distribution, then if each univariate conditional posterior distribution, $p(\theta_j | \{\theta_i, i \neq j\}, x)$ has a closed form, one can sample from these conditional posterior distributions and explore parameter space proportionally to the posterior probability, using Markov chain Monte Carlo (MCMC) according to Gibbs sampling [83, 98].

In cases where either there is no efficient mechanism for simulating from the conditional posteriors, or these conditionals are not available in a closed form, one can instead use Metropolis-Hastings sampling. In Metropolis-Hastings sampling, the likelihood is approximated using a general MCMC method that performs a random walk in parameter space, and again accept samples proportionally to the posterior probability [7, 34].

These likelihood-based algorithms rely on establishing a Markov chain that converges to the posterior distribution, however requiring a likelihood to form a posterior can be bypassed by using approximate Bayesian computation (ABC). Simulating the model many times with parameters sampled from a prior distribution allows an approximate posterior distribution to be obtained by retaining those parameters that result in summary statistics closely matching experimental summary statistics. Although computationally more demanding than previous methods, ABC bypasses disadvantages associated with the use of a Markov chain, namely assessing chain convergence, and correlations between samples from the posterior distribution [92]. The use of ABC relies on several important user choices, most importantly the means of

comparison between simulation and experiment, typically through summary statistics obtained from experimental data [93, 100].

ABC has been used previously in the biological sciences, particularly to target complex problems in systems biology [45, 51], population genetics [8], and ecology [99]. There is also growing popularity in the use of ABC, specifically in the investigation of cell spreading processes [40, 77, 107]. There are many ABC variants that have arisen, mainly to try and alleviate the computational burden of the standard rejection ABC approach, from sequential Monte Carlo ABC (ABC-SMC), to approximate ABC (aABC) to name a few [16, 92].

Another important consideration when linking experimental data with computational models is that of identifiability of parameters. A parameter is identifiable if the confidence interval of its estimate is finite. Identifiability is separated into two types, either structural identifiability or practical identifiability. Structural non-identifiability arises from a redundancy of the model parameterization, whereas practical non-identifiability implies the amount or quality of experimental data is inadequate. Understanding practical identifiability is important to understand whether the experimental data acquired in experiments is appropriate for the models adopted [21, 72].

1.4 Aims

Throughout the thesis I consider how to apply mathematical and computational approaches to help strengthen the link between experiments and quantitative models. It is through the application of inference methods to parameterize models that further insights into biological processes can be gained. I will tackle the question of optimal experimental design: how should an experiment be designed to maximise the usefulness of a quantitative model, for a variety of data and models.

In Chapter 2 I explore how experimental design impacts our ability to infer parameters for simple models of the growth of proliferative and motile cell populations. A Bayesian approach is adopted, which allows characterisation of the uncertainty associated with estimates of the model parameters. A lattice-based volume-exclusion model is parameterized in a variety of experimental geometries, with cell types that span a range of possible model parameters. A multitude of summary statistics are considered, in order to establish how sensitive parameter inference is to the choice of summary statistics. I also investigate an alternative Bayesian method for performing parameter inference, namely approximate Bayesian computation data-cloning (ABC-DC), that iteratively converges onto maximum likelihood estimates (MLEs).

In Chapter 3, previously introduced inference methods are applied to novel experimental investigations that employ an electric field to influence the direction of the spread of cells. For these experimental data, it is important to quantify the effect of the electric field on the migration of cells, specifically to understand the changes in directing the motility of cells for applications in wound repair. I aim to determine the summary statistics and the experimental geometries that maximise the insights one can obtain from experiments, in order to direct future experimental choices to provide the most value.

In Chapter 4, I investigate the identifiability of parameters in a stochastic, off-lattice, one-dimensional model of cell spreading incorporating mechanisms of growth, contact and proliferation. I adapt the IBM from Murray et al. [57] to include a stochastic mechanism for proliferation, and infer model parameters in light of both synthetic data and experimental data. By considering a range of different potential experimental protocols, I find the experimental design choices that could be employed to maximise information gained. After inferring model parameters in light of experimental data, the model is also used to predict behaviour in experiments with different initial conditions through posterior predictive checks.

In Chapter 5, the previously established off-lattice one-dimensional model derived by Murray et al. [57] is extended to incorporate a mechanism for proliferation. Through the derivation of an equation for the cell density with a proliferation mechanism, further realism is added to the coarse grained model, and therefore makes it applicable to a wider range of experimental data. The equation established for the density of cells in the domain can be solved to determine average behaviours of cells that proliferate and interact with one another in a general length-dependent manner. Extending the coarse graining approach of Murray et al. [58], the accuracy of predictions from the density model are also improved, quantified by comparison to the individual-based model.

The thesis concludes with an overview of the results from each chapter, followed by a discussion of the potential future avenues that can be considered.

1.5 Chronology of Thesis

This section gives an overview of the chronological progression of the thesis, to give a better understanding to the reader of how the thesis came to be in its current form. At the outset, the aim of the thesis was to apply inference methods to a range of computational and mathematical models of cell spreading, with various experimental data. The first project considered was that of the two-dimensional lattice-based model, seen in Chapter 2, which is a commonly applied cell spreading model, and to which ABC inference had already been attempted elsewhere [40]. Verifying those results would give a good platform to jump from with new experimental data and different computational models. In the course of that study, there were various avenues for improvement of the methodology, e.g. considering different experimental geometries or considering novel summary statistics.

After initial investigations, it became clear that extracting quantitative informa-

tion from experimental images quickly would allow for more data to apply these data intensive approaches on, so several months were spent developing automated procedures using image processing algorithms. After these algorithms were established, more focus was given to the development of the inference procedures (as uniform prior distributions were leading to long simulation times) and the speeding up of simulations (through parallelisation). This led to an investigation into different inference methods and practices, such as ABC-SMC, data cloning ABC (ABC-DC) and regression adjustment. After implementing these alternative inference techniques, the work in Chapter 2 was written up with particular attention paid to the results using synthetic data, as the automated extraction of quantitative data from experimental images was still underdeveloped.

The next avenue considered was a different computational model with new experimental data, and I had a preference for an off-lattice model, as I believed it would provide a more realistic view of the cells interaction with its local neighbourhood and neighbours. A straightforward off-lattice model would enable more analytical tractability, so the spring-based model of Murray et al. [57] was chosen, with the intention to extend the model to include a stochastic proliferation mechanism. Unfortunately, reproducing the results from the original paper were difficult because of failed attempts at adequately capturing the boundary condition on the free boundary. Approaches attempted include using a NAG (numerical algorithms group) PDE solver and MATLAB in-built PDE solvers before a custom method was adopted. The mathematical formulation of a proliferation mechanism in this model was attempted, and the details of this work are in Chapter 5. Only after many iterations of these formulations was it deemed necessary to reformulate the problem as one PDE for the density of cells through time (on a moving domain), and one PDE for the moving domain through time. The numerical solver for these PDEs took a long time to implement, and solving these PDEs on a fine enough spatial and temporal mesh for

convergence also required a long time, so only after this was completed was it deemed that the PDE formulation was unsuitable for parameter inference.

After the details of the works in Chapters 2 and 5 were finalised, and further clarity was added to the first inference project, there was little time left to further develop and validate the automated image extraction methods, so additional experimental data (such as in Chapter 3) was manually extracted. With the computational models and inference methods already established, the focus of the final pieces of research were to apply each model to new and relevant experiment data as a proof of principle. The work in Chapter 3 involved applying inference methods to an extension of the model in Chapter 2 to experimental data where electric fields are used to bias cell movements. The work in Chapter 4 was an inference application of the model in Chapter 5, using scratch assay experiments that could be assumed one-dimensional, without motility bias, and where proliferation drove the expansion of cells.

The application of inference to a variety of experimental data and computational models, the original intent of the thesis, makes up the works in Chapters 2 to 4, whereas the work in Chapter 5, that appears to be quite a large deviation from the rest of the thesis, has been left until last.

Chapter 2

The impact of experimental design on parameter inference

The work in this chapter has been published [63]: Andrew Parker, Matthew J. Simpson, Ruth E. Baker, *The impact of experimental design choices on parameter inference for models of growing cell colonies*, Royal Society Open Science, doi: 10.1098/rsos.180384. Figures in this chapter are reproduced from this publication with kind permission.

2.1 Introduction

In this chapter we utilise a two-dimensional lattice-based exclusion process model that incorporates both motility and proliferation mechanisms. Our goal is to assess how our ability to accurately infer model parameters is affected by changes in the experimental design. Parameter inference is performed in a Bayesian framework using approximate Bayesian computation (ABC), which arrives at model parameters by simulating the model many times to find parameter values that lead to simulations that closely match experiments. ABC allows quantification of the uncertainty of the parameter estimates and bypasses the need to compute a likelihood function for the

mechanistic model. By quantifying the information gained of the model parameters using the different experimental protocols, we are able to provide guidelines for experimental design in terms of the selection of experimental geometry and the collection of relevant quantitative summary statistics from imaging data.

The experimental data we consider in this chapter are two of the main types of two-dimensional *in vitro* experiments, considered at the level of the population. The first experiment, shown in Figure 2.1(a), is a *growth-to-confluence assay*. These experiments observe a population of cells seeded approximately uniformly. The second experiment considered is shown in Figure 2.1(b), and is a *scratch assay*. It involves perturbing a population of cells at, or near, confluence by scraping away a region of the population. In this chapter we will explore the extent to which models of these two types of assays can be parameterized using various summary statistics of these experimental data.

For the mechanistic model used in this chapter, the likelihood is intractable and so we use ABC to generate an approximate posterior distribution. When using ABC in practice there are several important user choices, most importantly the means of comparison between simulated and experimental data, typically through summary statistics and a suitable distance metric, and the threshold used for accepting or rejecting parameter sets [93, 100]. In particular, the summary statistics represent lower dimensional descriptions of the data and their choice is vitally important as they impact the amount of information gained about model parameters through the use of ABC [6].

The major new insights provided in this chapter are a quantitative understanding of how both the choices of experimental geometry and summary statistics impact the quality of the posteriors generated using ABC. The use of ABC rejection allows comparing the posteriors resulting from a wide range of possible summary statistics in a computationally efficient way, and we quantify the information gain using the

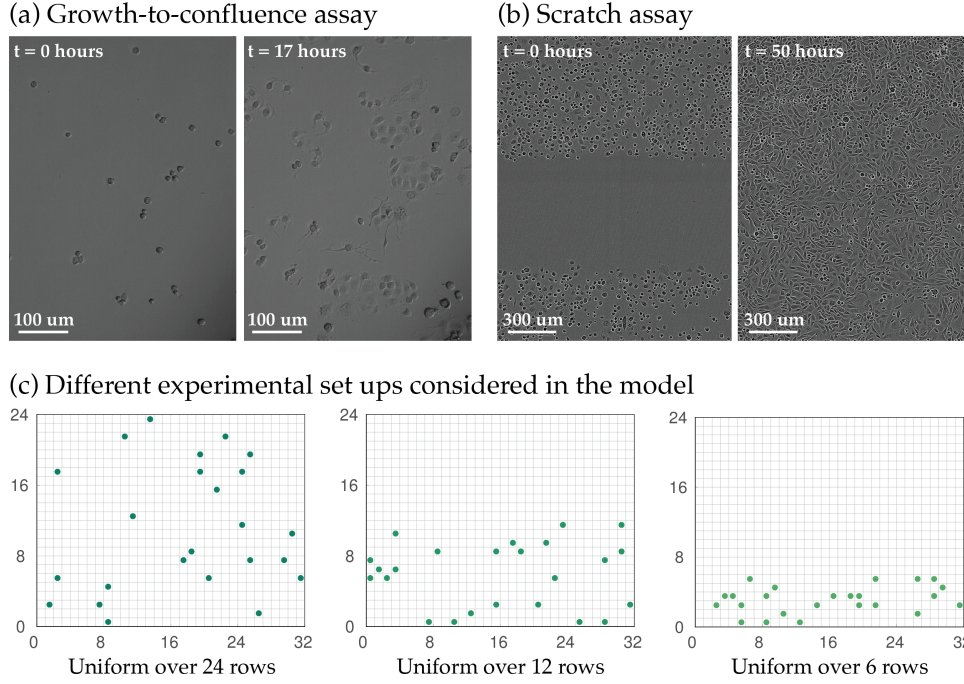


Figure 2.1: Two experimental designs considered in this chapter. (a) Images from a growth-to-confluence using MDA MB 231 breast cancer cells. See [90] for more information. (b) Images from a scratch assay with PC3 prostate cancer cells. See [41] for more information. (c) Schematics of the model. We use an on-lattice model, and vary the initial condition to replicate the experiments in (a),(b): we seed cells uniformly at random over 24 rows (left), 12 rows (middle) or six rows (right). The images in (a) are reproduced from Simpson et al. [90] with kind permission, whereas the images in (b) are reproduced from Johnston et al. [41] with kind permission.

Kullback-Leibler divergence [46]. We also demonstrate how ABC-DC [64] can potentially be used to obtain MLEs of model parameters more efficiently than ABC rejection. Instead of targeting the likelihood of the observed data, the data cloning approach targets the likelihood corresponding to a large number of copies (also known as clones) of the data, where each data clone is assumed independent of the others. Data cloning results in a posterior distribution that has the MLE as its mean, and the variance can be related to the asymptotic variance of the MLE [48]. ABC-DC uses ABC Markov chain Monte Carlo with data cloning to facilitate convergence towards a MLE [64].

2.1.1 Aims and outline

The aim of this chapter is to demonstrate the use of Bayesian inference methods to characterise the motile and proliferative behaviour of individual cells within growing cell colonies. In particular, we aim to estimate parameters of a lattice-based volume exclusion model for cell colony growth in a variety of experimental geometries. In addition, we demonstrate the use of ABC-DC for more efficient maximum likelihood estimation. In Section 2.2 we introduce the model, and the various ABC algorithms and summary statistics employed in this chapter. In Section 2.3 we assess how experimental design choices impact the quality of estimated posterior distributions, using both ABC rejection and ABC-DC, and we conclude with a discussion of the main results in Section 2.4. To confirm the accuracy of the inference process, and the relevance of the mechanistic model to the experimental data, we also sample from the posterior distributions and exploring whether key summary statistics predicted by the parameterized model, such as the size of the population and the displacement of individual cells, reflect the experimental observations within some reasonable confidence interval.

2.2 Methods

2.2.1 Mechanistic model

We employ a simple two-dimensional lattice-based exclusion process model akin to that of Simpson *et al.* [86], whereby $N(t)$ cells occupy a square lattice with R rows and C columns at time t . During each time step of duration τ , we choose $N(t)$ cells at random with replacement to attempt a movement or proliferation event into orthogonally adjacent lattice sites with probabilities P_m and P_p , respectively. For each cell we draw a uniform random number, $r \sim \mathcal{U}(0, 1)$. If $r \leq P_m$ the cell attempts a movement event into one of the four orthogonally adjacent lattice sites with equal probability, and if $P_m < r \leq P_m + P_p$ the cell attempts a proliferation event, whereby a daughter cell is placed into one of these lattice sites, each with equal probability. If $P_m + P_p < r \leq 1$ no movement or proliferation event is attempted. If a cell attempts to move or to place a daughter cell into an occupied lattice site, or outside of the domain, the attempted movement or proliferation event is aborted. To replicate experimental images, we use $R = 24$, $C = 32$, where lattice sites have length $\Delta = 18.75 \mu\text{m}$ (corresponding to the approximate cell diameter of the cells considered in typical experiments). Simulations are initialised with cell positions randomly distributed in the first $\hat{R}$ rows of the domain, where $\hat{R}$ is chosen to mimic potential experimental conditions. To interpolate between the *growth-to-confluence* and *scratch assay* designs, we choose three initial conditions (see Figure 2.1(c)) initialising cells uniformly at random across either 24, 12 or six rows of the domain. Initialisation over a smaller number of rows is intended to emulate a scratch assay where cells are initialised into a dense area of cells and will expand into an empty region of domain.

Note that, in reality, the size of the experimental domain is much larger than that captured in the experimental images [41]. To confirm that these results are insensitive to this choice of domain size, we repeated investigations with simulations carried out

on a larger domain and data collected on a subset of this domain that corresponds to the experimental image size. Results are shown in Appendix A.1.

2.2.2 *In silico* data

Since our aim in this chapter is to better understand how experimental design impacts upon our ability to infer model parameters, we first use the mechanistic model to generate *in silico* (observed) data that closely replicates that available from experiments (see Figure 2.1(a),(b)) and attempt to infer the parameters used to generate these *in silico* data. We use a time step of $\tau = 1/24$ hr, and model parameters $P_m = 0.25$ and $P_p = 0.0025$ (motivated by estimates for the diffusivity and proliferation rates from similar experiments [41], and applying the formula linking population-level diffusion and proliferation rates with individual-level parameters from Simpson et al. [88]). We record the final positions of cells in the *in silico* experiments after $T = 12$ hr, equating to 288 time steps. We also record trajectory data for five randomly chosen cells by recording their positions every eight time steps (corresponding to every 20 minutes [90]). As experiments are typically repeated several times to ensure reproducibility of results [90], we repeat simulations $M = 10$ times and average the resulting statistics. To confirm consistency of our results for larger values of the final simulation time, T , we repeat our methodology for longer periods of time, either $T = 24$ hr or $T = 36$ hr. Results are shown in Appendix A.2.

2.2.2.1 Model implementation

The mechanistic model is implemented with self-written code using `mex` functionality in MATLAB 2019a, where simulations are sped up using `Armadillo 9.600.6` [80], and the `parfor` function when performing inference.

Using a desktop Windows 10 PC with an Intel Core i5-3570 processor (3.40 GHz) and 16GB RAM, a simulation with $N(0) = 24$ cells initially, $P_m = 0.25$, $P_p = 0.0025$,

takes 0.0030 seconds to run. An set of simulations used in ABC rejection inference with prior distributions as in section 2.2.3, with $K = 10,000$ samples, and $M = 10$ replicates took 110.1 seconds to complete.

2.2.3 Inference of model parameters

We use ABC to estimate posterior distributions for the model parameters, $\theta = (P_m, P_p)$. ABC rejection is performed by repeatedly sampling from the prior distribution, $\pi(\theta)$, simulating the model, and accepting parameters that result in simulation output sufficiently close to the observed data. The accepted parameters are used to compute the approximate posterior distribution, $\mathbb{P}(\theta|\mathcal{D}^{obs})$. In order to assess how close the simulation and observed data are, we consider a range of summary statistics (detailed in Section 2.2.3.2) and an appropriate distance function (described in Section 2.2.3.3).

In order to quantify the performance of different summary statistics in inferring model parameters, we choose a uniform (uninformative) prior,

$$\pi(P_m) = \mathcal{U}(0, 0.99);$$

$$\pi(P_p) = \mathcal{U}(0, 0.01),$$

to ensure $P_m + P_p \leq 1$.

2.2.3.1 ABC rejection

Algorithm 1 describes the ABC rejection algorithm we use in this chapter. In Algorithm 1 we define the acceptance threshold, ϵ , as the 1st percentile of the distances, d^k , between the simulated and observed data, where the number of samples, $K = 20,000$, is sufficiently large to allow the 1st percentile of data to be a reasonable approximation to the true posterior. The accepted samples are used to estimate a posterior

distribution using bivariate kernel density estimation with a Gaussian kernel [13].

Algorithm 1 ABC rejection sampler.

```

1: for  $k = 1$  to  $K$  do
2:   sample  $\theta_k$  from  $\pi(\theta)$ 
3:   simulate  $\mathcal{D}^k$  from the model with parameter  $\theta_k$ 
4:   compute summary statistics,  $s_j^k(l)$ , for  $j = 1, \dots, 13$  and  $l = 1, \dots, L_j$  (see
      Section 2.2.3.3)
5:   calculate distance  $d^k$  (equations (2.15) and (2.16))
6: end for
7: calculate the acceptance threshold,  $\epsilon$  (see Section 2.2.3.1).
8: for  $k = 1$  to  $K$  do
9:   if  $d^k < \epsilon$  then
10:    accept  $\theta_k$ 
11:   else
12:    reject  $\theta_k$ 
13:   end if
14: end for

```

2.2.3.2 Summary statistics

In order to compare observed and simulated data, we reduce the dimension of the data using summary statistics. How best to choose summary statistics for parameter inference is an ongoing research question, and some automated procedures have been developed for this choice, typically involving minimising a loss of information function [1], or maximising the gain of information [6, 43], through a measure such as the Kullback-Leibler divergence [46].

The Kullback-Leibler divergence is a measure of the relative difference in two continuous probability distributions, F and G , and defined as

$$I_{\text{KL}}(F\|G) = \int_{\theta} F(\theta) \ln \left(\frac{F(\theta)}{G(\theta)} \right) d\theta. \quad (2.1)$$

In this chapter, we generate bivariate posteriors and prior distributions which are each functions of $\theta = (P_m, P_p)$. We discretise these distributions onto a fine mesh, 512×512 in size, for plotting, and compute the Kullback-Leibler divergence, I_{KL} , by

numerically integrating over both dimensions.

The observed data consists of the positions of all cells at the terminal time of the assay together with the tracks of five randomly chosen cells, $i = 1, \dots, 5$. Let $N(t_n)$ be the number of cells at time $t_n = \tau \times n$, i.e. at the n^{th} iteration of a simulation, so that $n = 288$ corresponds to the final simulation time, $T = 12$ h. We also let $\mathbf{X}_i(t_n) = (X_i(t_n), Y_i(t_n))$ be the (x, y) lattice coordinates of cell i at time t_n , and define,

$$\delta X_{ij}(t_n) = X_i(t_n) - X_j(t_n), \quad (2.2)$$

as the horizontal displacement between cell i and cell j at time t_n . We similarly define δY_{ij} . We summarise these data using statistics motivated by random walk models [6, 40, 90], considering thirteen statistics in total, labelled as in Table 2.1.

- The first summary statistic we consider is the final cell number, $N(T)$.
- In order to assess whether cells are forming clusters, the second and third summary statistics we consider relate to the size of the largest cluster of cells, computed using the **MATLAB** function `bwconncomp`. We consider either 4-connected clusters (cells orthogonally adjacent are part of a single cluster), $\kappa_4(T)$, or 8-connected clusters (cells diagonally adjacent are also part of a single cluster), $\kappa_8(T)$.
- Summary statistics four to six correspond to the binning variance [90], which quantifies the deviation from the average number of cells expected in quadrats of the domain, and is computed as

$$Q_w(T) = \sum_{b=1}^{B_w} \left(n(b, T) - \frac{N(T)}{B_w} \right)^2 \quad \text{for} \quad B_w = \frac{RC}{w^2} \quad \text{and} \quad w = 2, 4, 8, \quad (2.3)$$

where $n(b, T)$ is the number of cells in bin b at time T , and B_w the number of bins when the bin width is w . See Figure 2.2.

- As summary statistics seven and eight we consider trajectory statistics from the five tracked cells (labelled $i = 1, \dots, 5$): firstly, the total Manhattan displacement (the sum of horizontal and vertical distances travelled) of the cells, $\|x\|$, where

$$\|x\| = \frac{1}{5} \sum_{i=1}^5 \|x_i\| \quad \text{and} \quad \|x_i\| = \sum_{n=1}^{288/8} \left\| \mathbf{X}_i(t_{8n}) - \mathbf{X}_i(t_{8(n-1)}) \right\|_1, \quad (2.4)$$

recalling that there are 288 iterations, and cell positions are recorded every 8th iteration. Or, secondly, the tortuosity of the trajectory, Γ , where

$$\Gamma = \frac{1}{5} \sum_{i=1}^5 \frac{\|x_i\|}{\left\| \mathbf{X}_i(T) - \mathbf{X}_i(0) \right\|_1}. \quad (2.5)$$

- Summary statistic nine is the smallest eigenvalue, λ , of the gyration tensor [79],

$$G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}, \quad (2.6)$$

where

$$\begin{aligned} G_{11} &= \frac{1}{N(T)^2} \sum_{i=1}^{N(T)} \sum_{j=i+1}^{N(T)} \delta X_{ij}(T) \delta X_{ij}(T), \\ G_{12} &= \frac{1}{N(T)^2} \sum_{i=1}^{N(T)} \sum_{j=i+1}^{N(T)} \delta X_{ij}(T) \delta Y_{ij}(T), \\ G_{21} &= \frac{1}{N(T)^2} \sum_{i=1}^{N(T)} \sum_{j=i+1}^{N(T)} \delta X_{ij}(T) \delta Y_{ij}(T), \\ G_{22} &= \frac{1}{N(T)^2} \sum_{i=1}^{N(T)} \sum_{j=i+1}^{N(T)} \delta Y_{ij}(T) \delta Y_{ij}(T), \end{aligned} \quad (2.7)$$

with δX and δY defined as in equation (2.2). The smallest eigenvalue of G quantifies the spread of cells along the lesser of the two principal axes. See Figure 2.3.

- Summary statistics ten and eleven consider the distribution of pairs of cells, using the pairwise correlation functions, where the argument l indicates the number of pairs of cells separated by distance l . We consider pairwise correlations that measure only the vertical separation between cells, C_Y (due to the heterogeneity of initial condition in the y direction), or the total separation of cells, C_{XY} , where

$$C_Y(l) = \frac{1}{2} \sum_{i=1}^{N(T)} \sum_{j=1}^{N(T)} 1_{\{l\}} \left[\|\delta Y_{ij}(T)\| \right], \quad (2.8)$$

$$C_{XY}(l) = \frac{1}{2} \sum_{i=1}^{N(T)} \sum_{j=1}^{N(T)} 1_{\{l\}} \left[\|\delta X_{ij}(T)\| + \|\delta Y_{ij}(T)\| \right], \quad (2.9)$$

and $1_{\{l\}}$ is the indicator function which, in this case, counts pairs of cells separated by lattice distance l .

- Finally, we consider the correlation functions normalised by the expected number of pairs at each distance $l = 1, \dots, L$, which accounts for the density of cells in the domain,

$$\hat{C}_Y(l) = \frac{C_Y(l)}{q_Y(l)}, \quad \hat{C}_{XY}(l) = \frac{C_{XY}(l)}{q_{XY}(l)}, \quad (2.10)$$

where

$$q_Y(l) = C^2(R-l)\rho\bar{\rho}, \quad (2.11)$$

$$q_{XY}(l) = \left[C(R-l) + 2 \sum_{j=1}^{l-1} (C-j)(R-(l-j)) + (C-l)R \right] \rho\bar{\rho}, \quad (2.12)$$

$$\rho = \frac{N}{CR}, \quad \bar{\rho} = \frac{N-1}{CR-1}, \quad (2.13)$$

and C and R are the numbers of columns and rows in the lattice, respectively.

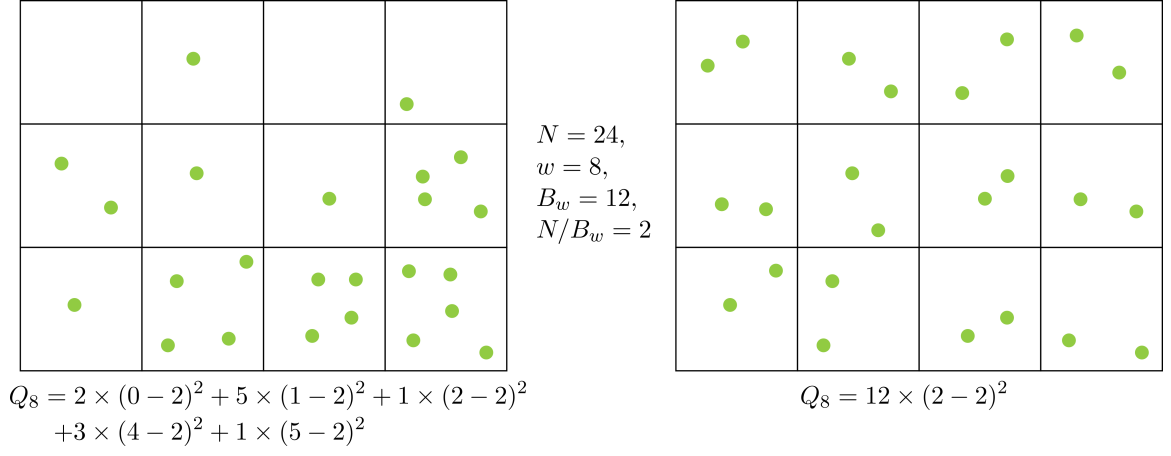


Figure 2.2: Two example distributions of twenty four cells in a domain with twelve bins. The binning variance statistic, Q_w quantifies the deviation from the average number of cells in each bin.

Summary statistic	Notation
Number of cells	N
Size of largest cluster, k -connected	$\kappa_k, k = 4, 8$
Binning variance, bin width = w	$Q_w, w = 2, 4, 8$
Manhattan displacement of cell i	$\ x\ $
Manhattan tortuosity of cell i	Γ
Smallest gyration tensor eigenvalue	λ
Correlation function	$C_{XY}(l)$
Normalised two-dimensional correlation function	$\hat{C}_{XY}(l)$
One-dimensional correlation function	$C_Y(l)$
Normalised one-dimensional correlation function	$\hat{C}_Y(l)$

Table 2.1: A list of the summary statistics and the corresponding notation adopted. Note, apart from cell trajectory statistics, all summary statistics are evaluated at the final time, T .

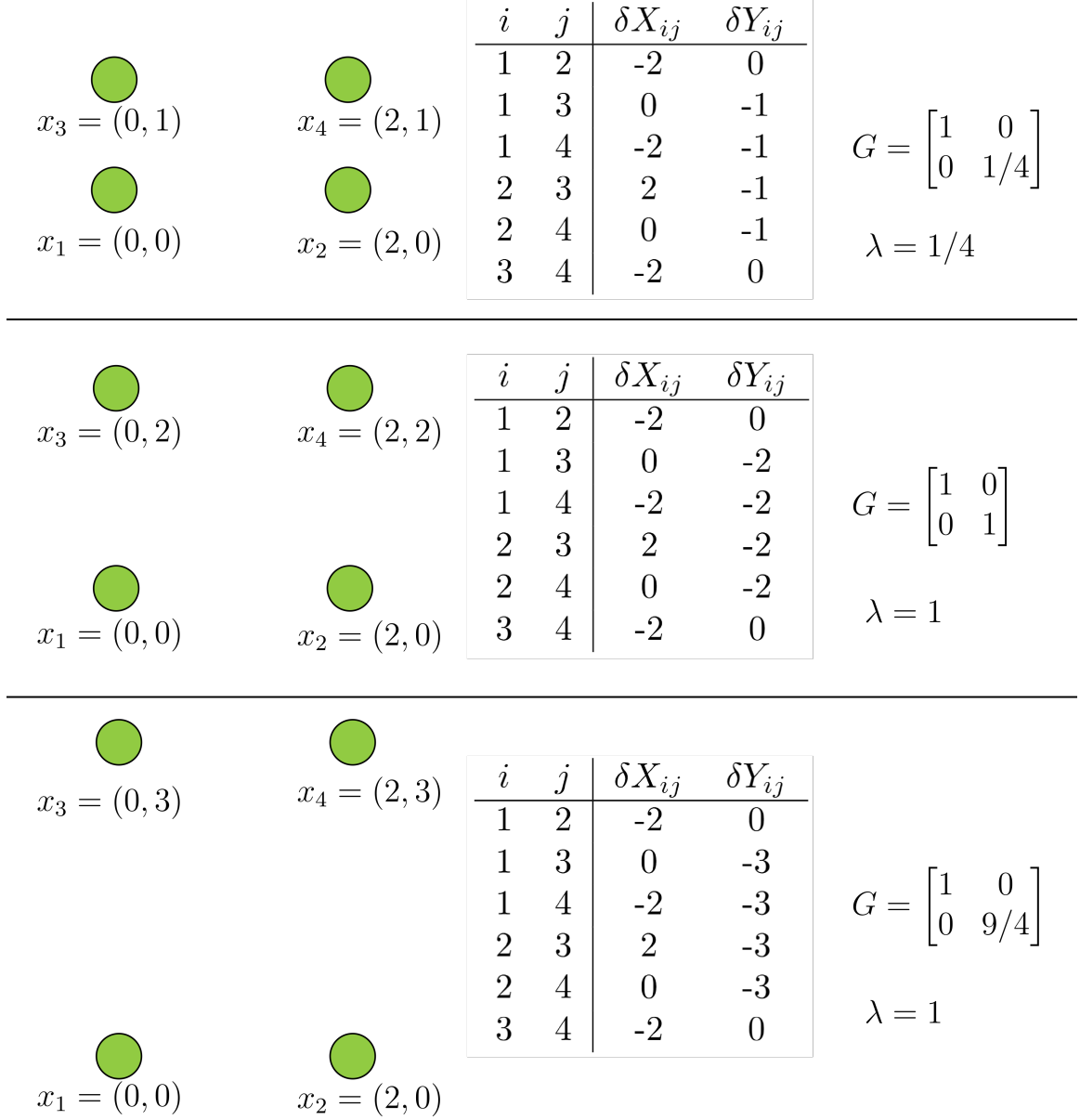


Figure 2.3: Three example distributions of four cells. The smallest eigenvalue of the gyration tensor quantifies the spread along the lesser of the two principal axes.

2.2.3.3 Computing distances

Experiments are typically performed multiple times under similar conditions, so we simulate multiple data sets, and call each set a replicate. The observed data, $\mathcal{D}^{obs}$, consists of M replicates, $\mathcal{D}^{obs} = \{\mathcal{D}_m^{obs} : m = 1, \dots, M\}$. The simulated data, $\mathcal{D}$, consists of M replicates and K samples, $\mathcal{D} = \{\mathcal{D}_m^k : m = 1, \dots, M; k = 1, \dots, K\}$. From the simulated data we compute each summary statistic, $s_j(l)$, where $j = 1, \dots, 13$ denotes the thirteen different summary statistics, and $l = 1, \dots, L_j$ is the l^{th} element of statistic ($L_j = 24$ for the pairwise correlation function statistics because there are 24 rows in the lattice, and $L_j = 1$ otherwise).

To calculate the distance, we firstly average each statistic over all replicates,

$$\bar{s}_j^k(l) = \frac{1}{M} \sum_{m=1}^M s_{j,m}^k(l) \quad \text{and} \quad \bar{s}_j^{obs}(l) = \frac{1}{M} \sum_{m=1}^M s_{j,m}^{obs}(l), \quad (2.14)$$

then we compute the median absolute deviation (MAD) for each statistic, $\sigma_j(l)$. This is defined for a univariate data set, $X = \{X_i\}$, as $\text{MAD} = \text{median}(|X_i - \text{median}(X)|)$. Then we compute the distances for each statistic, $j = 1, \dots, 13$, as

$$d_j^k = \sqrt{\frac{1}{L_j} \sum_{l=1}^{L_j} \left(\frac{\bar{s}_j^k(l) - \bar{s}_j^{obs}(l)}{\sigma_j(l)} \right)^2}. \quad (2.15)$$

Note that this choice of distance function weights vectors of summary statistics of different lengths equally. This is of particular importance when performing parameter inference utilising combinations of summary statistics, as in this chapter. We consider combinations of $A = 1, 2$ or 3 summary statistics from the thirteen statistics listed in Table 2.1 to give

$$d^k = \sqrt{\sum_{a=1}^A (d_a^k)^2}, \quad (2.16)$$

where the subscript a indexes the specific summary statistics used.

2.2.3.4 Regression adjustment

ABC rejection is an approximate method to arrive at the posterior distribution, accepting parameters that lead to simulations that closely match, but are not necessarily identical to, observed data. To account for distances that are non-zero, regression adjustment may be employed. Regression adjustment assumes that the relationship between parameter values and distances is linear near to $d = 0$. Any parameters that lead to non-zero distances can be adjusted using the regression line of best fit, to shift parameters according to where they would fall if they resulted in $d^k = 0$. Regression adjustment has particular benefit to ABC methods due to the inherent approximations involved in establishing a posterior distribution. It can both reduce the size of the confidence limits, and by adjusting parameters, one is more likely to obtain accurate parameter estimates [8, 11, 59, 64]. The method is not flawless however, as if any of the user choices for kernel width or threshold width for example are too large, parameters may be adjusted too far or too little. Any issues arising from using regression adjustment are seen clearly when comparing posteriors made with and without using the method.

The process is outlined mathematically below, or graphically in Figure 2.4. Assume that the vector of parameters, $\boldsymbol{\theta}$, satisfy the following regression model,

$$\boldsymbol{\theta}_k = \boldsymbol{\alpha} + (\mathbf{s}^k - \mathbf{s}^{obs})^T \boldsymbol{\beta} + \boldsymbol{\xi}_k, \quad k = I(1), \dots, I(B), \quad (2.17)$$

where the samples considered are all those within a given tolerance, δ , namely $I = \{k : d^k < \delta\}$, and B is the size of this set, $B = |I|$. In the above equation, $\mathbf{s}$ denotes the vector of summary statistics considered in the distance function (2.16) and the $\boldsymbol{\xi}_k$ are uncorrelated Gaussian random variables with zero mean and common variance σ^2 . The regression model is solved to find the least-squares estimates, $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$. Once

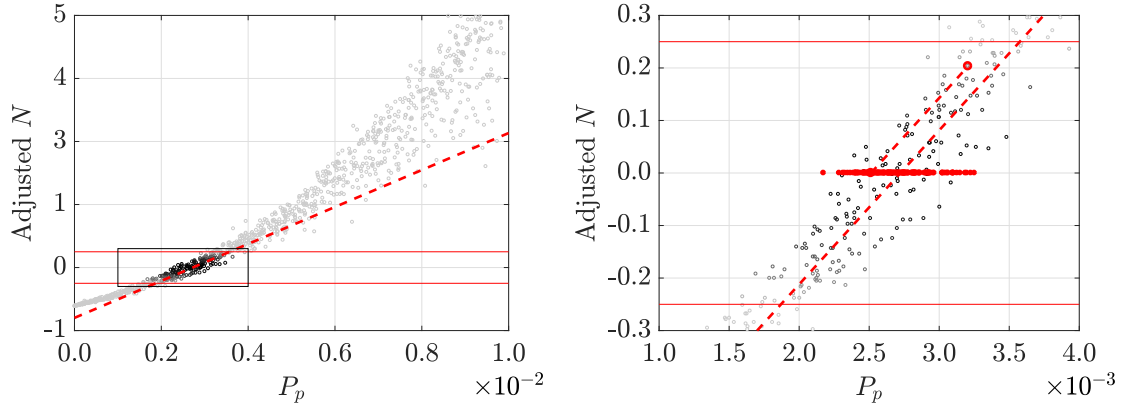


Figure 2.4: A schematic showing how regression adjustment works. Left: Exclude parameters that lead to statistics outside a tolerance ($\delta = 0.25$), then (optionally) weight those within this tolerance according to some weighting function (e.g. an Epachnikov kernel [59]), and find the line of best fit (red-dashed line). Right, a larger version of the box on the left: Adjust parameter sets falling within the tolerance so it were as if those parameter sets produced a statistic exactly matching the experimental statistic (black circles are adjusted to red dots).

these are found, the accepted parameter sets are adjusted according to

$$\theta_k^* = \theta_k - (\mathbf{s}^k - \mathbf{s}^{obs})^T \hat{\beta}, \quad k = I(1), \dots, I(B). \quad (2.18)$$

2.2.3.5 Data-cloning ABC

Data-cloning ABC involves considering a data-set, $\mathcal{D}_{K^{DC}}^{obs}$, containing K^{DC} clones of the experimental data, $\mathcal{D}^{obs}$, that is, $\mathcal{D}_{K^{DC}}^{obs} = (\mathcal{D}^{obs}, \mathcal{D}^{obs}, \dots, \mathcal{D}^{obs})$, where each clone is assumed independent of the others. The likelihood of the cloned data is then [48, 64],

$$\mathcal{L}(\mathcal{D}_{K^{DC}}^{obs} | \theta) = \left(\mathcal{L}(\mathcal{D}^{obs} | \theta) \right)^{K^{DC}}. \quad (2.19)$$

Hence the posterior distribution resulting from cloned data satisfies

$$\mathbb{P}(\theta | \mathcal{D}_{K^{DC}}^{obs}) \propto \mathcal{L}(\mathcal{D}_{K^{DC}}^{obs} | \theta) \pi(\theta) = \left(\mathcal{L}(\mathcal{D}^{obs} | \theta) \right)^{K^{DC}} \pi(\theta), \quad (2.20)$$

and the MLE of θ is equivalent to the mean of $\mathbb{P}(\theta|\mathcal{D}_{K^{DC}}^{obs})$ as $K^{DC} \rightarrow \infty$ [48]. Intuitively, it can be reasoned that if all of the independent model-generated data-sets are close to the experimental data, we are K^{DC} times as likely to have selected a sensible candidate parameter.

The ABC-DC algorithm was proposed by Picchini *et al.* [64], and it uses ABC Markov chain Monte Carlo [53]. The algorithm works in two stages, first the acceptance threshold, ϵ , is decreased according to a tolerance scheme $\{\epsilon_1, \epsilon_2, \dots, \epsilon_P\}$, and then the number of clones, K^{DC} , is increased according to a clones scheme $\{K^{DC}_{P+1}, K^{DC}_{P+2}, \dots, K^{DC}_{P+Q}\}$, where subscripts index the population number. The MLE is approximated by averaging the final parameter population.

For each sample, a proposal $\theta^\#$ is accepted or rejected according to the acceptance probability [64]

$$J_j^\epsilon(\mathcal{D}^\#, \mathcal{D}^{obs}) = \exp \left\{ -\frac{1}{2\epsilon^2} \left(\frac{1}{L_j} \sum_{l=1}^{L_j} \left(\frac{\bar{s}_j^\#(l) - \bar{s}_j^{obs}(l)}{\sigma_j(l)} \right)^2 \right) \right\} = \exp \left\{ -\frac{(d_j^\#)^2}{2\epsilon^2} \right\}, \quad (2.21)$$

which approximates a Gaussian centred on the experimental observation with variance ϵ , weighted according to the MAD, $\sigma_j(l)$, of the summary statistic j under consideration, where the MAD is estimated using ABC rejection.

The ABC-DC method is described in full in Algorithm 2. The number of samples in the decreasing tolerance stage is denoted by $\{r_1, \dots, r_P\}$, and $\{r_{P+1}, \dots, r_{P+Q}\}$ denotes the number of samples for the increasing clones stage. The algorithm is implemented using self-written code in **MATLAB**. The scheme we use has $P = 5$ and $Q = 4$, and other variables are as follows:

Number of samples, r = (1000, 1000, 1000, 1000, 1000, 1000, 1000, 1000, 2000);

Number of clones, K^{DC} = (1, 1, 1, 1, 1, 1, 2, 3, 4);

Value of tolerance, ϵ = (5, 2, 1, 0.5, 0.25, 0.25, 0.25, 0.25, 0.25).

2.3 Results

My aim in this chapter is to provide insights into how both the choices of experimental geometry and summary statistics impact the quality of the posteriors generated using ABC for experiments that are typically used to characterise growing and spreading cell populations. We generate *in silico* (observed) data using the model outlined in Section 2.2.1 and Section 2.2.2 for the range of experimental designs illustrated in Figure 2.1(c), and quantify the quality of the posteriors resulting from the use of ABC rejection with various summary statistics using the Kullback-Leibler divergence between the prior and posterior distributions (details of the methods used can be found in Section 2.2.3). In Section 2.3.1 and Section 2.3.2, we ask whether it is possible to quantify, using only a single summary statistic of the data, the relative contributions of proliferation and motility in a number of cell spreading experiments. In Section 2.3.3 we explore the extent to which parameter estimates can be improved using multiple summary statistics. Finally, in Section 2.3.4 we demonstrate the use of ABC-DC to efficiently find MLEs, and compare these estimates to those generated using ABC rejection.

2.3.1 Can we infer model parameters using a single summary statistic?

We first consider whether it is possible to infer model parameters, P_m and P_p , using a single summary statistic in the distance function (2.16). To do this, we perform ABC rejection on *in silico* (observed) data gathered from simulations with $P_m = 0.25$ and $P_p = 0.0025$, where we replicate different experimental designs by varying the model initial conditions: we initialise the model with 24 cells either in all 24 rows, in 12 rows, or in six rows, of the domain, as shown in Figure 2.1(c). Note that this means we change the average density within the region into which cells are initialised, but that

Algorithm 2 ABC-DC algorithm.

```

1: initialise  $\theta^* = (0.5, 0.005)$ 
2: for  $p = 1$  to  $P$  do
3:   set  $\theta_1 = \theta^*$ , and fix  $\epsilon = \epsilon_p$ ,  $K^{DC} = 1$ 
4:   simulate  $\mathcal{D}^*$  from the model with parameter  $\theta^*$ , and compute  $q^* = J_j^\epsilon(\mathcal{D}^*, \mathcal{D}^{obs})$ 
5:   for  $k = 1$  to  $r_p - 1$  do
6:     sample a new  $\theta^\# \sim N(\theta^*, (2.38)^2 \Sigma_p / 2)$ , where  $\Sigma_p$  is the covariance of the
       previous  $\theta$  population (or the prior if  $p = 1$ )a.
7:     simulate  $\mathcal{D}^\#$  from the model with parameter  $\theta^\#$ , and compute  $q^\# = J_j^\epsilon(\mathcal{D}^\#, \mathcal{D}^{obs})$ 
8:     compute  $\alpha = \min \left[ 1, \frac{q^\#}{q^*} \right]$  and generate  $\omega \sim U(0, 1)$ 
9:     if  $\omega > \alpha$  then
10:       set  $\theta_{k+1} := \theta_k$ 
11:     else
12:       set  $\theta_{k+1} := \theta^\#$ ,  $\theta^* = \theta^\#$  and  $q^* := q^\#$ .
13:     end if
14:     if  $p = P$  and  $\max \pi_P < q^\# \pi(\theta^\#)$  then
15:       update current maximum likelihood estimate: set  $\max \pi_{\epsilon_p} = q^\# \pi(\theta^\#)$ 
       and  $\theta_{MLE} = \theta^\#$ .
16:     end if
17:   end for
18: end for
19: perform regression adjustment on the final  $\theta$  population, set  $\Sigma_P$  as the covariance
   of the adjusted  $\theta$ .
20: for  $q = 1$  to  $Q$  do
21:   set  $\theta_1 = \theta^*$ , and fix  $\epsilon = \epsilon_P$ ,  $K^{DC} = K^{DC}_{P+q}$ 
22:   simulate  $K^{DC}$  independent data sets from the model,  $\mathcal{D}^{*(1)}, \dots, \mathcal{D}^{*(K^{DC})}$ , with
     parameter  $\theta^*$ , and compute  $q^* = \prod_{\kappa=1}^{K^{DC}} J_j^\epsilon(\mathcal{D}^{*(\kappa)}, \mathcal{D}^{obs})$ 
23:   for  $k = 1$  to  $r_{P+q} - 1$  do
24:     sample a new  $\theta^\# \sim N(\theta_{MLE}, \Sigma_{P+q-1})$ 
25:     simulate  $K^{DC}$  independent data sets from the model,  $\mathcal{D}^{\#(1)}, \dots, \mathcal{D}^{\#(K^{DC})}$ ,
       with parameter  $\theta^\#$ , and compute  $q^\# = \prod_{\kappa=1}^{K^{DC}} J_j^\epsilon(\mathcal{D}^{\#(\kappa)}, \mathcal{D}^{obs})$ 
26:     compute  $\alpha = \min \left[ 1, \frac{q^\#}{q^*} \right]$  and generate  $\omega \sim U(0, 1)$ 
27:     if  $\omega > \alpha$  then
28:       set  $\theta_{k+1} := \theta_k$ 
29:     else
30:       set  $\theta_{k+1} := \theta^\#$ ,  $\theta^* = \theta^\#$  and  $q^* := q^\#$ .
31:     end if
32:   end for
33:   update  $\Sigma_{P+q}$  as the covariance of the current  $\theta$  population
34: end for

```

^aThe factor $2.38^2/d$ is optimal for MCMC chain exploration [32].

the overall cell number, and hence the average total density, is constant. Figure 2.5(a)-(f) shows the posteriors generated for selected summary statistics of the data. The average information gain, and its deviation, for each of the three experimental designs and all summary statistics, is summarised in Figure 2.5(g).

Figure 2.5(a)-(b) shows that for a growth-to-confluence assay design (where cells are initialised across all 24 rows of the domain), an accurate estimate of P_p can be obtained using the number of cells, N , as a summary statistic, and an accurate estimate of P_m can be obtained using the Manhattan displacement, $\|x\|$, as a summary statistic.

Note that single-valued summary statistics such as the number of cells, N , will be unable to jointly infer P_m and P_p as a single constraint (in this case, that the simulated value is equal to the experimental value) applied to a two-dimensional parameter space leads to a one-dimensional manifold of solutions (and is a form of practical non-identifiability, as the extracted experimental data is insufficient) [72].

Figure 2.5(c)-(d), demonstrates that, in contrast to a growth-to-confluence assay design, using a scratch assay design (where cells are initialised only in six rows of the domain) enables both P_m and P_p to be accurately estimated using the one-dimensional correlation function summary statistic, C_Y (this is possible as C_Y is a vector with 23 values). Finally, Figure 2.5(e)-(f) demonstrates how the predicted posterior can vary for a single summary statistic as the experimental design is changed. We use as an example the binning variance statistic, Q_8 , which is the deviation in cell numbers from the mean when the domain is divided into 8×8 bins. Increasing the rate of movement, P_m , means clusters are broken up more rapidly, whilst increasing the rate of proliferation, P_p , makes clusters larger. As a result, *in silico* (observed) data generated using small P_m and P_p results in the same binning variance as data generated with large P_m and P_p . The effect is more pronounced for a scratch assay design due to the initial confinement of cells.

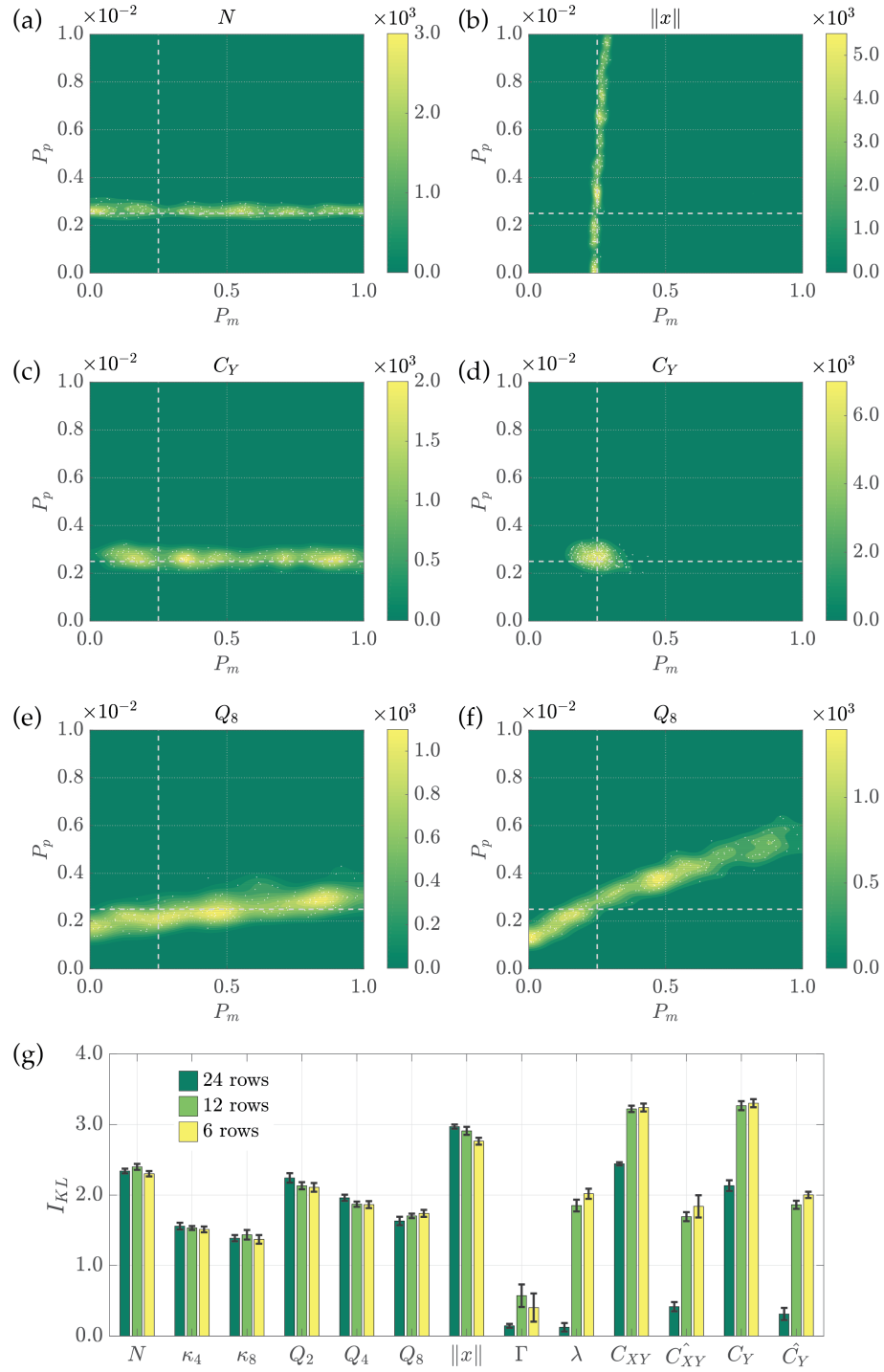


Figure 2.5: Posteriors generated using a single summary statistic as the experimental design is varied. The parameters used to generated the *in silico* (observed) data are indicated using dashed white lines. (a),(b) The posterior distributions corresponding to the most informative summary statistic for inferring P_p ((a), N) or P_m ((b), $\|x\|$) for the growth-to-confluence assay (cells initialised uniformly at random over 24 rows). (c),(d) The posterior distributions generated using the summary statistic C_Y for the growth-to-confluence assay (c) and the scratch assay (d) (cells initialised uniformly at random over six rows). (e),(f) The posterior distributions for summary statistic Q_8 for the growth-to-confluence assay (e), and the scratch assay, (f). (g) The information gain in moving from the prior to the posterior for each summary statistic and for each experimental design, where the error bars denote the standard deviation.

Figure 2.5(g) summarises the quality of the predicted posteriors resulting from the use of each of the summary statistics listed in Table 2.1 for each experimental design considered in Figure 2.1(c). Our results demonstrate that statistics based on cell numbers provide the same or less information as cells are increasingly confined initially, whereas the opposite is true for the summary statistics that relate to correlations in cell positions. Both model parameters, P_m and P_p , can be inferred using a single multi-valued summary statistic, without the need for cell trajectory information, but only when cells are initially confined to a scratch-assay-like geometry. For growth-to-confluence assays, cell trajectory data is necessary to accurately infer both model parameters.

2.3.2 What is the impact of the initial number of cells on the posteriors?

Next we consider how changing the initial number of cells in the assay (24, 48 or 72 cells) affects the quality of the predicted posterior distribution. We consider either a growth-to-confluence assay (cells initialised uniformly at random across the domain, Figure 2.6(a)-(c)) or a scratch assay design (cells initialised uniformly into six rows of the domain, Figure 2.6(d)-(f)). Figure 2.6(a)-(c) demonstrates that, for the growth-to-confluence assay, increasing the initial cell number does not facilitate accurate inference of both model parameters using a single summary statistic. In line with the results presented in Figure 2.5, Figure 2.6(d)-(f) indicates that both model parameters can be accurately estimated using a single summary statistic when the scratch assay is used; however, increasing the initial cell number does not significantly increase the quality of the posterior. We quantify the quality of the estimated posterior for each summary statistic and each experimental design in the bar plots in Figure 2.6(c),(f). For almost every summary statistic the quality of the posterior increases as the initial cell number increases, likely due to a reduction in the variance of each summary

statistic.

2.3.3 Are trajectory statistics necessary for maximal information gain?

It is natural to ask whether multiple summary statistics can be combined to increase the quality of the estimated posteriors and, further, whether combining different summary statistics can avoid the need to collect cell trajectory data to accurately infer parameters of a growth-to-confluence assay. To answer this question, we considered all combinations of two or three summary statistics from the 13 under consideration, computing the resulting posteriors and information gain for each combination. Figure 2.7 demonstrates that, for a growth-to-confluence assay, cell trajectory data are necessary to generate a posterior that allows for accurate estimation of both parameters P_m and P_p (Figure 2.7(a),(b)). In particular, trajectory information is necessary to estimate the motility parameter, P_m . It is interesting to note that, in both cases (with and without trajectory information), the number of cells summary statistic, N , is included in the best-performing two-summary-statistic combinations, likely because it facilitates accurate estimation of the proliferation parameter, P_p . For a scratch assay, cell trajectory information is not necessary to accurately infer both model parameters, however, as expected, including cell trajectory data enables more accurate inference of the cell motility parameter, P_m . Figure 2.7(e) quantifies the quality of the estimated posteriors for a range of experimental designs, initial cell densities and for combinations of one, two or three summary statistics. Results demonstrate that little extra information is gained by considering more than two summary statistics and that, as expected, accurate parameter inference for growth-to-confluence assays requires cell trajectory data.

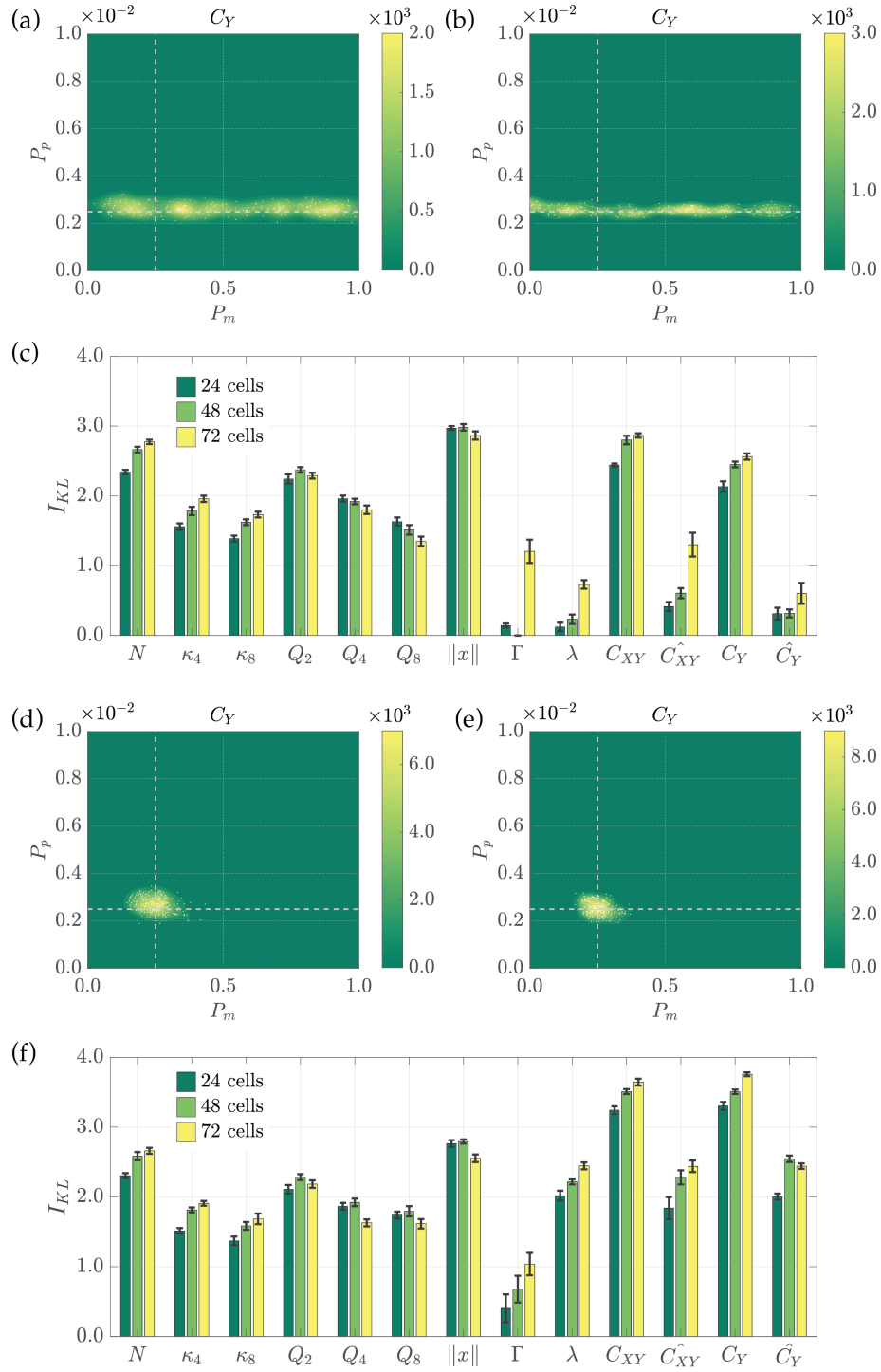


Figure 2.6: Posteriors generated using a single summary statistic as the initial cell number is varied. The parameters used to generated the *in silico* (observed) data are indicated using dashed white lines. (a)-(c) Results for a growth-to-confluence assay generated using 24 cells (a) and 72 cells. (d)-(f) Results for a scratch assay generated using 24 cells (d) and 72 cells (e). The posterior distributions in (a),(b) and (d),(e) are generated using the pairwise correlations summary statistic C_Y . (c),(f) The information gained in moving from the prior to the posterior for each summary statistic for each experimental design, where the error bars denote the standard deviation.

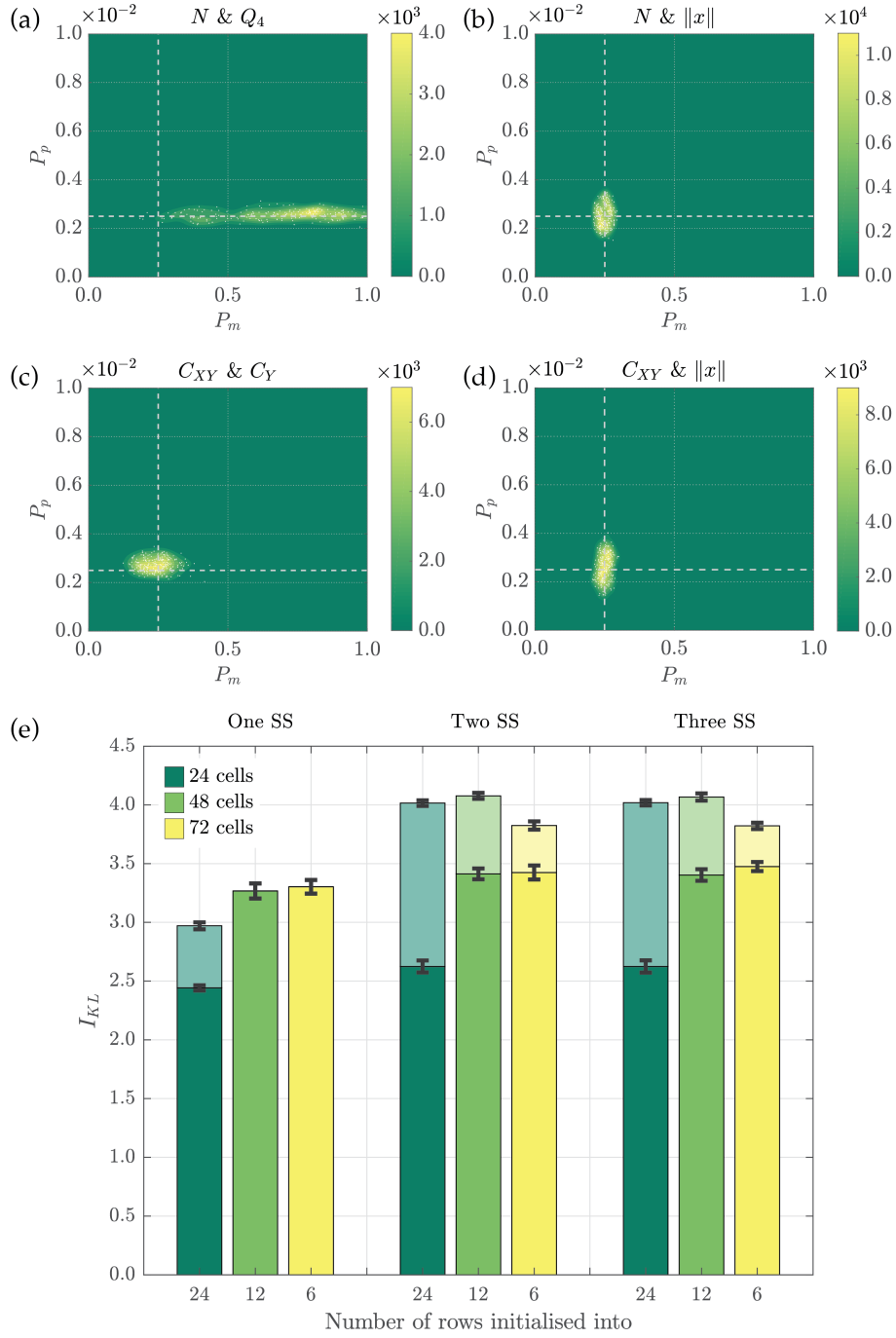


Figure 2.7: Posterior distributions and information gain when considering combinations of summary statistics (SS). (a)-(d) show posteriors resulting from the most informative two-summary-statistic combination when considering either a growth-to-confluence assay (a),(b), or a scratch assay, (c),(d), excluding trajectory information (a),(c), or including trajectory information (b),(d). (e) The information gain in moving from the prior to the posterior for each of the best performing summary statistic combinations, varying the experimental design, and excluding trajectory information (dark shading) or including trajectory information (light shading), where the error bars denote the standard deviation.

2.3.4 Can data-cloning ABC provides an efficient means to estimate model parameters?

Finally, we investigate whether the ABC-DC algorithm proposed by Picchini *et al.* [64] can provide an efficient means to estimate model parameters. Figure 2.8 shows the results of using ABC-DC with the two-dimensional pairwise correlation function, C_{XY} , to estimate posterior distributions for both P_m and P_p . Figure 2.8(a),(c) shows how the posterior changes as the number of clones is increased, relative to ABC rejection, whilst Figure 2.8(b),(d) shows the Markov chains of both P_m and P_p . Our results indicate that, as expected, the ABC-DC algorithm results in a tighter posterior focussed approximately about the MLEs found using ABC-DC with a single instance of the data, $K^{DC} = 1$. To understand how the MLEs generated using ABC-DC differ from those generated using ABC rejection we look at the posteriors generated using each algorithm together with three summary statistics: the number of cells, N ; the two-dimensional pairwise correlation statistic, C_{XY} ; and the Manhattan displacement, $\|x\|$. We calculate MLEs and the 5% and 95% empirical percentiles of the accepted parameters. Results in Table 2.2 indicate that the MLEs are similar for both ABC rejection and ABC-DC, though the credible intervals for ABC-DC are much smaller, and not a good representation of the true uncertainty in each parameter. However, ABC-DC can provide significant computational savings in estimation of the MLE, with maximal efficiency displayed with summary statistics that are relatively computationally demanding to compute, such as the two-dimensional pairwise correlations, C_{XY} .

2.4 Discussion

Mechanistic models together with experimental data from cell biology assays provide an excellent opportunity to characterise the extent to which cell proliferation and cell

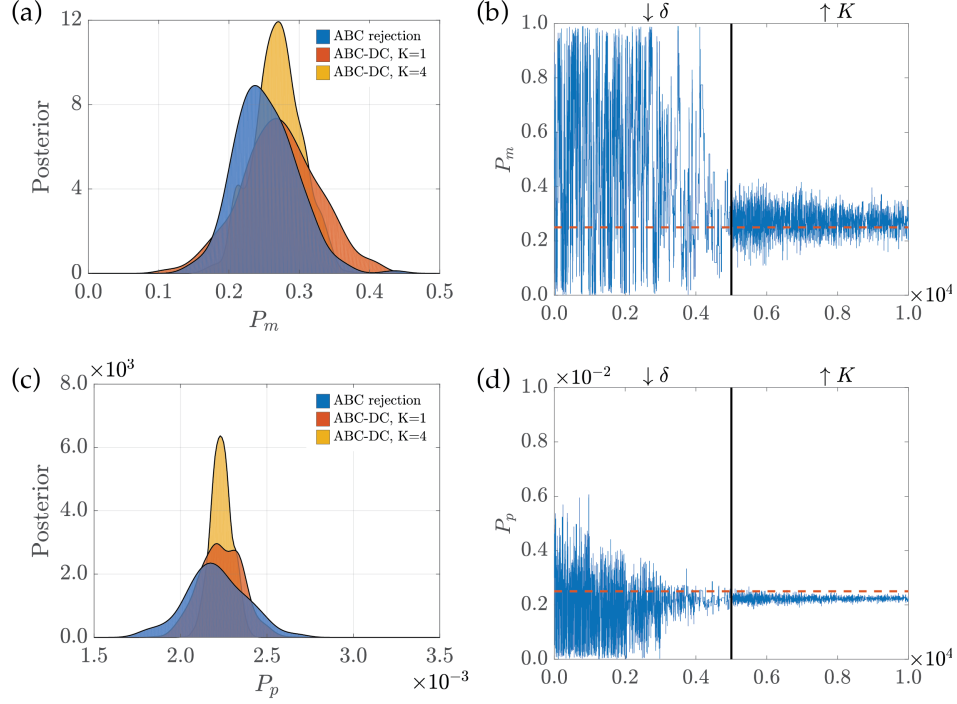


Figure 2.8: Example results from the ABC-DC algorithm for a scratch assay using the two-dimensional pairwise correlations statistic, C_{XY} . (a),(c) show estimates of the posteriors for P_m and P_p generated using kernel density estimation. We compare the posteriors for P_m and P_p generated using ABC-DC (yellow), ABC-DC (before the clones numbers are increased) (red) and ABC rejection (blue). (b),(d) show the trajectories of parameters accepted during the Markov chain Monte Carlo step.

Statistic: Parameter	ABC rejection MLE (5%-95% CI)	ABC-DC MLE (5%-95% CI)	Speed up
$\ x\ : P_m$	0.250 (0.231, 0.300)	0.247 (0.239, 0.255)	1.65
$C_{XY}: P_m$	0.239 (0.174, 0.310)	0.236 (0.186, 0.280)	4.50
$C_{XY}: P_p(\times 10^{-2})$	0.241 (0.215, 0.270)	0.228 (0.219, 0.236)	4.50
$N: P_p(\times 10^{-2})$	0.243 (0.220, 0.282)	0.246 (0.235, 0.257)	1.72

Table 2.2: ABC-DC for a scratch assay experiment. The MLEs and the 5%-95% empirical percentiles are shown for parameters accepted using ABC rejection and ABC-DC, for different summary statistics. The speed up of the ABC-DC method over ABC rejection is also given.

motility can drive the growth and expansion of a cell colony. To be able to draw quantitative distinctions between the relative contributions of cell proliferation and motility for different cell types and biological scenarios requires the accurate inference of model parameters using quantitative summaries of experimental data. In this chapter we have applied ABC to explore the extent to which different experimental designs and quantitative summary statistics impact our ability to accurately infer parameters of a simple mechanistic model of cell growth and spreading. Our results suggest that, for growth-to-confluence experiments, cell trajectory information, in addition to summary statistics such as cell numbers and spatial correlations, are required for accurate parameter inference. On the other hand, for a scratch assay geometry, it is possible to accurately infer both motility and proliferation parameters without recourse to tracking individual cells over multiple frames of a microscopy video. Since cell tracking provides the bottleneck in the analysis of many experimental studies, this renders experimental geometry an important aspect to consider the design of experiments.

In addition, our results demonstrate that increases in the initial cell number in cell biology assays can increase the quality of estimated posterior distributions by decreasing noise in the summary statistics, a result also found for a similar model by Ross et al. [77]. We also demonstrated that posteriors of significantly increased quality can be generated using a combination of two summary statistics, but that further increases in the number of summary statistics provided little increase in the quality of the posteriors. In Appendix A.3, we demonstrate that our results are consistent across a range of parameter values and a range of experimental durations. However, it should be noted that, because we collected data from only the final time point of the experiment, some summary statistics lose efficacy if experiments are run over a sufficiently long time period that the population grows to confluence. For example, in this case, the number of cells, N , at the final time point provides little information on

the model parameters. Finally, we have shown that the ABC-DC approach provides an efficient means to approximate the MLEs of model parameters, with significant improvements in computational efficiency provided where the summary statistics are time-consuming to calculate.

2.4.1 Predictions for experimental data

It is possible to use the established model to infer parameter distributions for different cell types in a growth-to-confluence assay, and subsequently make predictions about long-time behaviours, such as the time taken to reach confluence. In Appendix A.4 we consider both an automated means of extracting quantitative data from experimental images and a manual method. The following results use experimental data extracted manually. In Figure 2.9 we use data from the growth-to-confluence experiments detailed in [90] to infer motility and proliferation rates for two different cell types: 3T3 fibroblast cells and MDA MB 231 breast cancer cells. In particular, for each cell type we use ten experimental replicates to estimate the bivariate posterior distribution for P_m and P_p , with data collected 12 hr into the experiment. Figure 2.9(a),(b) shows the posteriors generated using ABC rejection with the cell number, N , and Manhattan displacement, $\|x\|$, as summary statistics, and the prior distributions $\pi_{3T3} = \pi_{231} = \pi_{exp}$

$$\pi_{exp}(P_m) = \mathcal{U}(0, 0.1);$$

$$\pi_{exp}(P_p) = \mathcal{U}(0, 0.01).$$

For the 3T3 fibroblast cell line we obtain $P_m = 0.0465 \pm 0.00341$ and $P_p = 0.00163 \pm 0.000403$, whilst for the MDA MB 231 breast cancer cells we obtain $P_m = 0.0227 \pm 0.00254$ and $P_p = 0.000931 \pm 0.000427$. In each case we give the posterior mean and the standard error. In real terms, for the 3T3 cells this gives a diffusivity

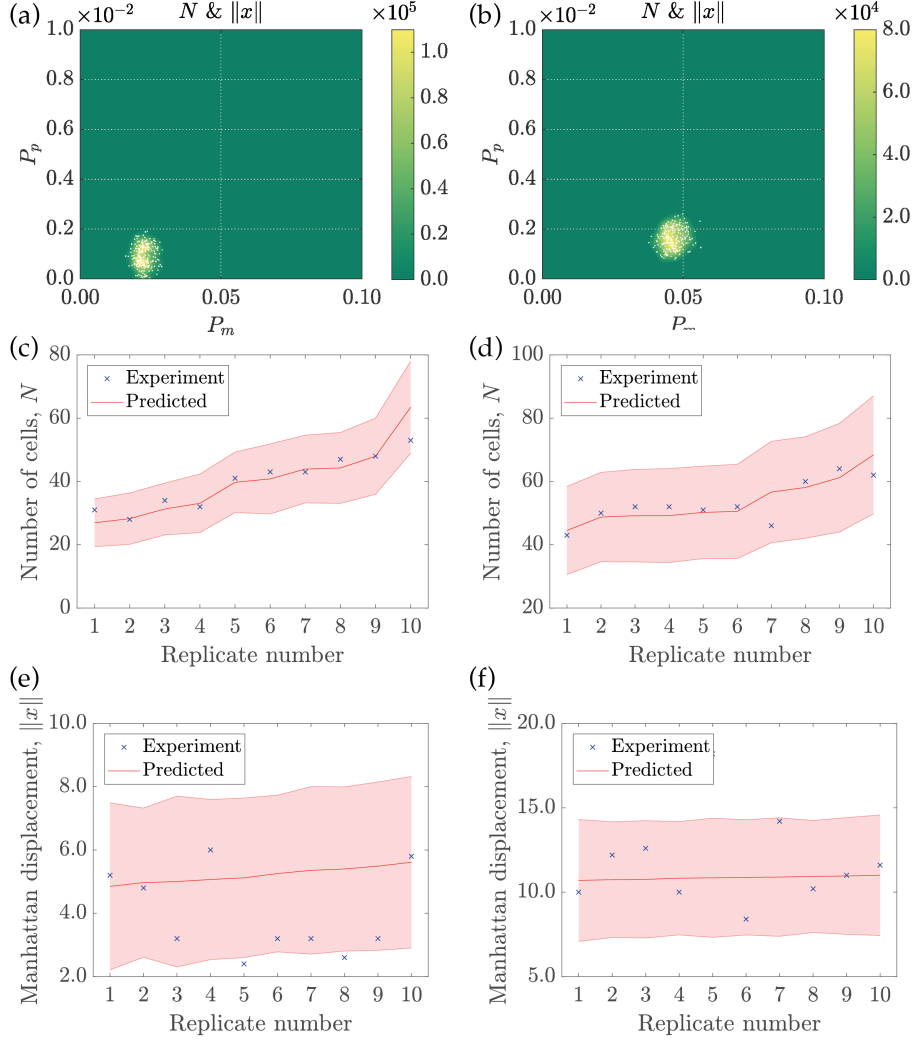


Figure 2.9: Posterior distributions and results of posterior predictive checks. (a),(b) Posterior distributions for 231 breast cancer and 3T3 fibroblast cell lines, respectively. (c),(d) Posterior predictive check using data collected 24 hr into the experiment, for the 231 breast cancer and 3T3 fibroblast cell lines, respectively, and the cell number, N as a summary statistic. The ten experimental replicates are ordered according to increasing cell number at 24 hr. For each replicate we plot the experimental data (blue cross), the mean cell number (red line) and the 95% confidence interval (shaded region). (e)-(f) Posterior predictive check using data collected 24 hr into the experiment, for the 231 breast cancer and 3T3 fibroblast cell lines, respectively, and the Manhattan displacement, $\|x\|$, as a summary statistic. Details are as for (c),(d).

of $D = 97.986 \pm 7.196 \mu\text{m}^2 \text{ hr}^{-1}$ and a proliferation rate of $\lambda = 0.0392 \pm 0.00968 \text{ hr}^{-1}$, whilst for the 231 cells this gives $D = 47.850 \pm 5.347 \mu\text{m}^2 \text{ hr}^{-1}$ and $\lambda = 0.0223 \pm 0.0102 \text{ hr}^{-1}$.

To check the validity of the model and inferred posterior distributions, we collected further data from the same experimental replicates at 24 hr and used these to perform a posterior predictive check. In Figure 2.9(c)-(f) we present the results of performing the posterior predictive check using both the cell number, N , and the Manhattan displacement, $\|x\|$, as summary statistics. We produce results using, for each data set, 1000 samples from the estimated posteriors, and visually compare the statistics for each experimental replicate with the average of the predicted statistics and their 95% confidence intervals. These plots show strong support for the model for each cell type. It can be seen that for the number of cells statistic, all of the experimental statistics fall within the 95% confidence intervals. For the Manhattan displacement statistic, two of the ten replicates for the 231 cells have an experimental statistic that falls outside the 95% confidence interval of the simulated statistic. In this case, we suspect translation of the positions of slow-moving cells (as in the case of the 231 cells) onto a lattice has introduced artefacts, causing discrepancy between simulation displacement and experimental displacement.

2.4.2 Conclusions

In summary, our results suggest that experimental design choices that incorporate initial spatial heterogeneities in cell positions facilitate parameter inference without the requirement of cell tracking, whilst those that seed cells uniformly initially require cell tracking for accurate parameter inference. As cell tracking is often a major technical limitation in many experimental studies of this type, our recommendations for experimental design choice could lead to significant potential time and cost savings in the analysis of these kinds of commonly-used experiments.

In the next chapter, we consider the use of these methodologies on different experimental data, where cell motility is influenced by an electric field. Understanding the impact of this new mechanism, and the different experimental domains considered, will help to demonstrate the applicability of these methods to investigate experimental design in a new context.

Chapter 3

The influence of motility bias on parameter inference

3.1 Introduction

In this chapter the inference methods of Chapter 2 are used to determine an appropriate experimental design for cell spreading experiments that influence the movement of cells using electric fields. This work forms the initial investigations from collaboration with Min Zhao (University of California San Diego), who is interested in quantifying the impact of electric fields on the spread of cells, to understand wound healing. I thank them for providing the experimental data for this chapter.

Movement of cells due to an electric field is termed electrotaxis (or galvanotaxis), and has been observed in a variety of cell types, including cells that would not typically experience electric fields in their usual environments. Understanding wound healing requires an understanding of electrotaxis, because it has been suggested that the polarisation of cells directs cell migration in tissues towards a more negatively charged wound centre [30].

Electrotaxis is recreated *in-vitro* by subjecting cells initialised on membranes to

electric fields of various strengths for a duration of time, and collecting time-lapse images of the domain. These images can be used to determine the positions of cells over several consecutive time points, and various statistics can be calculated; for example the trajectory speeds (trajectory distance moved divided by time), displacement speeds (displacement from the origin divided by time) or tortuosity of cell paths (trajectory distance moved divided by displacement from the origin) [73, 114].

These statistics can be compared for different cell types and a range of field strengths, to understand how field strength affects motility. However, to make predictions of motility bias in an unfamiliar context it is useful to quantify the change of parameters of a mathematical or computational model. By quantifying the uncertainty in parameter estimates using Bayesian inference, it is possible to capture different trajectories that cells may undergo when being exposed to an electric field, in the scope of the computational or mathematical model considered.

Notably, it is possible to build upon the investigations into appropriate experimental design from Chapter 2 for maximising information gain in electrotaxis experiments, to determine the optimal initialisation of cells. Though investigations into wound healing typically use scratch assays to investigate the expansion of a region of cells into an empty domain [17], barrier assays can also be used to initially confine cells [113]. Similar experiments can be performed where a specific shape of cells is removed from the domain, for example through either a stamp assay (used in Lee et al. [47] to investigate clearance of cell debris) or a sticker assay [37], (where various shaped regions of cells are removed from a confluent domain before observing the growth of the surrounding cells into this empty space). The experiments considered by Min Zhao consider cells initialised in various shapes (confined with barriers into either a square, circle or triangle), before being subjected to electric fields, though it is not currently known which initialisation shape is most useful to adopt to determine the parameters of a mathematical model with best accuracy.

The novelty of this work is twofold; firstly in determination of the most informative experimental design (the number of cells to initialise, the shape of assays to consider, and the statistics that should be captured) to infer the model parameters in the two-dimensional model adopted. Secondly, a demonstration of the inference of these parameters when using experimental data, outlining some of the potential pit-falls that a future user may experience, and how these can be addressed.

3.2 Methods

In order to quantify the impact of the introduction of a motility bias, it is important to link experimental data with a computational or mathematical model. Only through this link is it possible to understand the impact of the various experimental choices on the inference of model parameters. Determining the identifiability of a bias parameter is tantamount to knowing whether these experiments can effectively determine differences in cell spreading with the mathematical or computational tools available.

In this section, we outline the methods used to perform Bayesian inference investigations for either a synthetically created data or experimental data. we demonstrate the experimental data under consideration and how these data can be replicated in a computational model, before giving details of the inference algorithm. ABC-SMC is applied in the following section to synthetic data first, to understand the impact of experimental design choices, before applying the same methods to experimental data, allowing comparison between synthetic examples and experimental examples.

3.2.1 *In silico* data

The experimental data provided by Min Zhao investigates the electrotaxis of cells, under varying electric field strengths, electric field protocols, and initial tissue shapes.

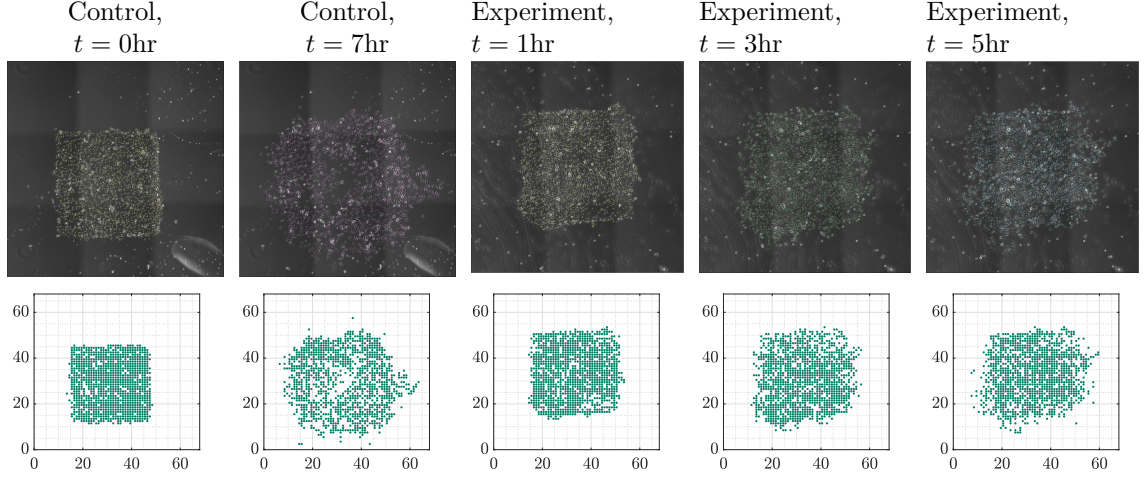


Figure 3.1: The top row displays experimental images at various times. In control experiments, there is no electric field, and cells spread freely for the duration of the experiment (seven hours). In the right-most three images, an electric field biasing cell motility downwards is introduced at $t = 1$ hr and lasts for two hours, until at $t = 3$ hr the direction of the electric field is reversed for a further two hours. In the bottom row, the cell positions in the experimental images from the top row are manually selected, and are approximated into the two-dimensional lattice model such that every cell occupies a unique lattice site as close as possible to its position.

In brief, experiments are performed by initialising a culture of human corneal epithelium cells onto a commonly used hydrogel material. Electric fields are applied along the vertical axis for a duration of two hours at a time through agar salt bridges, and images are captured at five minute intervals [30].

In total, 85 images are captured over a duration of seven hours. After one hour the electric field is turned on for two hours, before the polarity of the field is reversed for a further two hours. The electric field is then turned off for the final two hours of the experiment. Experimental images are shown in Figure 3.1.

In these initial investigations, only a subset of the available data are considered, because the manual tracking of cell positions is a time-consuming process. The data considered is the two hour time period when the electric field, and hence the motility bias, is first introduced. Cell positions are marked manually at the start and end of this two hour period.

3.2.2 Mechanistic model

Replicating experiments involves both accounting for the bias of cells, which can be achieved by using a bias parameter (such as those similarly adopted in previous cell spreading investigations [61]), and replicating the shape of the initial domain.

3.2.2.1 Introduction of bias parameter

The model adopted is the two-dimensional lattice-based model introduced in Chapter 2, whereby N_0 cells are initialised in a lattice with R rows and C columns. In each time-step, $\tau = 1/24$ hr, $N(t)$ cells are chosen at random with replacement and a random number is sampled for each to determine if it moves, $r \sim \mathcal{U}(0, 1)$. The model consists of a motility parameter, $P_m \leq 1$, and bias parameter, $P_b \in (-1, 1)$, such that a movement event is attempted if $r < P_m$, and movement upwards is attempted with probability $P_m(1 - P_b)/4$, and movement downwards is attempted with probability $P_m(1 + P_b)/4$ (movement left or right is attempted with probability $P_m/4$). There is no bias in the left-right direction to replicate the experiments under consideration.

A movement event attempted into an already occupied site, or outside the domain, is aborted (no flux boundary condition). However, as cells generally do not reach the boundary, this choice of boundary condition does not impact results herein. The model is iterated until a final time, $T = 2$ hr, is reached. Proliferation is neglected in the model as, during the short time period of experiments, and with a densely packed initial domain, there are very few proliferation events, (for example, in experimental data, the number of cells increased by less than 5% over the seven hour period).

To demonstrate the impact of the introduction of bias on the model, in Figure 3.2 different final domains are shown after two hours with fixed motility parameter, P_m and varying bias parameter, P_b , for domains of cells initialised in a triangular shape. When initialising in a triangle shape, it can be observed that for a strong downwards bias, the bottom of the triangle separates from the bulk, whereas for a strong upwards

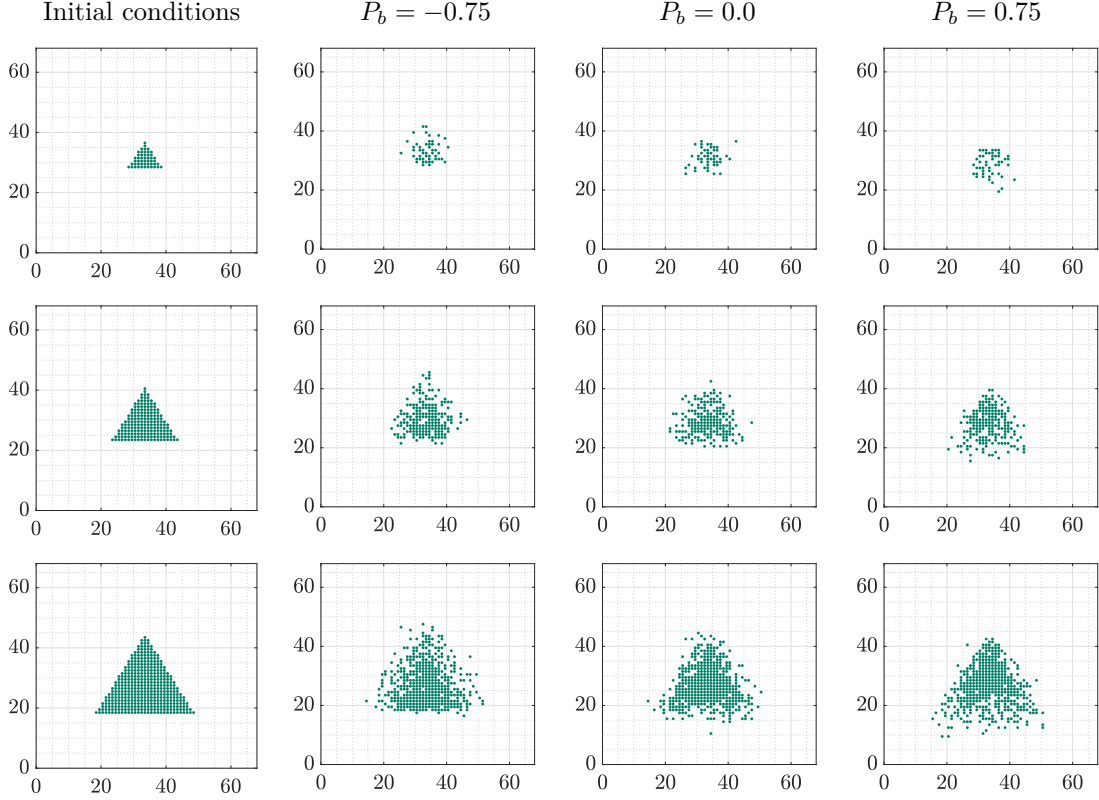


Figure 3.2: The impact of varying the bias parameter, P_b , on the final domain when initialising in the shape of a triangle with radii, $r = 5$ (top row), 10 (middle row) and 15 (bottom row). As a positive P_b represents a preference for movements downwards, different values of P_b correspond to different degrees of drift. The ratio of the expected number of upwards movements to downwards movements is 7:1, 1:1, or 1:7 for $P_b = -0.75$, 0.0, and 0.75 respectively. These images demonstrate the subtle variability of the final domain under large changes in the bias parameter, P_b .

bias, the cells on the left/right sides of the triangle disperse. Understanding whether an upwards or downwards pointing triangle is more informative for obtaining accurate parameter estimates is evaluated in Section 3.3.2.

3.2.2.2 Replication of experimental domain

Cell positions are extracted from experimental data by manually marking cell positions in images using the `imageJ` software [70], before importing cell positions (in pixels) into `MATLAB` where they are overlaid onto the $R \times C$ lattice, and the routine demonstrated in Appendix A.4 is used to place each cell into the nearest possible

unique lattice site. Images are $1350\text{px} \times 1344\text{px}$, and the scale of the image is such that there are $1/0.61$ microns per pixel. Cell diameters are approximately $20\text{px} \approx 33\mu\text{m}$, estimated using `imageJ`. Using this estimate for cell diameter, the experimental domain is approximated to a lattice with $1350/20 \approx 67$ lattice sites, therefore $R = 67$ and $C = 67$ matches experimental data.

These procedures are demonstrated in Figure 3.1. Note that if a cell cannot be placed into its closest site, or the four orthogonally adjacent sites, it is not placed. Non-placement occurs very few times (less than 1% of cells are unable to be placed, results not shown) as we chose the size of lattice sites to approximately match the size of a typical cell. Conversion onto a lattice allows simulation and experimental statistics to be computed in the same manner, and hence they are easily comparable.

3.2.2.3 Synthetic replication of experimental domain

To investigate all possible experimental geometries, we generate *in-silico* data using the model. In order to replicate experiments where cells are initialised into a square, circle or (equilateral) triangle, the simple `MATLAB` codes in B.1 are used. The initial domains used are shown in Figure 3.3. We vary the shape (circle, triangle or square), the size of the shape ($r = 5, 10$ or 15) and the degree of confluence of the shape (75% or 100%). The cell number is not kept constant when creating different synthetic initial domains, which means later work on determining the most informative experimental domain is flawed, so should be addressed in future work.

3.2.3 Inference of model parameters

To infer parameters of the two-dimensional computational model introduced, we adopt the Bayesian methodologies introduced in Chapter 2, allowing estimation of uncertainty in parameter estimates. ABC rejection is used, as it is easy to implement, and it is straightforward to estimate posterior distributions for model parameters,

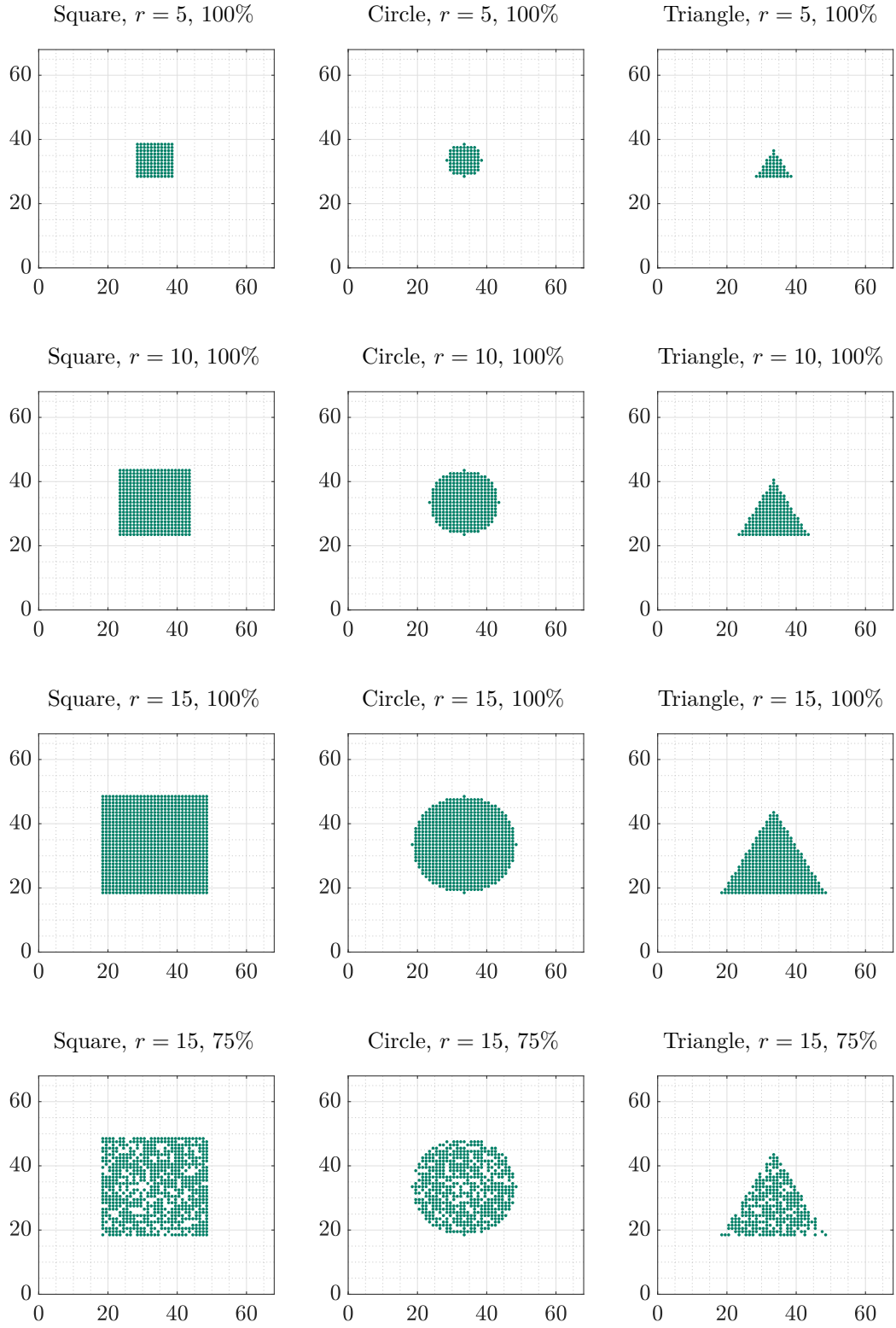


Figure 3.3: Cells are initialised into either a square, circle or triangle shape with various radii, r , and initial degrees of confluence, either 100% or 75% (bottom row). These are potential experimental domains at the beginning of experiments considered by Min Zhao.

P_m and P_b . The algorithm is presented in Chapter 2, so only a short description is provided here.

To perform ABC rejection, sample $\theta = (P_m, P_b)$ 20,000 times from uniform (uninformative) prior distributions,

$$\begin{aligned}\pi(P_m) &= \mathcal{U}(0, 1); \\ \pi(P_b) &= \mathcal{U}(-1, 1).\end{aligned}$$

Simulate the model with the sampled parameters, to obtain data that is used to compute statistics, from which a distance, d , between simulated and experimental statistics is calculated. The parameters that result in the smallest 1% of distances are retained to approximate a posterior distribution (these posterior distributions are visualised with the aid of a two-dimensional kernel density estimation). The Kullback-Leibler information gain is computed to quantify the information gain in parameters P_m and P_b when moving from the prior distribution to the posterior distribution.

3.2.3.1 Summary statistics

Without a proliferation mechanism and with the inclusion of bias, some of the statistics used in Chapter 2 are no longer informative, and it is therefore important to consider new statistics that can account for bias in movements upwards or downwards. In particular, we no longer consider the number of cells statistic, N , and further, with an initially dense domain, tracking of randomly chosen cells is less effective as a randomly chosen cell is likely confined in the centre of the cluster of cells, and so would move very little. We could instead investigate the trajectories of cells near the edges of the group of cells, but with the introduction of bias it is hoped that other statistics will be informative. We therefore opt not to consider the tracking statistics; Manhattan displacement, $\|x\|$, and tortuosity, Γ . The following twelve statistics are

considered:

- The leading edge position is approximated by considering the average vertical position of the five top-most, or bottom-most cells, for the top leading edge, $\overline{LE}$ or bottom leading edge $\underline{LE}$, respectively.
- The vertical position of the centre of mass of cells is given by the average of all the vertical positions of cells in the domain, ρ_Y (accounting for cell drift), or the standard deviation of those vertical positions, σ_{ρ_Y} (accounting for cell separation).
- The gyration tensor is again considered, but instead of the smallest eigenvalue, the largest gyration tensor eigenvalue is considered, as this accounts for the spread of cells across their longest axes. The smallest gyration tensor eigenvalue was appropriate in Chapter 2 for scratch assays where cells are initially confined to a small number of rows but here it is more appropriate to consider the largest gyration tensor eigenvalue, as larger bias in motility leads to larger separation of cells in the vertical direction.
- The binning variance statistic, Q_2 , is computed as the sum of differences of number of cells in quadrats (of size 2×2) from the average number of cells expected in a quadrat, and exclude the 67th row and column from its computation (which makes little difference as cells rarely reach the boundary). Previously, Q_2 , Q_4 , and Q_8 were considered, but they performed similarly, and so we consider only Q_2 in the following results to minimise the number of similar statistics considered.
- The vertical one-dimensional pairwise correlation function, C_Y , or two-dimensional pairwise correlation function, C_{XY} , along with their normalised counterparts ($\hat{C}_Y$ and $\hat{C}_{XY}$), as defined in Chapter 2. As cells are initialised with a high

Summary statistic	Notation
Position of upper leading edge	$\overline{LE}$
Position of lower leading edge	$\underline{LE}$
Vertical position of centre of mass	ρ_Y
Vertical standard deviation of cell positions	σ_{ρ_Y}
Size of largest cluster, k -connected	$\kappa_k, k = 4, 8$
Binning variance, bin size = 2	Q_2
Largest gyration tensor eigenvalue	λ_l
Correlation function	$C_{XY}(l)$
Normalised two-dimensional correlation function	$\hat{C}_{XY}(l)$
One-dimensional correlation function	$C_Y(l)$
Normalised one-dimensional correlation function	$\hat{C}_Y(l)$

Table 3.1: A list of the summary statistics and the corresponding notation adopted. All summary statistics are evaluated at the final time, T .

degree of correlation (i.e. there are many cell pairs at low distances), these statistics should distinguish between different P_m and P_b values.

- The largest contiguous area of cells is used to determine the overall connectivity of the cells. This is defined as either the size of the largest connected cluster of cells allowing only orthogonal connections (κ_4), or also allowing diagonal connections (κ_8).

The notation adopted when referring to these statistics in figures is shown in Table 3.1.

3.2.3.2 Computing distances

The distance between simulated data and experimental data is computed using the same distance function as introduced previously, however the process is slightly simpler because only one replicate is considered (due to the sparsity of available data). Labelling each statistic as $s_j^k(l)$, where $j = 1, \dots, 12$ differentiates the statistic used, $k = 1, \dots, 20000$ refers to the sample considered, and $l = 1, \dots, L_j$ are the values of the statistic ($L_j = 132$ for C_{XY} or $\hat{C}_{XY}$; $L_j = 66$ for C_Y or $\hat{C}_Y$; $L_j = 1$ for all others).

First compute the median absolute deviation (MAD) for each statistic, $\sigma_j(l)$. The MAD is defined for a univariate data set, $X = \{X_i\}$, as $\text{MAD} = \text{median}(|X_i - \text{median}(X)|)$. Then compute the distances for each statistic, $j = 1, \dots, 12$, as

$$d_j^k = \sqrt{\frac{1}{L_j} \sum_{l=1}^{L_j} \left(\frac{s_j^k(l) - s_j^{obs}(l)}{\sigma_j(l)} \right)^2}. \quad (3.1)$$

In this work, combinations of $A = 2$ summary statistics are sometimes considered from the twelve statistics listed in Table 3.1, and in that case, the distance is calculated as

$$d^k = \sqrt{\sum_{a=1}^A \left(d_{j_a}^k \right)^2}, \quad (3.2)$$

where j_a is the index the a^{th} statistic.

We use regression adjustment using the methods in Section 2.2.3.4. Without using regression adjustment, the posterior distributions provide much less information about model parameters, as demonstrated in Appendix B.2.1.

To quantify the information gained in moving from the prior distributions to each posterior distribution, the Kullback-Leibler, divergence, I_{KL} , is computed in exactly the same manner as outlined in Section 2.2.3.2. I_{KL} can be computed either from the full two-dimensional posterior distributions or the marginal one-dimensional posterior distributions (obtained by integrating the two-dimensional posterior distributions over

one of the parameters).

3.3 Results

3.3.1 Is it possible to infer the bias parameter?

We investigate whether the bias parameter is identifiable in the model using the statistics outlined. Initially, consider the inference of model parameters using synthetic data with an initial cluster of cells forming a square in the centre of the domain with radius $r = 15$. The experimental data considered in the next section is most similar to this initialisation. The synthetic cell initialisation is simulated with $P_m = 0.25$ and $P_b = -0.5$ for $T = 2$ hr, and the final domain is used to generate synthetic experimental statistics, which are compared to simulation statistics.

In Figure 3.4 we show the posterior distributions resulting from the 12 statistics under consideration, as well as four posteriors generated by combining statistics. Statistics leading to posterior distributions that are shaped differently are chosen for combination, and the posteriors for some combinations are shown. It is expected that differently shaped posteriors complement each other when combined, so are likely to increase the Kullback-Leibler information gain. The information gains in moving from the prior to the posterior distribution for each of the 16 posterior distributions are shown in the bar plots.

The two-dimensional information gains are much larger when combining statistics. In particular, combining statistics is necessary in order to maximise the one-dimensional information gain in parameter P_b . The combination of statistics that maximises the information gained about P_m and P_b is ρ_Y , the vertical position of centre of mass, and Q_2 , the binning variance. This combination of statistics results in the following maximum likelihood parameter estimates:

- $P_m = 0.2201 \pm 0.0225$,

- $P_b = -0.5522 \pm 0.0746$,

which are close to the values that the synthetic data are generated with of $P_m = 0.25$ and $P_b = -0.5$.

These results demonstrate the identifiability of both bias and motility parameters in experiments, without requiring a trajectory-based statistic such as the Manhattan displacement or the tortuosity (that were previously considered necessary to establish an estimate for the motility parameter, P_m). We believe this is due to the larger initial cell number, and the densely packed nature of the initial domain, which have allowed correlation-based statistics to demonstrate high sensitivity to changes in parameter P_m .

3.3.2 What bias direction leads to more informative statistics?

We investigate the impact of changing bias direction on the information obtained about the model parameters, P_m and P_b . For the shapes considered, only triangle shapes are asymmetric, so for these it is important to question whether one particular bias direction is more informative than another. In Figure 3.5 we show posterior distributions for a subset of statistics or statistic combinations for both an upwards bias and a downwards bias, the information gained for each statistic, and some of the most informative statistic combinations.

It can be seen that combining statistics is necessary to maximise the information gain and, that for the most informative statistics combinations, an upwards bias yields more information than the downwards bias, with a triangle oriented pointing upwards. For some statistic combinations, using an upwards bias is not more informative than a downwards bias, however, an upwards bias results in posterior distributions that overlap orthogonally from the bottom leading edge statistic, $\underline{LE}$,

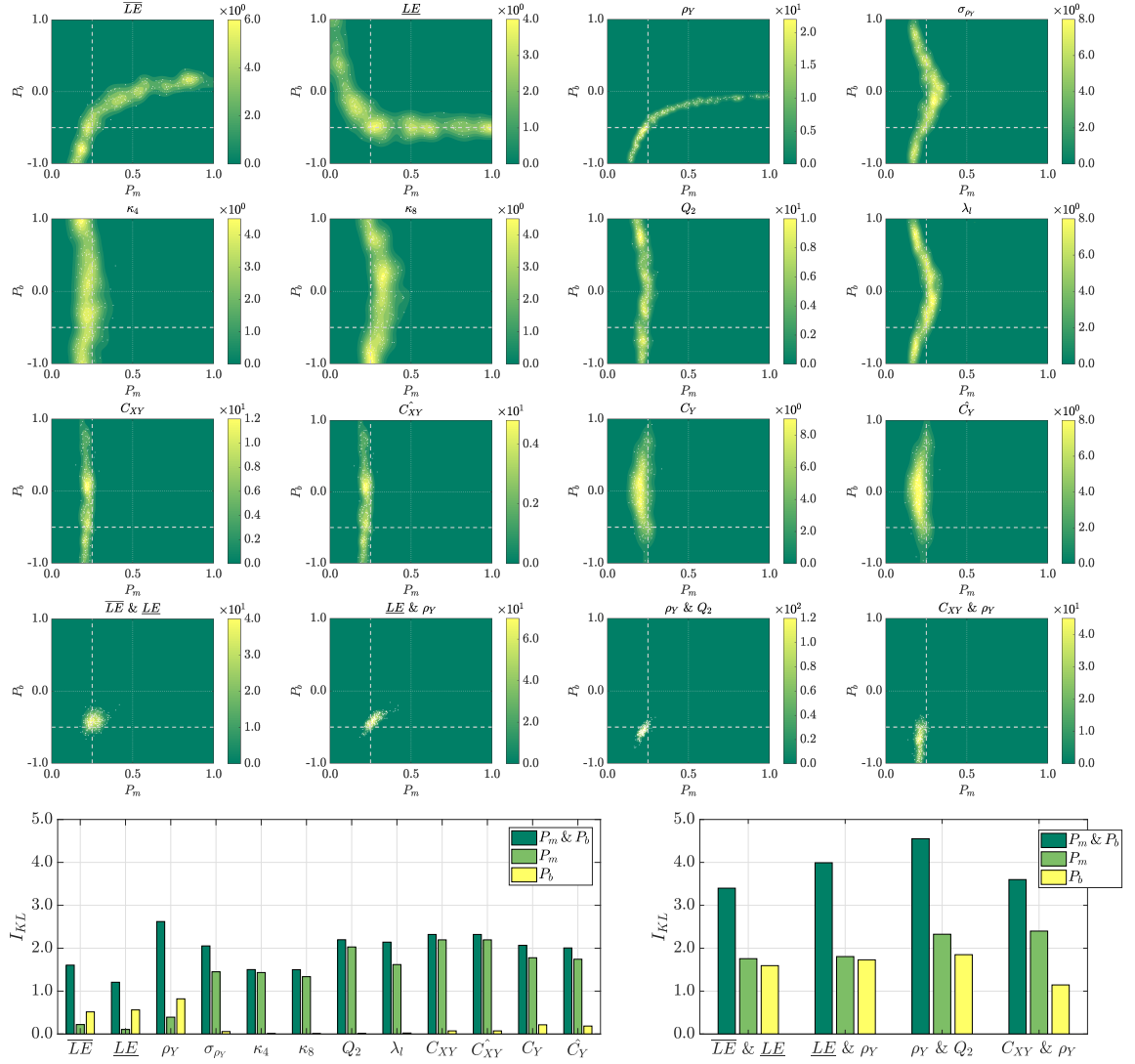


Figure 3.4: Posterior distributions and information gains resulting from various summary statistics (and combinations) are shown for synthetically created data initialised in a square with $r = 15$ and at 100% confluence. In the first three rows, posterior distributions for single summary statistics are shown. Each posterior distribution corresponds to the statistic titled, where the statistic shorthand is used (see Table 3.1). The fourth row posterior distributions correspond to those generated when using a combination of statistics. Posterior distributions are generated using kernel density estimation on the samples from prior distributions that resulted in the lowest distance between simulated and (synthetic) experimental data. In each posterior distribution, larger posterior densities are more yellow (according to the colourbar), the parameters used to generate the posterior distribution are shown as white dots, and the parameters used to generate the synthetic experimental data are shown as grey dashed lines. The bar plots show the corresponding two-dimensional (or one-dimensional) information gains in moving from the prior to the posterior distributions (or marginal one-dimensional posterior distributions).

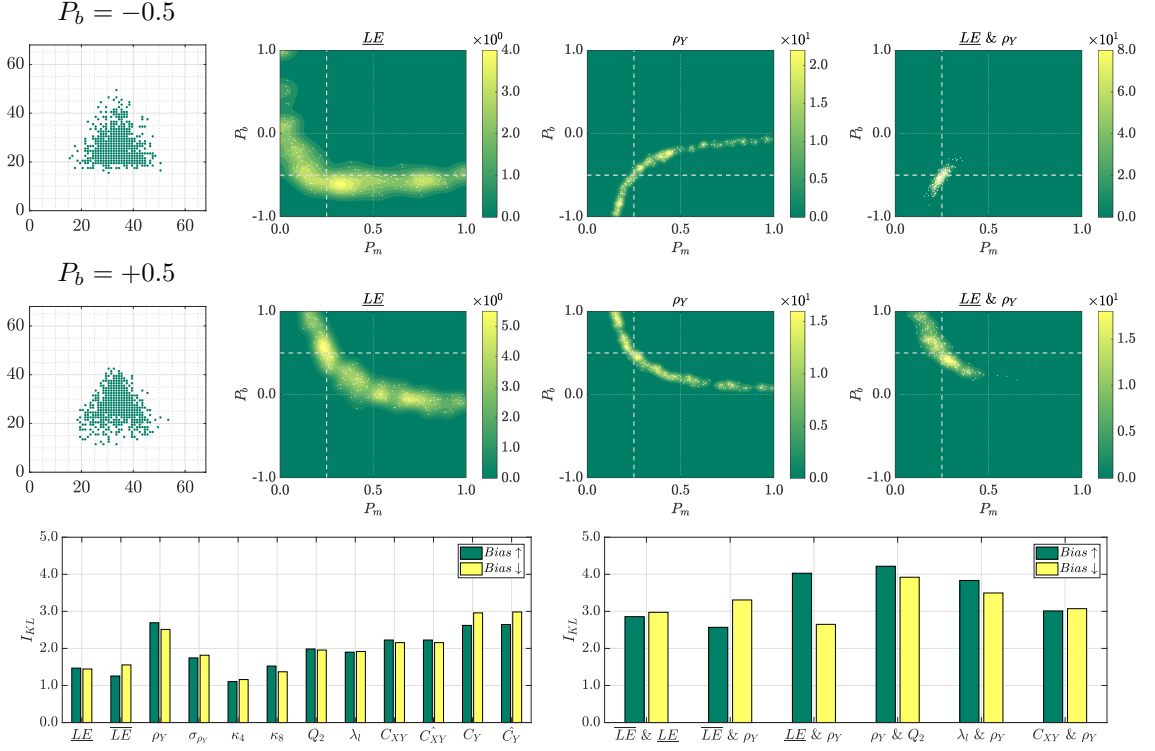


Figure 3.5: Posterior distributions for inference investigations when initialising in a triangle with radius $r = 15$ and at 100% initial confluence, with either an upwards or downwards bias. The posterior distributions are estimated from the parameters resulting in the smallest (1% of) distances between simulated statistics and experimental statistics (regression adjustment is applied to these parameters). We also display the information gain when moving from the prior to the posterior distributions generated by each statistic considered, and for a selection of two-statistic combinations, with experimental data generated with either an upwards bias or downwards bias.

and the vertical position of the centre of mass, ρ_Y , i.e. these posterior distributions are approximately right-angled to one-another, so they complement each other when multiplied or otherwise combined. Therefore, an upwards facing triangle with an upwards bias provides more informative statistics.

3.3.3 What initial configuration maximises information gain?

In order to determine the optimal initial configuration of cells to maximise the information gained in the bias and motility parameters, further synthetic data are generated for different potential experiments. The following experimental choices are

varied: different radius, $r = 5, 10$ or 15 ; different initial densities of cells, so the shape is 75% confluent or 100% confluent; or different initial shapes, either square, circle or triangular.

In Figures 3.6, 3.7, and 3.8, the information gain when varying different initial radii, initial densities, or initial shapes is shown. In each case, the information gain is only shown for the posteriors generated by the statistics combinations investigated earlier ($\overline{LE}$ & LE , LE & ρ_Y , ρ_Y & Q_2 , or C_{XY} & ρ_Y) as these statistics consistently present themselves as the most informative statistics for any of the initial configurations considered (results not shown).

Clear trends appear in the information gain plots, in Figure 3.6 it can be seen that increasing the initial radius of the shape generally leads to a larger information gain; in Figure 3.7 initialising at 100% confluence provides more information than initialising at 75% confluence; and finally in Figure 3.8 a square shape is more informative than either a circular or triangular shape. It is surprising that these trends (larger, more confluent domains are more informative) are not true for all statistic combinations.

Overall, the most informative configuration is a square initial domain at 100% confluence initialised at the largest radius, $r = 15$. The corresponding posterior distributions have been shown previously in Figure 3.4. Note that this domain contains the largest number of cells initially, so it is possible that this result is due to the increase in the number of cells; the stochasticity of cell migrations are less prominent with more cells, so there is less variation in statistics. However, information gains from initialising in circles or triangles are similar, despite having different numbers of cells initially.

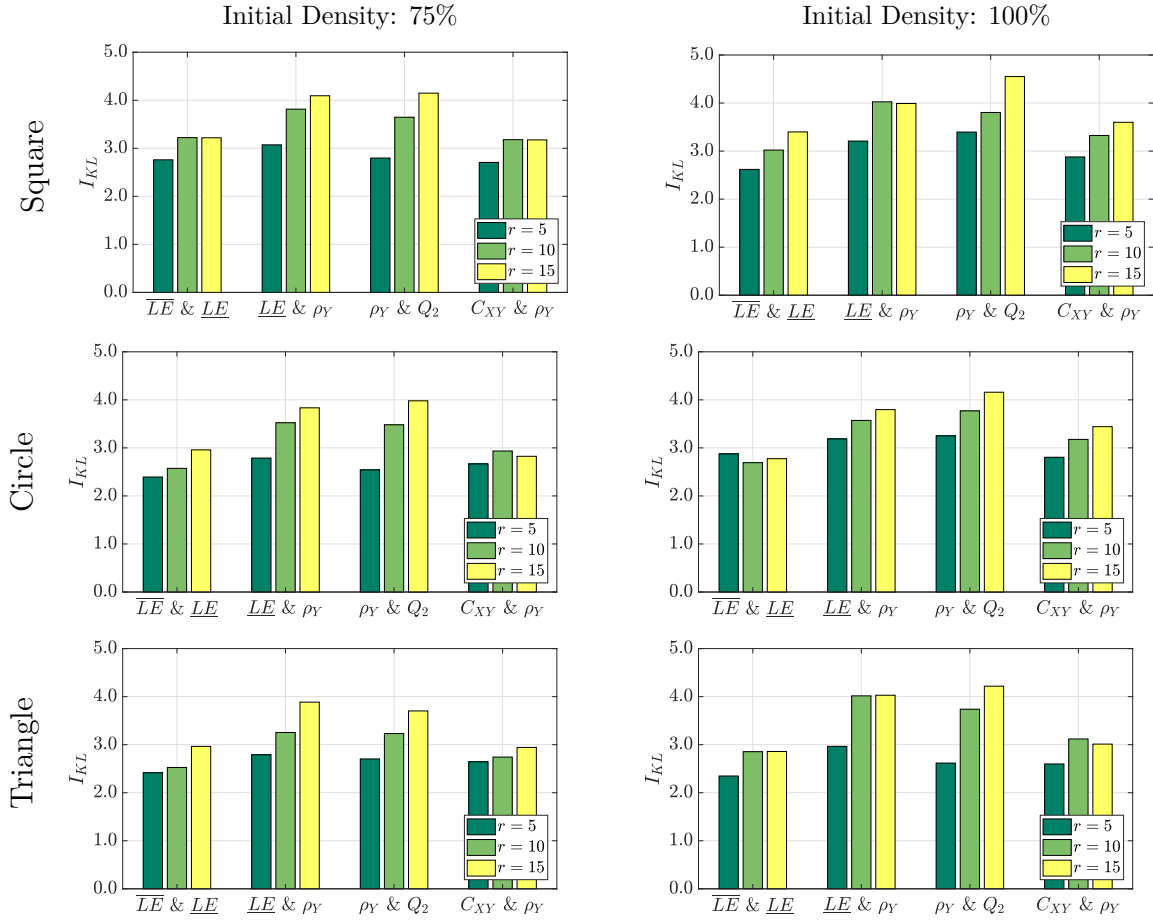


Figure 3.6: For each initial density, initial shape and initial shape radius considered, the two-dimensional information gain is computed in moving from the prior to the posterior distributions, for two-statistic combinations that are consistently more informative. The information gains when r is varied are shown, in order to see any trends that emerge.

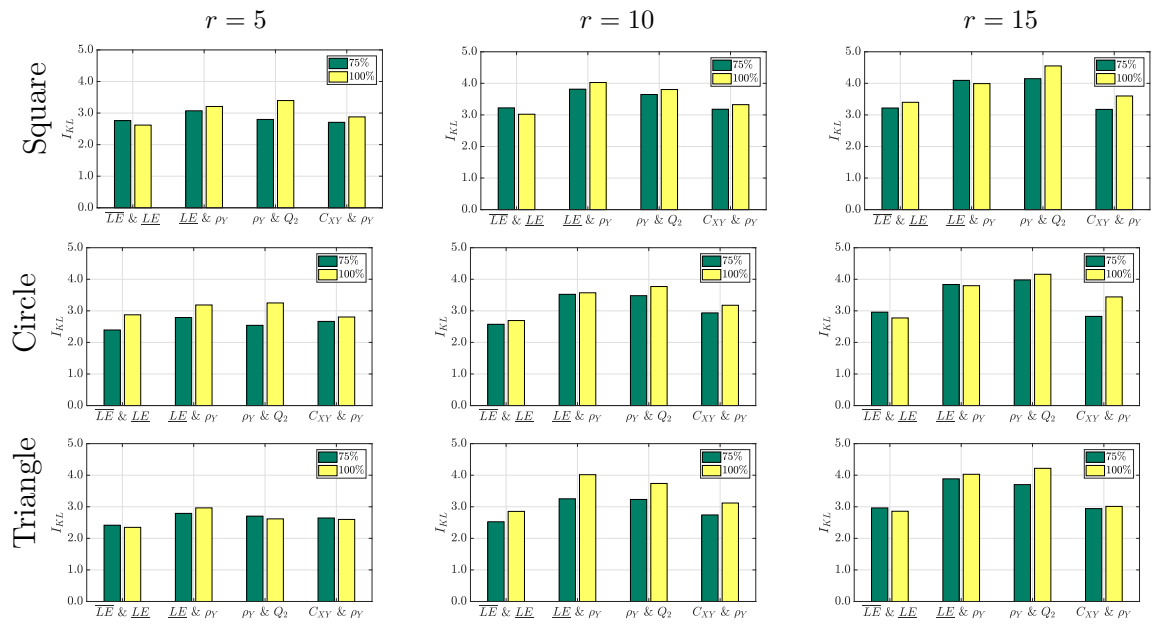


Figure 3.7: For each initial density, initial shape and initial shape radius considered, the two-dimensional information gain is computed in moving from the prior to the posterior distributions, for two-statistic combinations that are consistently more informative. The information gains when the degree of initial confluence is varied are shown, in order to see any trends that emerge.

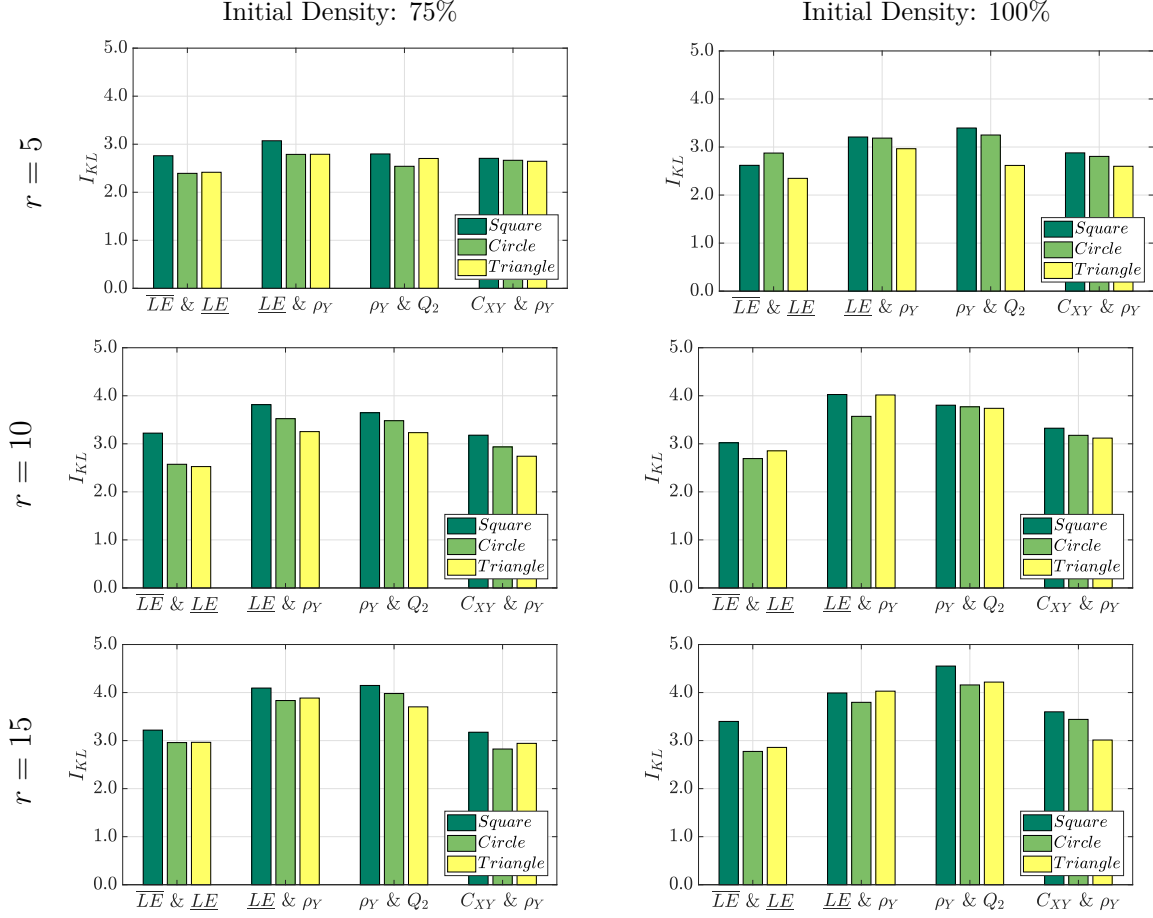


Figure 3.8: For each initial density, initial shape and initial shape radius considered, the two-dimensional information gain is computed in moving from the prior to the posterior distributions, for two-statistic combinations that are consistently more informative. The information gains when the shape of the domain is varied are shown, in order to see any trends that emerge.

3.3.4 Inference using experimental data

Now we investigate the identifiability of model parameters when using experimental data instead of synthetic data simulated from the model. The positions of cells at the start of the two-hour period when an electric field is applied are used to initialise simulations, and the positions of cells at the end of the two-hour period are used to compute experimental statistics. The experiment considered experiences a downwards bias, and the initial and final domains are the 1 hr and 3 hr domains in Figure 3.1. Without using regression adjustment, the resulting posterior distributions are shown in Figure 3.9.

Single-statistic posterior distributions when using experimental data are not centred on the same P_m and P_b values (see for example posteriors for ρ_Y and κ_8), suggesting that the model is not able to accurately capture the experimental data, or that posteriors are being calculated inappropriately (for example, if the distances between experiments and simulations are large). When considering combinations of statistics, it is possible to observe posterior distributions that overlap; for example the posterior from the statistic combination ρ_Y and Q_2 results in estimates of P_m and P_b of $P_m = 0.2511 \pm 0.0500$ and $P_b = 0.8774 \pm 0.0947$ for experiments with an upwards bias. These issues are discussed in the following section.

3.4 Discussion

Overall, synthetic data investigations have shown that information gain is maximised in experiments initialised with a large, square, fully confluent group of cells. However these results are potentially tainted through not keeping the number of cells fixed. It would be informative in a future investigation to consider different shapes that encompass either the same initial area, or the same initial number of cells. The new summary statistics introduced have been essential to estimate the bias parameter and

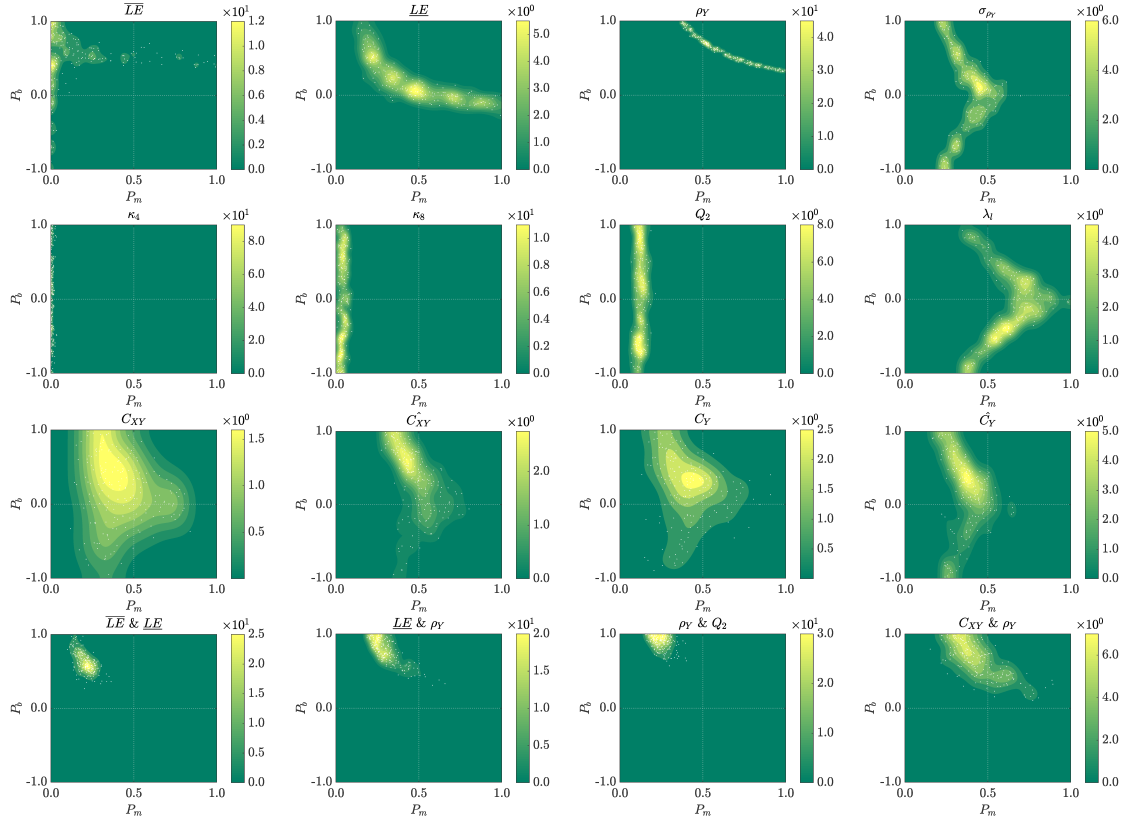


Figure 3.9: Posterior distributions resulting from various summary statistics (and combinations) are shown for experimental data. Simulations are initialised using the experimental domain at $t = 1$ hr, and experimental statistics are computed from the experimental domain at $t = 3$ hr, after two hours of being subjected to an electric field, as shown in Figure 3.1. In the first three rows, posterior distributions for single summary statistics are shown. Each posterior distribution corresponds to the statistic titled, where the statistic shorthand is used (see Table 3.1). The fourth row posterior distributions correspond to those generated when using a combination of statistics.

in general, the leading edge statistics and the position of the centre of mass prove very informative. The two-statistic combinations: the lower shape edge position, $\underline{LE}$ and the vertical position of the centre of mass, ρ_Y ; and ρ_Y and the binning variance, Q_2 , have been found to consistently lead to the highest amounts of two-dimensional information gain.

Investigations with experimental data have demonstrated that different statistics can lead to posteriors centred on different P_m and P_b values. To investigate this further, the distances, d_j^k , computed for each statistic, j , are ordered and plotted in Figure B.3. This reveals that some simulation statistics never match the corresponding experimental statistics. Note that this impacts the applicability of regression adjustment when using experimental data, and this is discussed further in Appendix B.2.1 and Appendix B.2.2.

It is possible that the model has been misspecified, so is unable to account for some aspect of the experimental data that is vital to ensure statistics of both are comparable. In this work, realism may be introduced by adding proliferation, however as there are only very few proliferation events, it is unclear if the addition of a proliferation mechanism will help or hinder inference attempts, as inference becomes more computationally demanding as the number of model parameters increases.

By reconsidering the experimental data, it is clear that we are not capturing a delay in the onset of a motility bias. From experimental images of a higher temporal resolution, a 10 to 15 minute delay can be seen before the mass migration of cells in the vertical direction, as shown in Figure 3.10.

As experiments are only considered for two hours, this onset delay could be important to capture, and we recommend that future work focus on addressing this, potentially estimating the length of delay using images of sufficiently high temporal resolution, or attempting to infer it as an additional model parameter. It may also be the case that the model cannot accurately capture the behaviour of cells in highly

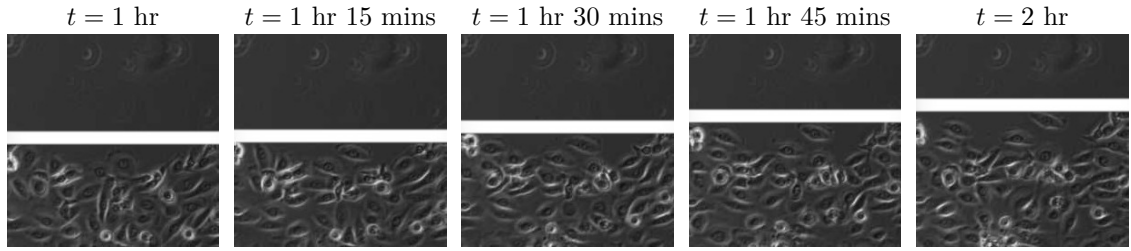


Figure 3.10: Example of experimental data demonstrating delay in onset of motility bias. Approximate leading edges in each image are overlayed. Approximately 15 minutes after the electric field is introduced (at $t = 1$ hr), cells begin to move upwards.

packed environments, i.e. the volume exclusion mechanism used to restrict movement in the model may be inappropriate to fully capture contact inhibition of locomotion [115].

In this chapter we have applied Bayesian inference methods to examine experimental data of cell spreading under the influence of an electric field. In the next chapter, we continue these investigations in a model that has not been previously used for inference purposes, that accounts for the mechanics of cell growth and contact with neighbouring cells.

Chapter 4

Bayesian inference applied to a force-based off-lattice model

4.1 Introduction

In the previous chapters we have parameterized simple on-lattice models accounting for mechanisms of motility, proliferation and motility bias. In general, however, models used to understand cell spreading processes will vary in complexity, and so it is important to understand how successful the inference methods are when applied to more complex models. By increasing the complexity of a model in a gradual way, it is possible to understand when difficulties arise, and develop ways to address concerns before further model development.

In this chapter, we consider the data gathered by Jin et al. [36], shown in Figure 4.1, where time-lapse images are taken from scratch assays of PC-3 prostate cancer cells. In that work, one-dimensional cell density data is extracted from these images by manually counting cells in $50\text{ }\mu\text{m}$ columns, before linking these data to a one-dimensional Fisher-Kolmogorov model [28]. The model parameters inferred are sensitive to the initial degree of confluence, however this sensitivity decreases when

adopting a Porous-Fisher model [82]. The most appropriate model for these experimental data is therefore unclear, however it is possible that for a different model, parameterization is not as sensitive to the initial conditions.

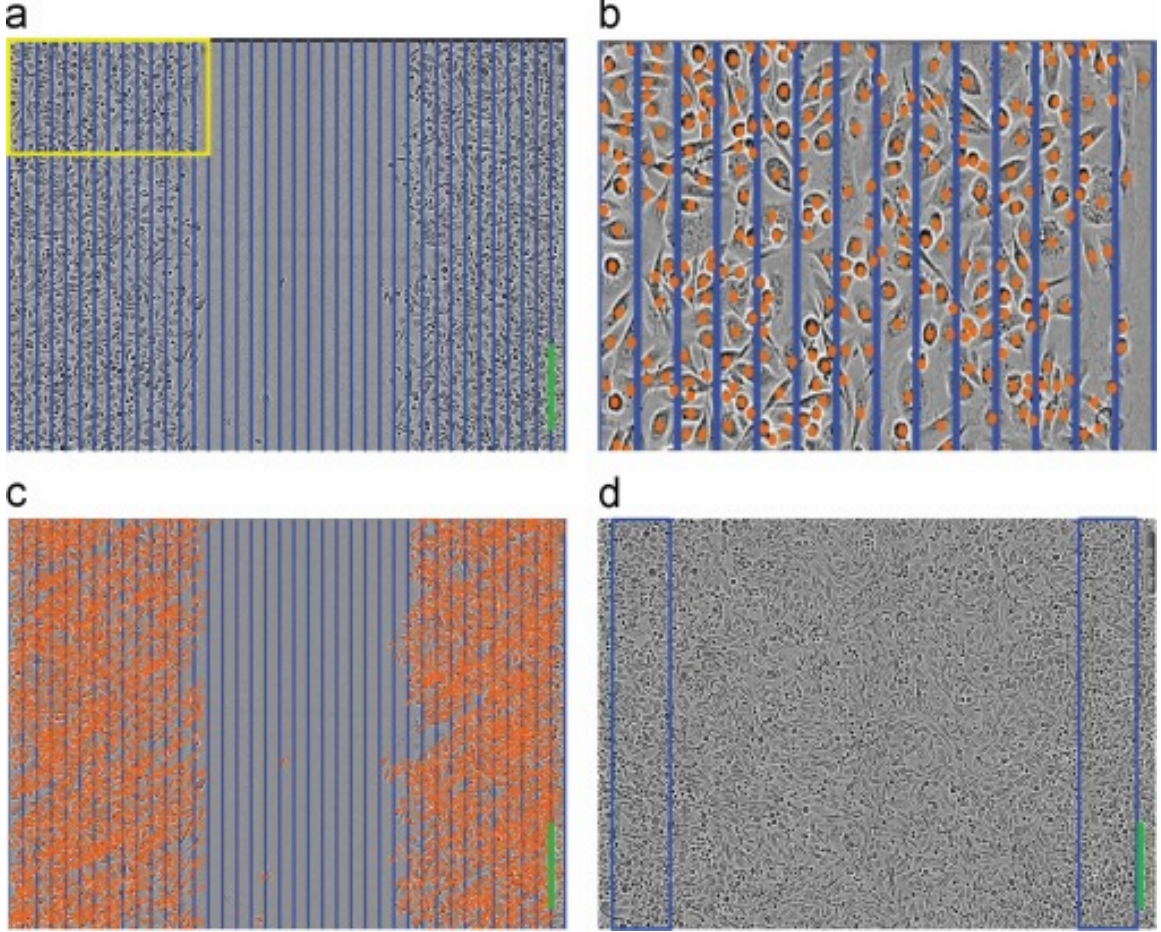


Figure 4.1: The method to analyse images from experiments in Jin et al. [36]. Scale bars in each image are $300\mu\text{m}$. PC-3 prostate cancer cells are imaged growing to confluence after a scratch, and cells are counted in those images to arrive at a cell density. (a) Focus on subregions of the domain for purposes of manual cell counting. (b) Each cell is labelled with a red marker in the subregion. (c) Repeat manual cell counting for the whole image, to obtain a cell density estimate at 38 uniformly distributed positions between $0\mu\text{m}$ and $1900\mu\text{m}$. This is repeated for three experimental replicates, and for six time-points. (d) The region behind the scratch is used to estimate the carrying capacity density, q^{max} , the maximum density for a monolayer of cells. Reproduced from Jin et al. [36] with kind permission.

The main mechanisms driving cell spreading in scratch assay experiments are motility and proliferation, however in the experimental data considered, PC-3 cells remain in close proximity for the entire duration, suggesting that an appropriate

model should include a mechanism for adhesion. Although it is relatively simple to account for proliferation and adhesion effects in crowded on-lattice cell spreading models, it is potentially more realistic to consider a model derived from the underlying mechanics involved in these experimental data.

We consider the one-dimensional model of Murray et al. [57], where a series of springs are attached to one another at node points. Cells move according to the forces they are subjected to by their surrounding cells, as shown in Figure 4.2. Using this formulation, cells can be thought of as springs [29], or cell centres can be seen as node points [57]. We adopt the former approach, as the proliferation mechanism adopted lends itself more naturally to springs describing cells.

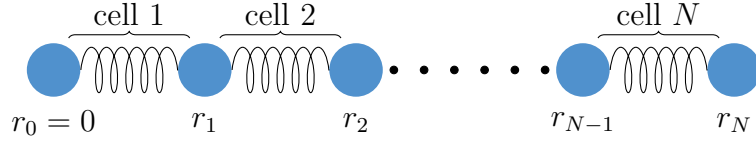


Figure 4.2: Schematic of the cell-based model, where springs represent cells, and nodes represent the point where two cells touch. There are N cells, and node positions are denoted by r_i , for $i \in \{0, \dots, N\}$, with the left boundary of the first cell fixed at $r_0 = 0$.

A stochastic proliferation mechanism is introduced into this model to account for the proliferation that occurs in the experimental data. In Figure 4.3, we show the proliferation mechanism adopted, and a proliferating cell splits in two from its centre.

Statistics traditionally extracted from scratch assay experiments are cell densities, leading edge positions or number of cells over time [36, 39]. We have used the experimental cell density data shown in Figure 4.1 from Jin et al. [36] in this Chapter.

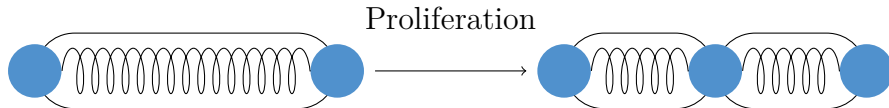


Figure 4.3: We introduce a proliferation mechanism into the off-lattice one-dimensional line of cells model (see Figure 4.2), such that when a cell proliferates, it divides into two equally sized cells, creating a new node point (nodes denote the join of two cells) at the previous cell centre.

From the cell density, it is possible to extract the leading edge position and number of cells statistics.

The method used in this Chapter to perform inference is slightly different that used previously. In previous chapters, ABC rejection or ABC-DC were the inference methods adopted, but here the model introduced is more computationally expensive to simulate than those used previously, so we adopt the more efficient inference method, approximate Bayesian computation - sequential Monte-Carlo (ABC-SMC) [92, 100]. ABC-SMC works by iteratively converging to a posterior distribution by accepting parameters leading to areas of higher posterior density, and has been used effectively in applications including dynamical systems, cell migration and in model selection [42, 51]. ABC-SMC is faster than ABC rejection but harder to customise, however, there are software packages available that implement ABC-SMC, for example through ABC-Sysbio [50].

4.2 Methods

Here we outline the one-dimensional model we explore in this chapter, the experimental data under consideration, and the inference procedure used to compute posterior distributions of model parameters.

4.2.1 Model

We adopt the one-dimensional cell spreading model first introduced by Murray et al. [57], where a series of springs are attached to one another at node points. Cells are represented in this model as the springs, so that the relaxation of each spring describes a cell returning to its natural resting length, a , and node positions are the points where two cells meet. We assume that cells never detach from one another, either due to a tissue-level compression of the cells or because the cells adhere to each

other.

A model can be formulated using the positions of each node via the indexing variable, i , describing the index of node i , as in Figure 4.2 where subscripts denote indices. In the case where there are N cells, there are $N + 1$ nodes; the left-most node has index $i = 0$ (the left-hand edge of the first cell), and the right-most node has index $i = N$. The positions of node i is given by r_i , for $i = 0, \dots, N$. Accounting for the forces on each node using Newton's second law allows arrival at the equation

$$m_i \frac{d^2 r_i}{dt^2} = \sum_{j \neq i} F_{i,j}^{int} + F_i^{visc}, \quad \text{for } i = 1, \dots, N, \quad (4.1)$$

where the interaction force between cells i and j is $F_{i,j}^{int}$, and the viscous force acting on the i^{th} node is F_i^{visc} .

Assume that the viscous force is composed of drag proportional to velocity with constant η , and that cells move in an over-damped regime, so the acceleration term of the left-hand side of equation (4.1) can be neglected. Further assume that as cells are in a one-dimensional line, they only interact with their adjacent neighbours and they never swap positions, so they necessarily remain in sequential order, $r_0 < r_1 < \dots < r_{N-1}(t) < r_N(t)$, so that,

$$\begin{aligned} r_0 &= 0, \\ \eta \frac{dr_i}{dt} &= F_{i,i-1}^{int} + F_{i,i+1}^{int}, \quad \text{for } i = 1, \dots, N-1, \\ \eta \frac{dr_N}{dt} &= F_{N,N-1}^{int}, \end{aligned} \quad (4.2)$$

where η is a drag parameter between the cells and their substrate. Only adjacent cells interact, therefore the right-hand node is free to move into the unbounded space to its right. The force law is chosen so that cells move toward a resting cell length, a .

A proliferation mechanism can be introduced whereby a proliferating cell splits

in two from its centre, as in Figure 4.3. To align this mechanism with the model introduced above, we re-index nodes whenever a cell proliferates, ensuring $r_0 < r_1 < \dots < r_{N(t)-1}(t) < r_{N(t)}(t)$. The full model and the re-indexing that occurs at a proliferation event is demonstrated in Figure 4.4

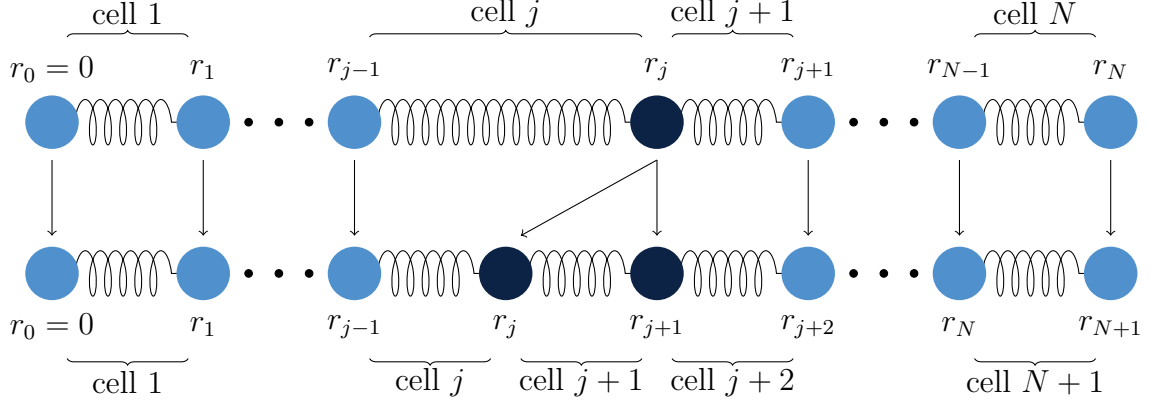


Figure 4.4: Proliferation of cell j causes a new node to be introduced at its centre, and cells are then relabelled to maintain their order ($r_i < r_{i+1}$ for $i \in \{0, \dots, N\}$).

We elect to adopt the simplest force law possible, a linear force law. Other potential force laws are considered in Chapter 5, however there is no consensus for the force law appropriate to cell spreading data. The force exerted by node i due to cell j is therefore,

$$F_{i,j}^{int} = k(a - (r_i - r_j)),$$

where k is the relaxation constant. If we define $\alpha = k/\eta$, we simplify the resulting model, and avoid potential structural non-identifiability of model parameters [21]. Now α represents the ratio of relaxation to drag in the domain, and describes relatively how quickly cells relax to equilibrium. It is therefore simply coined as a motility parameter.

It is possible to extend the equations (4.2) following the introduction of a proliferation mechanism to derive an equation describing the evolution of the cell density, and this analysis is shown in Chapter 5. In this Chapter, node positions are simulated

by solving equations (4.2) in increments of a small discrete time-step $dt=0.001$, using a forward Euler method.

The forward Euler method is implemented with self-written code using **MATLAB** 2019a. Using a desktop Windows 10 PC with an Intel Core i5-3570 processor (3.40 GHz) and 16GB RAM. A simulation with $N = 45$ cells initially, $a = 1$ and $\alpha = 15$, until final time $T = 100$, took 0.829 seconds to run.

We adopt a proliferation scheme where each cell proliferates at a constant rate, β , per unit time. The implicit assumption with this stochastic proliferation is there is no tracking of cell age, and that age plays no part in the likelihood of a cell to proliferate, different proliferation laws are considered in the next chapter. The proliferation mechanism is implemented by checking after each time interval, $[t, t+dt)$, whether a random number, $r \sim U(0, 1)$, satisfies $r < N(t)\beta dt$. If this condition is met, a randomly chosen cell is chosen to undergo a proliferation event: the chosen cell splits in two from its centre, creating a new cell edge (node point) at this position, as shown in Figure 4.4, and causing all nodes to the right of the proliferating node to be relabelled. This proliferation scheme only allows one proliferation to occur per time-step, introducing an error of order $\mathcal{O}(dt^2)$ in the model, however β , dt and N are chosen such that $N\beta dt \ll 1$.

4.2.2 Experimental data

Here we outline how to extract data from experimental images and how that data can be input into the IBM or compared with synthetic data obtained from the IBM. We consider a subset of the experimental scratch assay density data from Jin et al. [36]. In the scratch assay experiments, a vertical strip of cells is scraped off the central portion of a confluent layer of PC-3 prostate cancer cells. Images of the experimental domain of size $1.9 \text{ mm} \times 1.43 \text{ mm}$ are captured at five different time points, between 0 hr and 48 hr. The number of cells in the dimension perpendicular to the scratch is

recorded by counting cells within columns of width $50 \mu\text{m}$, between $0 \mu\text{m}$ and $1900 \mu\text{m}$, as shown in Figure 4.1. The number of cells are manipulated to give the average cell density from three repeated experiments, for six different initial cell numbers, between 10,000 cells and 20,000 cells. The resulting average cell densities are shown in Figure 4.5.

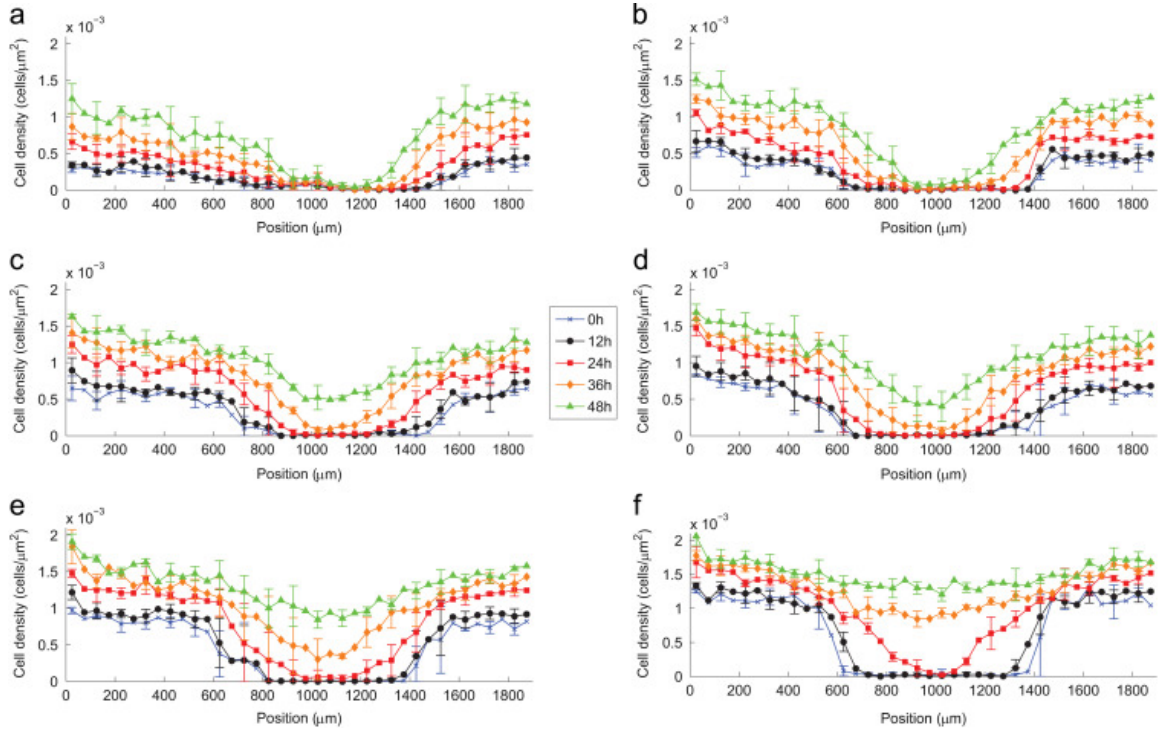


Figure 4.5: The scratch assay cell density data, reproduced from Jin et al. [36] with kind permission. Cells are counted in images by dividing the domain into $50 \mu\text{m}$ columns, as shown in Figure 4.1. The resulting average cell density profiles are shown here with error bars (the standard deviation over three experimental repeats), at five times, 0 hr, 12 hr, 24 hr, 36 hr and 48 hr, and for six different initial number of cells, (a) 10,000, (b) 12,000, (c) 14,000, (d) 16,000, (e) 18,000 or (f) 20,000.

4.2.3 Statistics

Comparison of the output of the computational model and experimental data is done by computing statistics from both sources. The statistics computed are the number of cells, N , the position of the leading edge, LE , and the average cell density, q . The computational model records positions of node points, whereas the data from

Jin et al. [36] provides the average density of cells at the points $0 \mu\text{m}$, $25 \mu\text{m}$, $75 \mu\text{m}$, $\dots$, $1825 \mu\text{m}$, $1875 \mu\text{m}$, $1900 \mu\text{m}$, therefore it is necessary to establish algorithms to convert from node positions to cell densities and vice versa, so these data can be compared. These algorithms are outlined in Appendix C.1, are visualised in Figure 4.6, and ensure that features are extracted from experimental data for only one-half of a scratch assay, and only while the cell density is high enough so the assumption of cell-cell contact is reasonable.

4.2.3.1 Comparing cell densities

Henceforth we use subscripts *exp* and *sim* to differentiate between experimental and simulation statistics. Using the methods in Appendix C.1.1 we can estimate the cell density, q_{sim} , at any position $x \in [0, LE_{sim}]$. Further, using the methods in Appendix C.1.3 we can estimate the cell density, q_{exp} at any position in $x \in [0, LE_{exp}]$.

In order to compare q_{exp} and q_{sim} we must interpolate both cell densities onto the same domain. It is most logical to interpolate onto the $x \in [0, LE_{exp}]$ domain, as the x positions at which experimental data are recorded is fixed. When computing the deviation between q_{exp} and q_{sim} , we must also consider how many positions to interpolate the cell densities at.

For simplicity, we interpolate using the MATLAB `interp1` function on a mesh with eleven space points evenly distributed in $x \in [0, LE_{exp}]$. If the cell density, q_{sim} , is not defined at any of the eleven space points, it is given a value equal to the equilibrium value there, $1/a_{sim}$, where a_{sim} is the resting spring length used in the simulation generating cell density q_{sim} .

Finally, we note that in the case where multiple replicates are used, the experimental leading edge is the average of all experimental leading edges, $LE_{exp} = \overline{LE_{exp,rep}}$ and so all subsequent computations of the cell density, q , are interpolated onto $r \in [0, LE_{exp}]$ before averaging, $q_{exp} = \overline{q_{exp,rep}}$

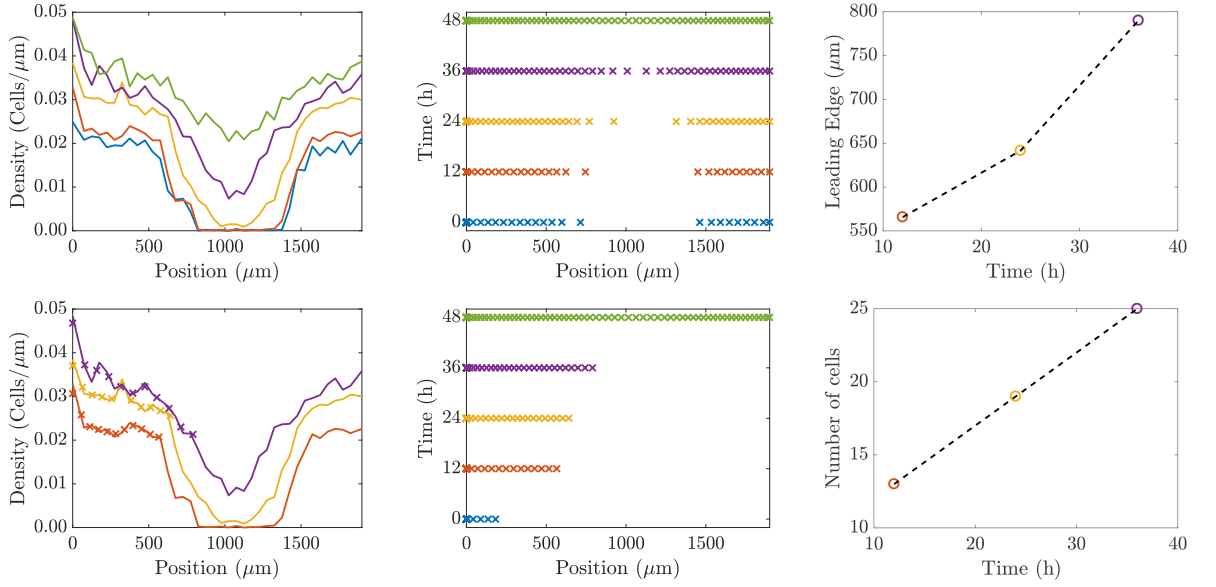


Figure 4.6: The experimental data (top-left) is converted to node positions via the algorithm in Appendix C.1 (top-middle). A subset of the resulting node positions are considered to give a single advancing wave-front. The resulting positions are used to compute experimental data in the form of a density, (bottom-left, crosses, shown with comparison to original densities) a leading edge position (top-right), or a number of cells (bottom-right). We only consider data where the two sides of the scratch assay have not met, and where the cell density is at least half the carrying capacity, $q^{max} = 1.7 \times 10^{-3} \text{ cells} / \mu\text{m}^2$.

4.2.4 Inference - ABC-SMC

In this section we outline the algorithm we use for inference, ABC-SMC, which has been applied in many areas of systems biology [92]. The algorithm iteratively converges to the posterior distribution, by gradually reducing the tolerance between experimental and simulation statistics. By sampling parameters from previously accepted population parameters, the algorithm retains parameters resulting in simulations that most closely match experiments.

4.2.4.1 Algorithm

Starting with a prior distribution of parameters, θ , we seek to compute the posterior distribution of parameters of the model given experimental data, D . Bayes theorem provides a means of doing this, namely the relationship, $\pi(\theta|D) = \pi(D|\theta)\pi(\theta)$, where $\pi(\theta)$ is the prior distribution, $\pi(\theta|D)$ is the posterior distribution, and $\pi(D|\theta)$ is the likelihood of observing experimental data given model parameters.

Approximate Bayesian inference methods utilise Bayes theorem by converging on the likelihood function rather than the posterior distribution, as it is simpler to estimate computationally. We use ABC-SMC with summary statistics, S , so that instead of attempting to find the likelihood of the full experimental data, D , we can instead reduce the dimensionality of D and deal with a more tractable set of data, $S(D)$.

We therefore make the assumption that $\pi(\theta|D) \approx \pi(\theta|S(D))$, which is acceptable when the statistics under consideration are sufficient. Determining sufficiency of statistics is an ongoing area of research [6, 69], but not further considered here.

We demonstrate the ABC-SMC algorithm in Algorithm 3, with the following definitions:

- M is the number of populations;

- L is the number of parameters accepted in each population;
- $\mathcal{K}$ is a perturbation kernel for adjusting sampled parameters;
- γ is the percentile value used to determine the distance threshold for subsequent populations ¹;
- d is the distance function defined in Section 4.2.4.2.

In all that follows, $\mathcal{K}$ is a multivariate normal distribution with zero mean and standard deviation based on the previous accepted population of parameters and an optimised acceptance rate [27, 84].

The ABC-SMC algorithm results in a sequence of populations of θ and weights w for each population m . The final population $m = M$ corresponds to accepted parameters with the smallest distance between simulation and experiment, which are used to generate a posterior distribution. The algorithm is implemented using self-written code in MATLAB.

4.2.4.2 Distance calculation and information gain

We record the following statistics at multiple times during simulations, $t_1, \dots, t_Z$, where Z is the number of time points considered. For experimental data, we choose $Z = 2$, with $t_1 = 12$ hr, $t_2 = 24$ hr. Denote the duration of the simulation by $T \equiv t_Z$. The statistics considered are as follows:

1. cell density interpolated at 11 points, $q_{j,z} = q(\tilde{r}_j, t_z)$, for $j = 0, 1, \dots, 10$, $\tilde{r}_j = LE_{exp} \times j/10$;
2. leading edge position, $LE_z = r_{N(t_z)}(t_z)$;
3. number of cells, $N_z = N(t_z)$.

¹We use a percentile value rather than pre-defining a sequence of decreasing tolerances, $\epsilon = \epsilon_1, \dots, \epsilon_M$ for simplicity.

Algorithm 3 Approximate Bayesian computation - sequential Monte Carlo algorithm

repeat

 Sample parameters from the prior distribution, θ^* . These samples are known as particles.

 Simulate a data set, D^* , from the model with parameters, θ^* .

 Compute the distance between summary statistics obtained from simulated data, $S(D^*)$, and those obtained from experimental data, $S(D)$, using distance function d .

if $d(S(D^*), S(D)) < \epsilon_1$ **then**

 Accept particle θ^* into population 1.

end if

until L particles have been accepted.

Particles accepted in population 1 have uniform weights, $w_1 = 1$.

for $m = 2, \dots, M$ **do**

 Set ϵ_m to be the γ percentile of the previous distances of accepted parameters.

repeat

 Simulate a data set, D^* , from the model with parameters sampled from population $m - 1$ with weights w_{m-1} and perturbed with perturbation kernel $\mathcal{K}$;

 Compute the distance between summary statistics obtained from simulated data, $S(D^*)$, and those obtained from experimental data, $S(D)$, using distance function d ;

if $d(S(D^*), S(D)) < \epsilon_m$ **then**

 Accept particle θ^* into population m ;

 Compute the weight for the accepted particle as $w = \pi(\theta^*) / \sum_{l=1}^L (w_{m-1}^l * \mathcal{K}(\theta^* | \theta_{m-1}^l))$;

end if

until L particles have been accepted.

 Normalise the weights w_m so they sum to unity.

end for

The experimental statistics are computed at each time point and used to calculate median absolute deviation statistics, $\sigma(S_{z,exp})$, and are utilised in the distance function to weight different statistics equally, therefore allowing for combining of statistics without bias. The distance function is computed for each statistic as follows,

$$\begin{aligned} d_z^q &= \frac{q_{j,z} - q_{j,z,exp}}{\sigma(q_{j,z,exp})\sqrt{Z \times 11}}, \\ d_z^{LE} &= \frac{LE_z - LE_{z,exp}}{\sigma(LE_{z,exp})\sqrt{Z}}, \\ d_z^N &= \frac{N_z - N_{z,exp}}{\sigma(N_{z,exp})\sqrt{Z}}. \end{aligned}$$

After computing the distance for each time point, compute the overall distance for the simulated statistic, $S \in \{q, LE, N\}$, via,

$$d^S = \sqrt{\sum_z (d_z^S)^2}. \quad (4.3)$$

When combining statistics, simply compute the normed distance, for example,

$$d^{q\&LE} = \sqrt{\sum_z ((d_z^q)^2 + (d_z^{LE})^2)}. \quad (4.4)$$

These distance functions are used in the ABC-SMC algorithm 3. The final population corresponds to the parameters, θ , resulting in the smallest distances between experiment and simulation statistics, are used to approximate a posterior distribution. The posterior distribution is approximated via kernel density estimation, which computes the sum of multivariate Gaussian kernels centred on each accepted parameter, to form a smoothed distribution. As θ consists of three parameters (a, α, β) , the bandwidth (standard deviation) for the multivariate Gaussian kernel is calculated via Silverman's

rule,

$$h_i = \hat{\sigma}_i^5 \left(\frac{4}{5 \times L} \right)^{\frac{1}{7}}, \text{ for } i = 1, 2, 3,$$

where L is the number of samples and $\hat{\sigma}_i$ is the standard deviation of the i^{th} variate. [85]

Using this bandwidth and the MATLAB function `mvksdensity` (for multi-dimensional kernel density estimation), we arrive at a three-dimensional posterior approximated on a fine mesh with 50 grid points in each dimension, which can be integrated numerically to obtain two- and one-dimensional posterior distributions, along with the approximate information gain of each posterior, when moving from the prior distribution to the posterior distribution. Information gain is computed via the Kullback-Leibler divergence,

$$I_{KL} = \int_{\theta} p(\theta|D) \log \left(\frac{p(\theta|D)}{p(\theta)} \right) d\theta,$$

akin to computations when performing inference on the two-dimensional on-lattice model in Chapter 3. Note that I_{KL} can be computed for the full posterior distribution, or marginal posterior distributions, to quantify how informative a statistic is of a specific parameter or a pair of parameters.

4.2.4.3 Inference particulars

The experimental data used is of cell densities at two times, $t_1 = 12$ hr and $t_2 = 24$ hr, so we consider synthetic experiments with $Z = 2$ and $T = 24$ hr initially, before considering a series of experimental protocols that may help improve inference. The initial inference protocols are shown in Table 4.1, and describe cells with a cell width, motility rate, and proliferation rate analogous to the experimental data considered later.

Table 4.1: The parameter choices adopted in initial inference procedures, influenced by the experimental data considered. Any deviation from these values in Section 4.3 are clearly outlined.

Description	Symbol	Value
Number of populations of ABC-SMC	M	6
Number of particles in each population	L	200
Tolerance scheme quantile	γ	0.5
Prior a	$\pi(a)$	$\mathcal{U}(15\mu\text{m}, 50\mu\text{m})$
Prior α	$\pi(\alpha)$	$\mathcal{U}(0.01 \mu\text{m}^2/\text{hr}, 10 \mu\text{m}^2/\text{hr})$
Prior β	$\pi(\beta)$	$\mathcal{U}(0\text{hr}^{-1}, 0.04\text{hr}^{-1})$
Experimental a	a_{exp}	25 μm
Experimental α	α_{exp}	5 hr^{-1}
Experimental β	β_{exp}	0.02 hr^{-1}
Force law	F_i	Linear, see Table 5.3
Proliferation law	G_i	Constant, see Table 5.4
Length of experiment	T	24 hr
Number of time points	Z	2
Number of replicates	-	10
Initial number of cells	N_0	30
Initial length of domain	R	750 μm

To estimate typical values that α may take in order to estimate a prior distribution, we consider the equation from Murray et al. [58], that determines a relationship between a population-level diffusion coefficient, D , and the motility parameter, α . For a linear force law, this relationship is $D = \alpha/q^2$, where q is the density of cells in the domain, which (at equilibrium) is equal to $1/a$. Therefore, $\alpha \approx D/a^2$, hence, we can estimate α using previously obtained values of D from the work of Jin et al. [36] on prostate cancer PC-3 cells, as $D \approx 625 \mu\text{m}^2/\text{hr}$, $a \approx 25 \mu\text{m}$, hence $\alpha \approx 1$.

4.3 Results

In this section parameter inference investigations are carried out on both synthetically created data, and experimental data from Jin et al. [36]. Inference of model parameters using synthetic data establishes the identifiability of parameters of the model. Further, considering a variety of synthetic experiments allows us to determine the most informative experimental design to determine each model parameter. Following on from these synthetic experiments, inference methods are applied on experimental data to find model parameters for a fixed force and proliferation law.

4.3.1 Inference using synthetic data: experimental design

In this section we consider how experimental design can be improved to increase the accuracy of parameter estimates obtained from ABC-SMC, quantified using the Kullback-Leibler divergence. There are numerous experimental adjustments that can be made, from the simple, such as changing the duration of experiment, to the complex, such as manipulating the experimental domain. To understand the effectiveness of each design choice, synthetic experimental data is computed, and ABC-SMC is applied to estimate model parameters in each design, and the Kullback-Leibler divergence values are compared with those from a ‘control’ experimental design.

The ‘control’ experimental design is of synthetic data generated according to the parameters in Table 4.1, where we investigate if we can recover model parameters using either cell density data, q , leading edge data, LE , or number of cells data, N . We try to infer the spring constant, α , resting spring length, a , and proliferation parameter, β , from an initially uniformly distributed line of cells.

We display the experimental data, the one- and two-dimensional posterior distributions, and the information gained in moving from the prior distribution to the posterior distribution in Figure 4.7. We consider the simple case with final time

$T = 24$ hr, and number of time points $Z = 2$, matching conditions in experimental data. We find one-dimensional posterior distributions are centred around true values for the resting spring length, a , using either statistic q or LE , and the proliferation parameter, β , using statistic N . In Appendix C.2.1 we determine that using a distance function that is a combination of statistics does not considerably increase the information gained about each parameter. Therefore we opt not to consider combinations of statistics for the other inference experiments considered.

4.3.1.1 More time points

We now perform the same investigation with data recorded for $Z = 4$ time points instead of $Z = 2$, so we record the positions of nodes at times $t = 6$ hr, 12 hr, 18 hr, 24 hr. We expect this experimental modification to be relatively straight-forward to implement in practice. In Figure 4.8 we can see that the information gain has increased marginally over the base experiment for all parameters and every statistic considered, meaning better estimates of the parameters are likely to be gained from an experiment with more time points. A simple recommendation from this result is to measure each statistic with a higher temporal resolution where possible. However, one should take care not to increase the temporal resolution too greatly, else the data becomes highly dimensional, and as data from each time point is equally weighted, we may be introducing excess data, and instead decrease the information gain.

4.3.1.2 A longer experiment

We now investigate the extent to which conducting a longer experiment changes the information obtained. We keep all experimental parameters consistent except now measure statistics at times $t = 24$ hr and $t = 48$ hr. Results shown in Figure 4.9 demonstrate that when considering a longer experiment, significantly more information is obtained about the proliferation parameter, β , seemingly at the cost of

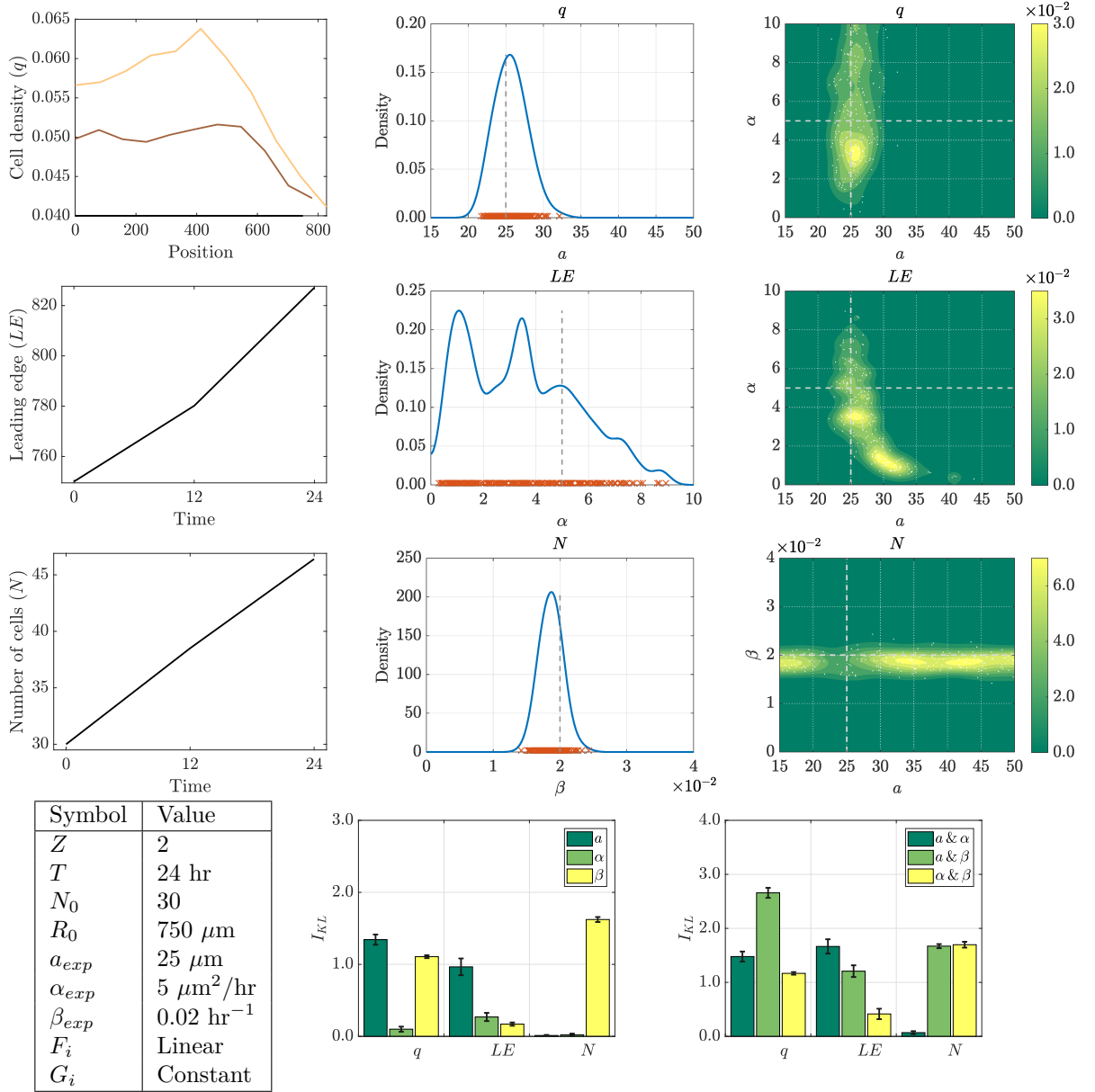


Figure 4.7: We perform ABC-SMC with synthetic data computed using the experimental parameters shown. Prior distributions used and other inference choices are outlined in Table 4.1. We display the synthetic experimental data in the left-most plots, for either the cell density, the leading edge or the number of cells. The central plots show the marginal one-dimensional posterior distributions for parameters a , α and β computed using statistics q , LE or N , which generally provide the most one-dimensional information gain. The right-most plots show marginal two-dimensional posterior distributions for some two-dimensional parameter combinations. The information gains computed from each marginal one-dimensional posterior distributions or marginal two-dimensional posterior distributions are shown for each statistic in the bar plots, where each information gain is shown with error-bars computed from three independent runs of ABC-SMC.

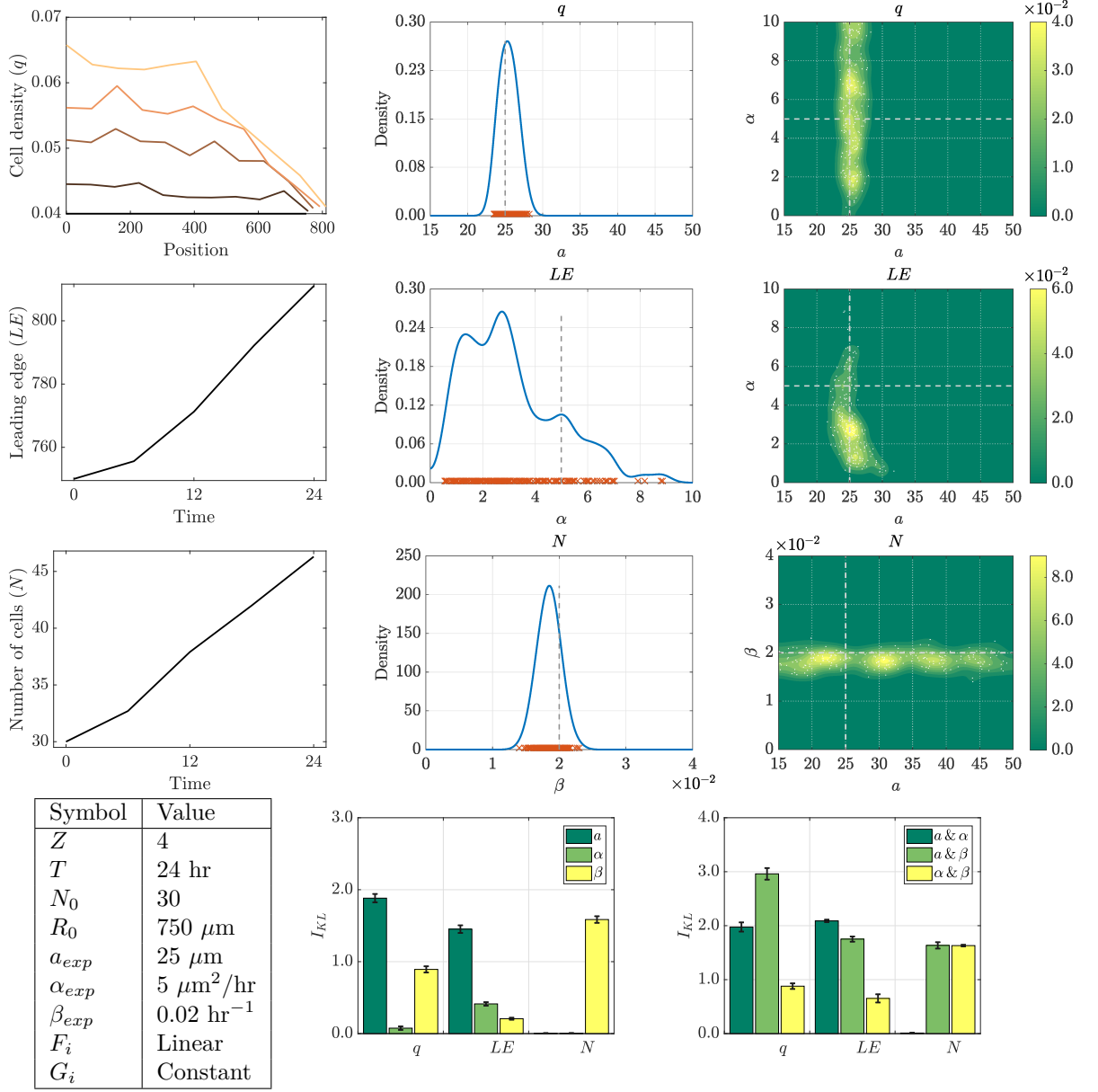


Figure 4.8: We perform ABC-SMC with synthetic data computed using the experimental parameters shown. Prior distributions used and other inference choices are outlined in Table 4.1. We display the synthetic experimental data in the left-most plots, for either the cell density, the leading edge or the number of cells. The central plots show the marginal one-dimensional posterior distributions for parameters a , α and β computed using statistics q , LE or N , which generally provide the most one-dimensional information gain. The right-most plots show marginal two-dimensional posterior distributions for some two-dimensional parameter combinations. The information gains computed from each marginal one-dimensional posterior distributions or marginal two-dimensional posterior distributions are shown for each statistic in the bar plots, where each information gain is shown with error-bars computed from three independent runs of ABC-SMC. Comparison to Figure 4.7 clearly shows the changes in information gain when using $Z = 4$ time points instead of $Z = 2$ time points.

information about the spring length, a , (and to a lesser extent the motility parameter, α).

Over a long time scale, proliferation dominates the evolution of the tissue, so it is unsurprising that a more precise β estimate can be obtained using a longer experiment. A recommendation from these results is that in order to obtain precise estimates for both a and β , a longer duration experiment should be considered, with images captured more frequently initially (to estimate a), and less frequently later (to estimate β). For example we could consider an experiment with time points $t = 6$ hr, 12 hr, 18 hr, 24 hr and 48 hr.

4.3.1.3 Different initial cell numbers

By keeping the initial cell density consistent, we can investigate how increasing or decreasing the number of cells initially impacts the inference of model parameters. In Figure 4.10 it is clear that for lower cell numbers, more information is obtained about the spring length, a , from the cell density statistic, q . We expect this is due to the fact that, for smaller cell numbers (and hence a smaller domain), deviations in q are more easily captured than when there are larger cell numbers. We also note that for higher cell numbers we obtain more confidence in the proliferation parameter, β , from the number of cells statistic, N , which is unsurprising as for larger initial numbers of cells we see more proliferation events.

Comparing this investigation with those already considered suggests that the information gained from model parameters is maximised when considering a range of time-steps and experiment durations, rather than a variety of initial cell numbers.

4.3.1.4 Initially confined cells

Instead of initialising with cells initially at their resting lengths, we could consider an experiment that stretches or compresses a line of cells at confluence. Although simple

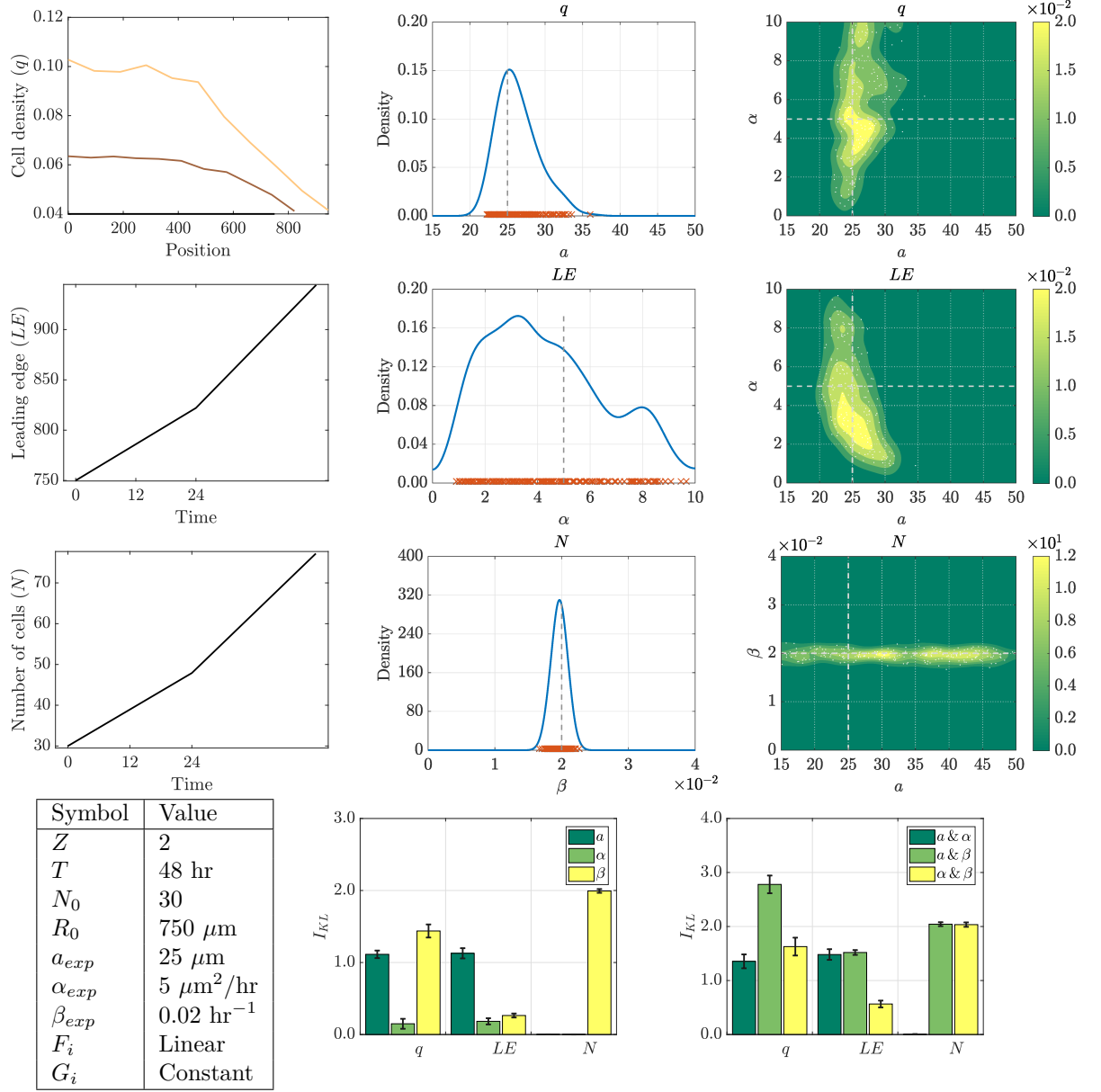


Figure 4.9: We perform ABC-SMC with synthetic data computed using the experimental parameters shown. Prior distributions used and other inference choices are outlined in Table 4.1. We display the synthetic experimental data in the left-most plots, for either the cell density, the leading edge or the number of cells. The central plots show the marginal one-dimensional posterior distributions for parameters a , α and β computed using statistics q , LE or N , which generally provide the most one-dimensional information gain. The right-most plots show marginal two-dimensional posterior distributions for some two-dimensional parameter combinations. The information gains computed from each marginal one-dimensional posterior distributions or marginal two-dimensional posterior distributions are shown for each statistic in the bar plots, where each information gain is shown with error-bars computed from three independent runs of ABC-SMC. Comparison to Figure 4.7 clearly shows the changes in information gain when using $T = 48$ hr instead of $T = 24$ hr.

Symbol	Value
Z	2
T	24 hr
N_0	20
R_0	500 μm
a_{exp}	25 μm
α_{exp}	5 $\mu\text{m}^2/\text{hr}$
β_{exp}	0.02 hr^{-1}
F_i	Linear
G_i	Constant

Symbol	Value
Z	2
T	24 hr
N_0	30
R_0	750 μm
a_{exp}	25 μm
α_{exp}	5 $\mu\text{m}^2/\text{hr}$
β_{exp}	0.02 hr^{-1}
F_i	Linear
G_i	Constant

Symbol	Value
Z	2
T	24 hr
N_0	40
R_0	1000 μm
a_{exp}	25 μm
α_{exp}	5 $\mu\text{m}^2/\text{hr}$
β_{exp}	0.02 hr^{-1}
F_i	Linear
G_i	Constant

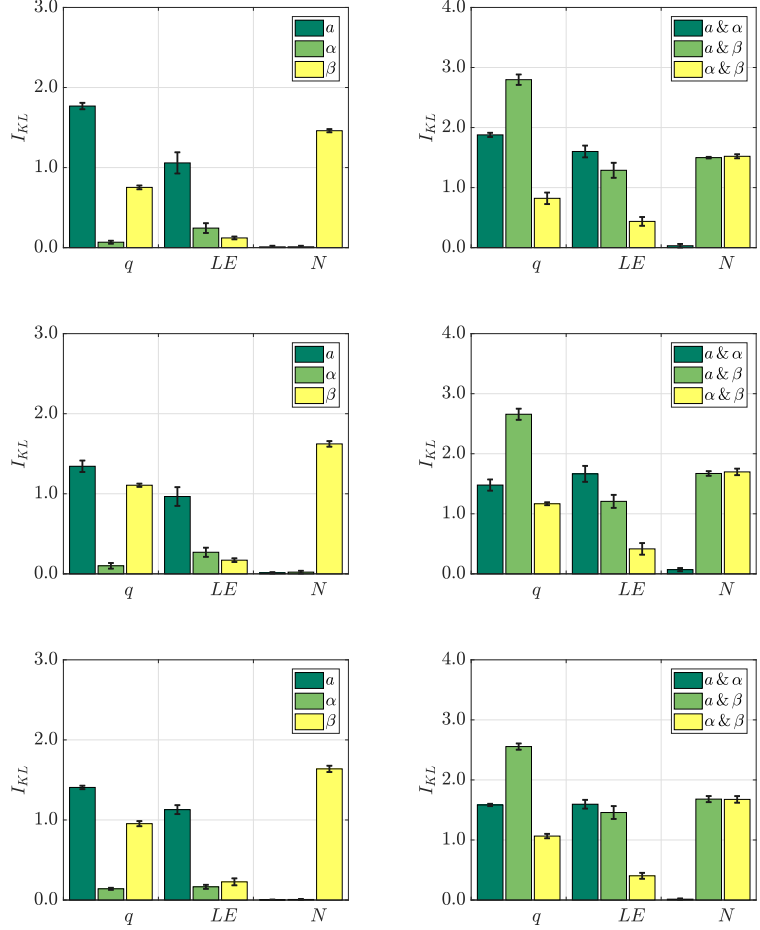


Figure 4.10: We perform three runs of ABC-SMC with synthetic data computed using the experimental parameters shown in the left-most tables, where we change the number of cells initially and the domain size those cells occupy (to ensure they are initially at their resting length). Prior distributions used and other inference choices are outlined in Table 4.1. The central plots show the information gain in the marginal one-dimensional posterior distributions for each parameter and each statistic, q , LE or N . The right-most plots show the information gain from the marginal two-dimensional posterior distributions for each parameter combination and each statistic. Each information gain is shown with error-bars computed from three independent runs of ABC-SMC. Note that the middle row of plots are the same as those in Figure 4.7, for ease of comparing the information gain when starting with more or less cells.

to implement in the computational model, this is harder to achieve in practice, and requires materials and experimental equipment such as those described in Gudipaty et al. [33]. These experiments seed cells onto a stretchable membrane, before uniformly stretching that membrane. Replicating these methods in the computational model (and creating the appropriate synthetic data), we are able to quantify the potential information gain from these types of experiment and compare to the ‘control’ experiment.

In Figure 4.11 the information gain when initialising $N_0 = 20, 30$ or 40 cells on a fixed domain of size $750\mu\text{m}$ is shown. The cells are either uniformly stretched ($N_0 = 20$), or uniformly compressed ($N_0 = 40$) with the aid of experimental apparatus [33]. Disappointingly, we find that there is only a slight increase in the information obtained in spring length parameter a , when using the cell density statistic q . However, when initialising cells compressed or stretched the two-dimensional (a, α) posterior distributions are shaped differently (overlap at right angles to each other), so combining them may allow us to obtain the previously unidentifiable motility parameter, α . This is investigated further in following section.

4.3.1.5 Combining experiments

By combining the statistics of the stretched and compressed experiments on a fixed domain ($R = 750\mu\text{m}$), we hope to identify the motility parameter α . In Figure 4.12 the information gain from these combined experiments, and the two-dimensional (a, α) posterior distributions using the leading edge statistic LE , are shown from the compressed, stretched and combined experimental protocols. The information gain for α is much larger than the previously considered experiments, suggesting it may be possible to identify α with these experiments.

Symbol	Value
Z	2
T	24 hr
N_0	20
R_0	750 μm
a_{exp}	25 μm
α_{exp}	5 $\mu\text{m}^2/\text{hr}$
β_{exp}	0.02 hr^{-1}
F_i	Linear
G_i	Constant

Symbol	Value
Z	2
T	24 hr
N_0	30
R_0	750 μm
a_{exp}	25 μm
α_{exp}	5 $\mu\text{m}^2/\text{hr}$
β_{exp}	0.02 hr^{-1}
F_i	Linear
G_i	Constant

Symbol	Value
Z	2
T	24 hr
N_0	40
R_0	750 μm
a_{exp}	25 μm
α_{exp}	5 $\mu\text{m}^2/\text{hr}$
β_{exp}	0.02 hr^{-1}
F_i	Linear
G_i	Constant

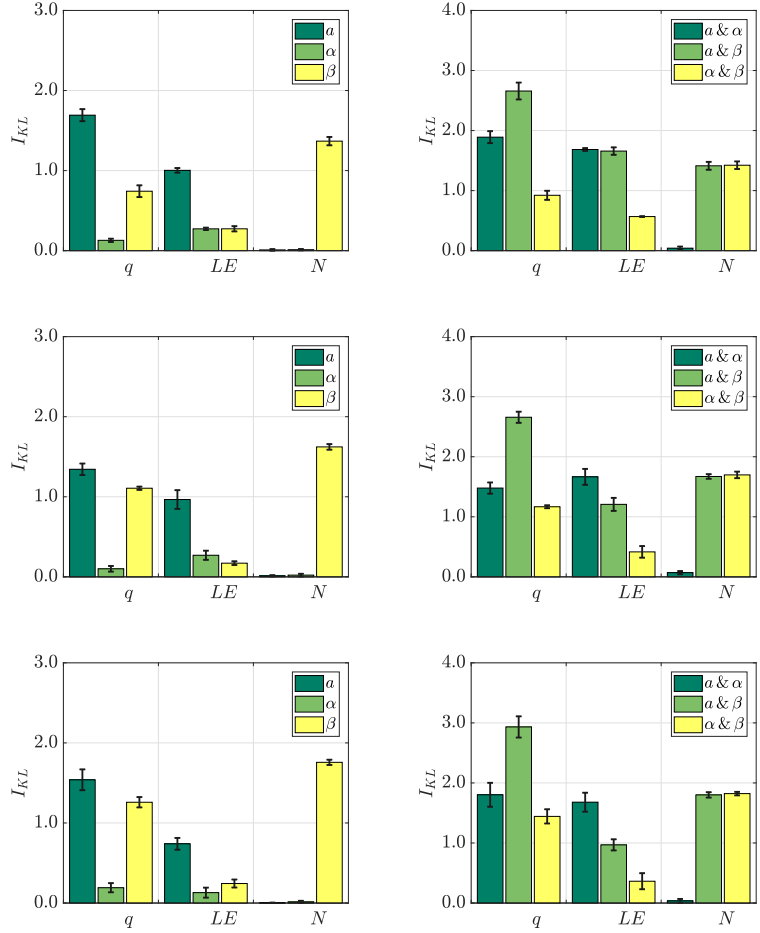


Figure 4.11: We perform three runs of ABC-SMC with synthetic data computed using the experimental parameters shown in the left-most tables, where we change the number of cells initially, keeping the size of domain fixed, so cells are either stretched (top-row), at their resting length (middle-row), or compressed (bottom-row) initially. Prior distributions used and other inference choices are outlined in Table 4.1. The central plots show the information gain in the marginal one-dimensional posterior distributions for each parameter and each statistic, q , LE or N . The right-most plots show the information gain from the marginal two-dimensional posterior distributions for each parameter combination and each statistic. Each information gain is shown with error-bars computed from three independent runs of ABC-SMC. Note that the middle row of plots are the same as those in Figure 4.7, for ease of comparing the information gain when starting with stretched or compressed cells.

Symbol	Value
Z	2
T	24 hr
N_0	20 and 40
R_0	750 μm
a_{exp}	25 μm
α_{exp}	5 $\mu\text{m}^2/\text{hr}$
β_{exp}	0.02 hr^{-1}
F_i	Linear
G_i	Constant

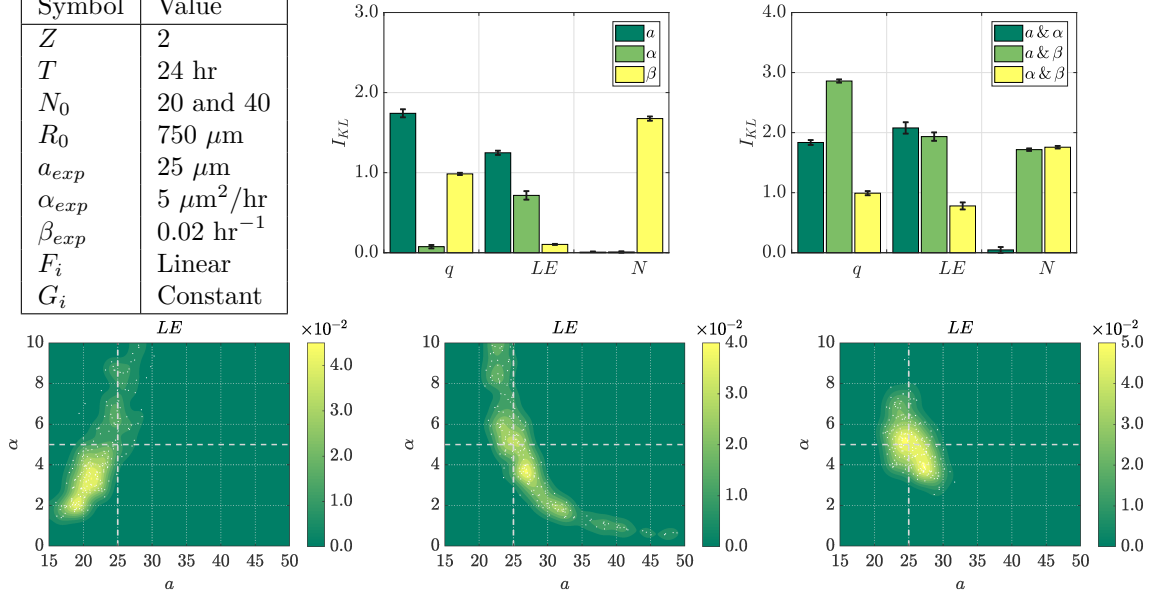


Figure 4.12: We perform ABC-SMC with combined synthetic data computed using the experimental parameters shown from an experiment with $N_0 = 20$ and with $N_0 = 40$, where we fix the domain size, so we have data from both stretched cells and compressed cells. Prior distributions used and other inference choices are outlined in Table 4.1. The bar plots show the information gain in the marginal one-dimensional or two-dimensional posterior distributions for each parameter combination and each statistic, q , LE or N . The bottom-row displays the marginal two-dimensional posterior distribution in a , α when using the leading edge statistic, from the initially stretched experiment (left, $N_0 = 20$), the initially compressed experiment (middle, $N_0 = 40$), or the combination of both of these experiments, (right). We can clearly see that the opposing shapes of the posterior distributions from the two experiments causes their combination to yield a larger information gain in parameter α .

4.3.1.6 Summary

To summarise investigations, Figure 4.13 shows the difference between the three-dimensional information gain in each experimental design considered and the ‘control’ experimental design shown in Figure 4.7. The simplest means to increase the information gain from the leading edge statistic, LE , and the number of cells statistic, N , is to use an experiment that captures images at a high frequency initially, and a slow frequency later, for example by observing at times $t = 0$ hr, 6 hr, 12 hr and 24 hr. Further, the information from the leading edge, LE , is substantially increased when combining the fixed-domain $N_0 = 20$ and $N_0 = 40$ experiments. In order to obtain an estimate for α , we determined that a combination of stretch/compressed experiments should be performed. In practice, however, such experiments are difficult to perform without the requisite experimental apparatus, so it may not routinely be possible to estimate the motility parameter, α . The motility parameter α is hard to interpret physically, as it is a combination parameters, one of which is the drag coefficient of the underlying substrate, and is therefore difficult to quantify. In order to understand the impact of the lack of confidence in α has on future experimental predictions, posterior predictive checks are considered in the next section for a range of experimental data.

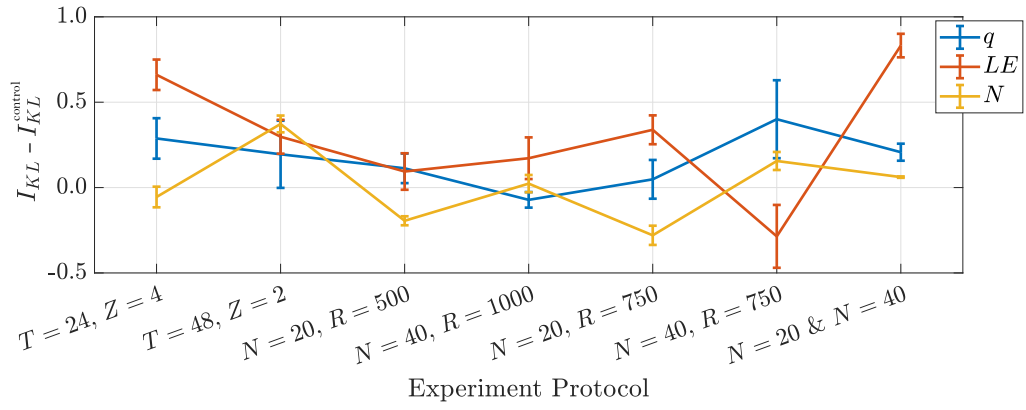


Figure 4.13: The ABC-SMC experiments performed lead to three-dimensional posterior distributions for each statistic, from which we compute the (three-dimensional) information gains, and display here the difference in the information gain from the experiment considered and the “control” experiment we started with in Figure 4.7, denoted $I_{KL}^{control}$ ($T = 24$ hr, $Z = 2$). The error-bars are computed from three independent runs of ABC-SMC in each case. We observe that some experiments lead to less information gain for some statistics, whereas some lead to marginal or considerable increases in the information gained. This is discussed further in Section 4.3.1.6

4.3.2 Inference using experimental data

The data from Jin et al. [36] describes the evolution of the density of cells in a scratch assay over time. This cell density data is converted to node positions using the methods in Appendix C.1, allowing us to initialise the model with non-uniform initial cell densities, replicating experimental data that is captured with $Z = 2$ time points over a period of $T = 24$ hr. We consider data taken from an experiment initialised with 14,000, 16,000 and 18,000 cells. We infer parameters of each set of experimental data before choosing one of the resulting posterior distributions to attempt posterior predictive checks, to ascertain the predictive potential of the model.

4.3.2.1 Choice of inference particulars

We choose the following prior distributions:

$$\begin{aligned}\pi(a) &= \mathcal{U}(25, 75)\mu\text{m}; \\ \pi(\alpha) &= \mathcal{U}(0.001, 10)\text{h}^{-1}; \\ \pi(\beta) &= \mathcal{U}(0, 0.08)\text{h}^{-1}.\end{aligned}$$

We use ten replicates, as this results in statistics with a variation similar to experimental data.

4.3.2.2 Varying initial cell density

ABC-SMC is applied to each experimental data set (with either 14,000, 16,000, or 18,000 cells initialised), and a subset of the resulting one-dimensional posterior distributions and the resulting Kullback-Leibler information gains in moving from the prior to the posterior are shown in Figure 4.14. For the spring length a , the posterior distribution resulting from the use of the cell density statistic, q , is used; for the motility parameter, α , the leading edge statistic, LE , is used; and for the prolifera-

tion parameter, β , the number of cells statistic, N , is used to generate the posterior distribution. The analogous plots for two-dimensional posterior distributions and information gains are shown in Figure 4.15, and display the (a, α) distributions for q and LE , or the (a, β) distribution for N . For each parameter (or parameter combination), the posterior distribution displayed generally corresponds to the posterior distribution with a maximal information gain out of the three statistics considered.

We note that for each experimental data set, the one- and two-dimensional posterior distributions are slightly different to one another when using the same statistic. Looking down columns of Figure 4.14, for example, clearly shows varying widths of posterior distributions for the spring length, a and the motility parameter, α , and appears to show a shifting of the posterior distribution for the proliferation parameter, β depending on the initial number of cells.

However, in either one- or two-dimensions, there is overlap in these posterior distributions. Similarly to the results in the previous section, the cell density statistic, q provides the most information about the spring length parameter, a , the leading edge statistic, LE , provides the most information about the motility parameter α , and the number of cells statistic is most informative with the proliferation parameter, β . Using the one-dimensional posterior distributions for a , α and β for each statistic, we can obtain estimates for a , α and β from the experimental data sets considered, and these are given in Table 4.2 for ease of comparison.

The Kullback-Leibler information gains of the full posterior distributions from each statistic are given in Table 4.3. In particular, we note that the cell density statistic, q , is the most informative statistic in three dimensions on average, though consideration of one- and two-dimensional information gains is also vital in making this conclusion, as the number of cells statistic is also very informative, but one- and two-dimensional posterior distributions make it clear that it provides no information about the model parameters a and α .

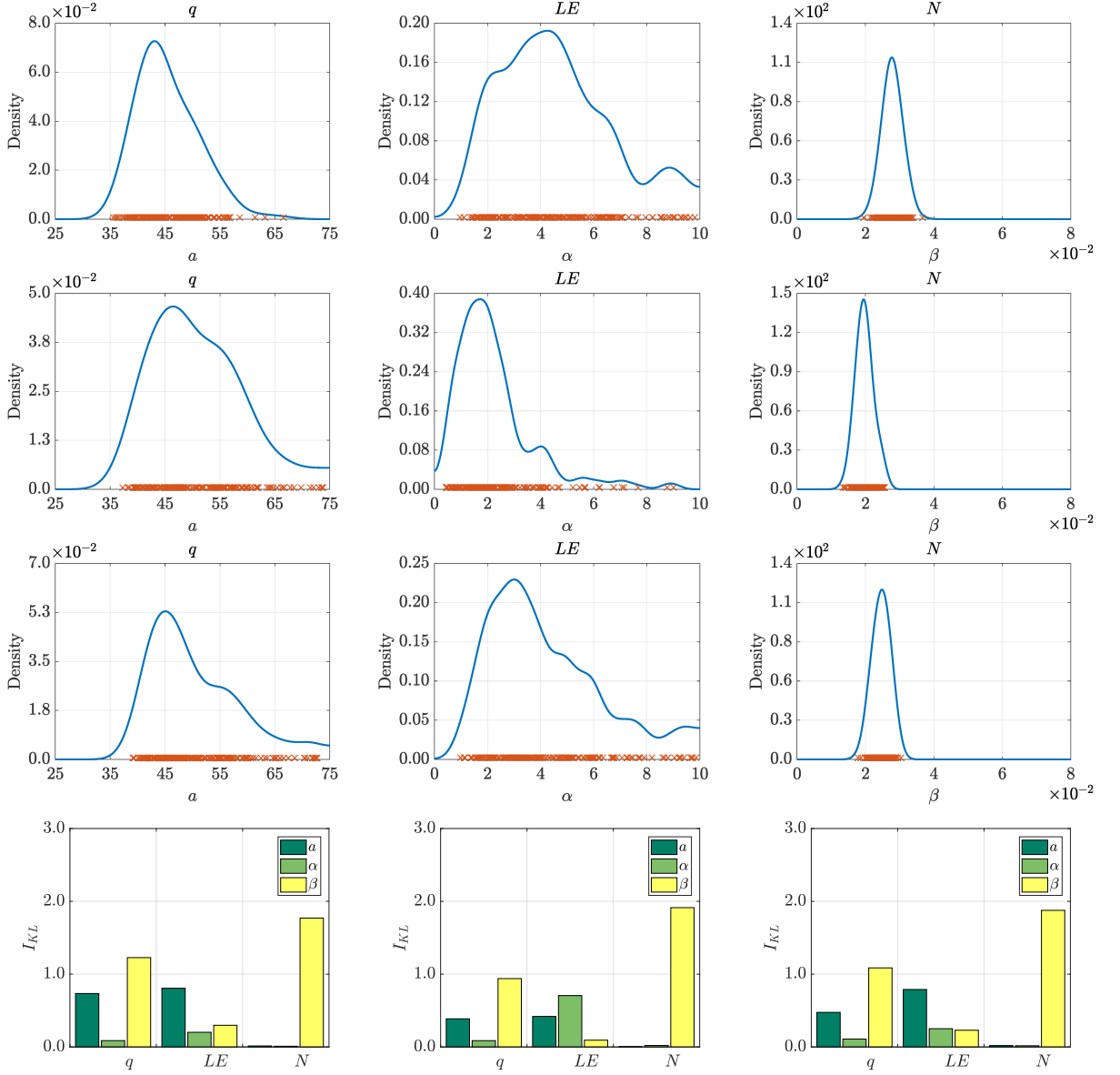


Figure 4.14: We perform ABC-SMC using the experimental data from Jin et al. [36] extracted using the methods in Appendix C.1.2. The inference particulars we use are described in Section 4.3.2.1. Experimental data is initialised with either 18000 cells (top row), 16000 cells (second row) or 14000 cells (third row). Experimental data consists of two time points, and has a duration of 24 hours. We find that each initialisation results in slightly different marginal one-dimensional posterior distributions, suggesting a sensitivity of the parameters on the initial number of cells, and is discussed further in Section 4.3.2.2.

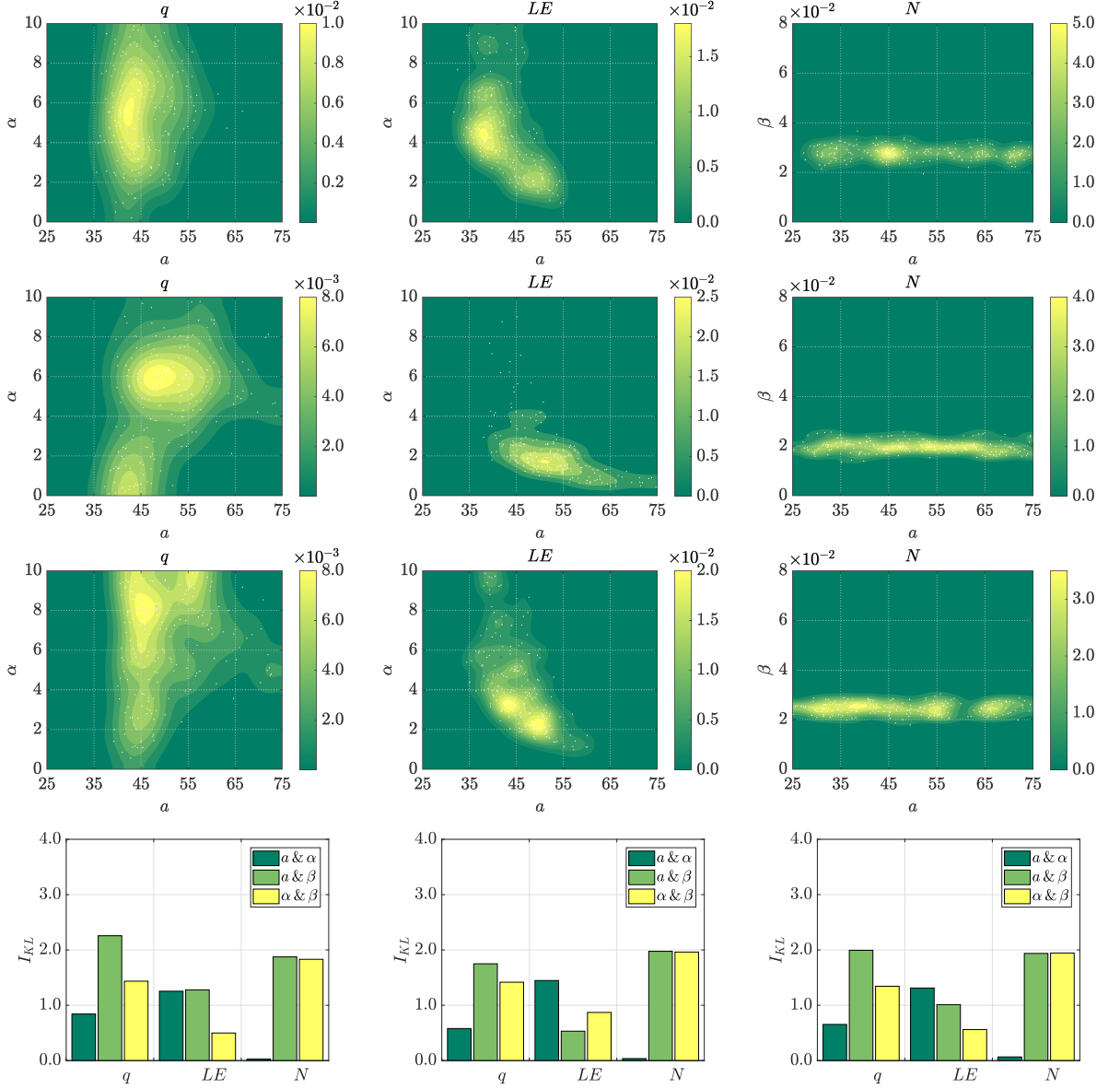


Figure 4.15: We perform ABC-SMC using the experimental data from Jin et al. [36] extracted using the methods in Appendix C.1.2. The inference particulars we use are described in Section 4.3.2.1. Experimental data is initialised with either 18,000 cells (top row), 16,000 cells (second row) or 14,000 cells (third row). Experimental data consists of two time points, and has a duration of 24 hours. We find that each initialisation results in slightly different marginal two-dimensional posterior distributions, suggesting a sensitivity of the parameters on the initial number of cells, and is discussed further in Section 4.3.2.2.

Table 4.2: We compute the means and standard deviations of the model parameters from the experimental data inference (see Figure 4.14), using the statistic that most commonly leads to maximum one-dimensional information gain in each case.

Parameter / Statistic	18,000 cells	16,000 cells	14,000 cells
a / q	45.44 ± 5.47	50.80 ± 8.04	50.37 ± 8.03
α / LE	4.55 ± 2.09	2.26 ± 1.51	4.18 ± 2.11
β / N	0.0278 ± 0.0027	0.0198 ± 0.00244	0.0246 ± 0.00237

Table 4.3: Three-dimensional Kullback-Leibler information gain from each experimental data set and statistic combination.

Experiment/Statistic	q	LE	N
I_{KL} (18000 cells)	2.3917	1.6823	2.0125
I_{KL} (16000 cells)	1.9259	1.5069	2.1379
I_{KL} (14000 cells)	2.0417	1.6169	2.1148
Average I_{KL}	2.1198	1.6020	2.0884

4.3.2.3 Posterior predictive checks

The posterior distribution from one of experimental data sets can be used to predict experimental statistics from a different experimental data set using a posterior predictive check. A check is performed by repeatedly sampling from one posterior distribution and running the model to obtain a distribution of statistics. Comparing the resulting distribution of statistics with the true statistics determines the predictive ability of the model in an unfamiliar context.

In Figure 4.16 parameters are sampled from the posterior distribution using the cell density statistic, q , established from the experimental data initialised with 16,000 cells. Predictions are made of the statistics of the experimental data sets initialised with 14,000 cells, 16,000 cells, or 18,000 cells initially. We find that we are able to predict the statistics from each of the experimental data sets. We note that the error bars displayed are one standard deviation, so experimental data (blue lines) are well within confidence intervals for each statistic. Similar results are seen if we use the posterior distributions from other initial numbers of cells to predict (e.g. using the posterior distributions from the experiment with 14000 cells initially to

predict the statistics from the experiment with 18000 cells initially). These results give us confidence that the one-dimensional stochastic model incorporating proliferation can reasonably be used in an experimental context to infer parameters and make predictions on the behaviour of cells under different initial data.

4.4 Discussion

In this chapter we have applied ABC to explore experimental design in a one-dimensional off-lattice model that accounts for cell mechanics and where cells proliferate stochastically. When using synthetically created data, we have demonstrate a series of experiments that can increase the information gained in moving from an uninformative prior distribution to a posterior distribution for the model parameters.

We have also considered the inference of model parameters when using experimental data, and found that despite obtaining information gains that are generally lower than in synthetic experiments, we are able to predict statistics from experiments with different initial distributions of cells. This contrasts slightly with the conclusion reached in Jin et al. [36], where different experimental data sets resulted in inferred model parameters that depended on the initial density of cells.

Note that in this work we have used a simple linear force law and a proliferation law with a constant proliferation rate. It is uncertain what force or proliferation law would best capture the behaviour of cells. One future direction of this work would be to try and utilise model selection techniques [51] to infer between different force laws and proliferation laws.

To further investigate the model established in this Chapter and its behaviour under different force and proliferation laws, we attempt to establish a continuum model in the next Chapter. This will allow us to compare statistics resulting from the model under different initial conditions and parameter combinations.

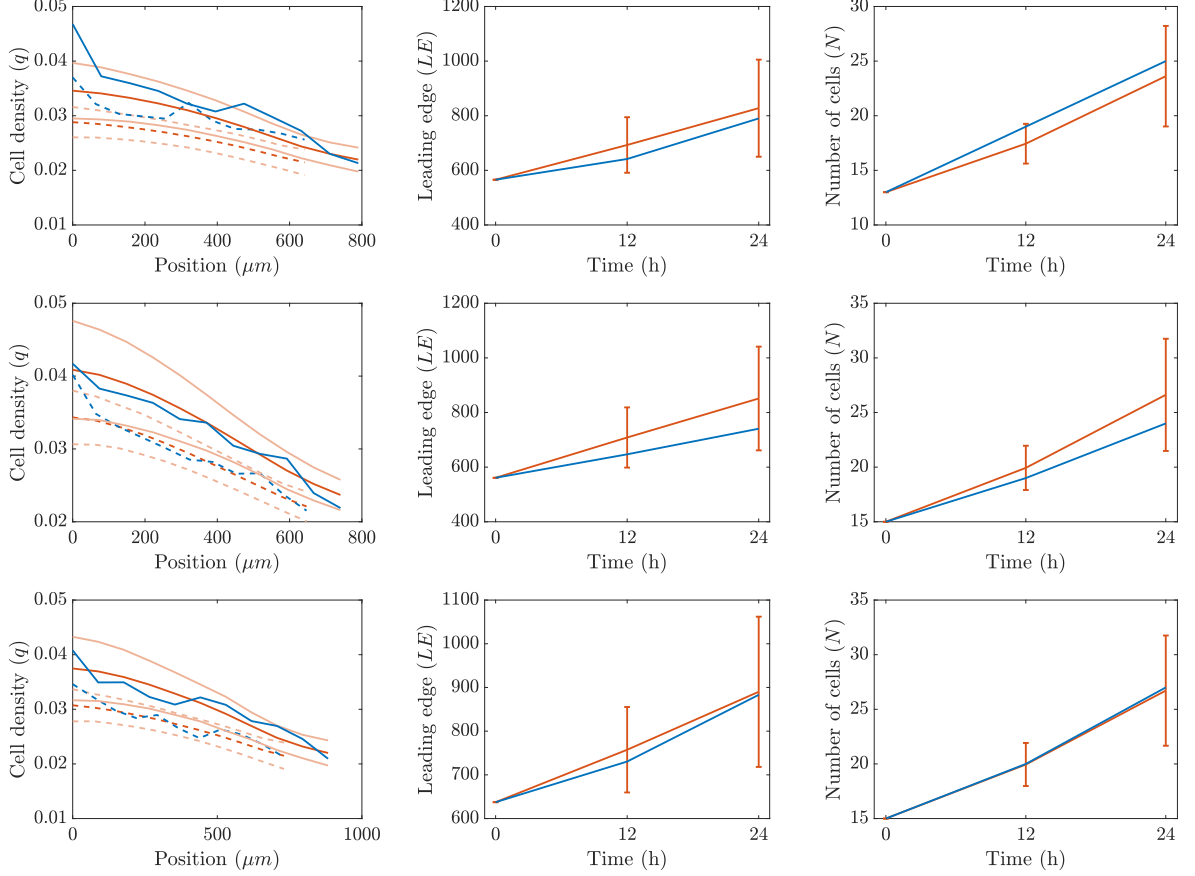


Figure 4.16: We perform posterior predictive checks by sampling from the three-dimensional posterior distribution for cell density, q , from the experiment with 16,000 cells initially (Figure 4.14 shows the corresponding marginal one-dimensional posterior distributions). In the top row, we attempt to predict the statistics of the 18,000 cells initially experiment (blue) - we show the average and standard deviation of the statistics computed from 1000 independent runs of the model (red). In the middle row, we attempt to predict the statistics of the 16,000 cells initially experiment (blue) from 1000 independent runs of the model (red). In the bottom row, we attempt to predict the statistics of the 14,000 cells initially experiment (blue) from 1000 independent runs of the model (red). In the case of the leading edge and number of cells, standard deviations are shown as error bars, whereas for the cell density, we show the statistics at $t = 12$ hr (dashed) and $t = 24$ hr (solid), and demonstrate the errorbars on the data via the pink lines at either time point.

Chapter 5

Coarse graining of an off-lattice model with proliferation

The work in this chapter has been submitted for publication in Journal of Theoretical Biology, under the title *A free boundary model of epithelial dynamics*.

5.1 Introduction

Understanding cell spreading in biological systems is achieved by establishing computational or mathematical models formulated through the guidance of experiments. Models are usually established at the level of an individual cell, but in order to link them with experimental data it is beneficial to coarse grain individual-based models in order to yield a density formulation that can be compared directly with experiments.

Coarse graining of cell-spreading models usually entails deriving a partial differential equation (PDE) for the cell density that incorporates the cell-level mechanisms specified in the IBM. On-lattice models of cell spreading allow for less computationally demanding simulations and analytical tractability. However, the introduction of lattice-based artefacts in these models can render them less realistic than off-lattice models. In particular, off-lattice models allow more realistic force laws to be accounted

for [56, 58].

The difficulty in coarse graining off-lattice models means that most coarse-graining attempts are performed on one-dimensional models [26], and there are few coarse-grained models that account for both mechanisms of cell proliferation and cell-cell interaction. Coarse graining of proliferation mechanisms in off-lattice models, to our knowledge, has only previously been attempted at low densities [65], though one could postulate proliferation laws that recapitulate individual behaviours (e.g. according to a cell cycle, [57]).

The work in this chapter extends the work of Murray et al. [58], using the model established in Chapter 4, by deriving continuum models for the evolution of the density of cells for a general interaction force law and a general proliferation law. Through the careful coarse-graining and solving of a cell-density PDE on a growing domain, we demonstrate the ability to accurately describe population-level behaviours with individual processes in a wide variety of potential cell phenotypes.

5.2 Linear force law model

The model considered here has been previously established in Chapter 4. In brief, we consider the one-dimensional cell spreading model first introduced by Murray et al. [57], where a series of springs are attached to one another at node points. Springs in this model formulation can be thought of as cells, and node points represent the boundary of cells.

Cells move according to the forces they are subjected to by their surrounding cells, and cells are assumed to be in contact with their neighbours at their boundaries. The left-hand edge of the first cell is fixed at the origin, and the right-hand edge of the final cell is free to move. The full model is visualised during a proliferation event in Figure 4.4, however, in this section the model is considered in the absence of proliferation.

Deriving a model from a balance of forces argument results in equation (4.1) in Chapter 4. After simplifying assumptions (outlined in the previous paragraph) are taken into consideration, the resulting model is

$$\begin{aligned} r_0 &= 0, \\ \eta \frac{dr_i}{dt} &= F_{i,i-1}^{int} + F_{i,i+1}^{int}, \text{ for } i = 1, \dots, N-1, \\ \eta \frac{dr_N}{dt} &= F_{N,N-1}^{int}, \end{aligned} \tag{5.1}$$

where i is the node indexing variable, the interaction force between cells i and j is given by $F_{i,j}^{int}$, N is the number of cells, t is the current time, and η is the drag constant. To progress with the coarse graining of this model, assume that the interaction force is a linear force law with spring constant k .

$$F_{i,j}^{int} = k(a - |r_i - r_j|) \frac{r_i - r_j}{|r_i - r_j|}.$$

Note that i is integer valued, and ranges between 0 and N , the number of cells. r_i is a continuous variable that represents position, has dimensions of length, and ranges between 0 and $N \times a$ when cells are at equilibrium. By writing $\alpha = k/\eta$, the equation of motion of node $i \in \{1, \dots, N-1\}$ is given by

$$\frac{dr_i(t)}{dt} = \alpha [r_{i-1}(t) - 2r_i(t) + r_{i+1}(t)], \quad i = 1, \dots, N-1, \tag{5.2}$$

The position of the final node, r_N , is only influenced by the relaxation of the N^{th} cell, so the equation of motion is

$$\frac{dr_N}{dt} = \alpha [r_{N-1}(t) - r_N(t) + a]. \tag{5.3}$$

The equation of motion for $i = 0$ is still $r_0 = 0$, as it is fixed at the origin. We assume

that the initial positions of nodes are given by a strictly increasing function of the index variable.

5.2.1 Numerical solution of node positions model

The initial system of ODEs model, as described in equations (5.2) and (5.3), can be solved analytically, but for simplicity we solve for the position of each node, r_i , numerically using a simple forward Euler method with time-step $dt = 0.001$, and node positions initially uniformly distributed in $[0, R]$.

The forward Euler method is implemented with self-written code using MATLAB 2019a. Using a desktop Windows 10 PC with an Intel Core i5-3570 processor (3.40 GHz) and 16GB RAM. A simulation with $N = 45$ cells initially, $a = 1$ and $\alpha = 15$, until final time $T = 100$, takes 0.829 seconds to run.

5.2.2 Continuum approximation

To make progress in deriving an equivalent continuum, coarse-grained model, with a slight abuse of notation we extend the node positions, $r_i(t)$, which is only defined for discrete $i \in \{0, \dots, N\}$, to a smooth function, $r(i, t)$, which is defined for $i \in [0, N]$. The function $r(i, t)$ will approximate r_i when i is an integer:

$$r_i(t) \approx r(i, t), \quad \text{for } i = 0, \dots, N, \text{ and } t \geq 0$$

To facilitate coarse graining, we first non-dimensionalise the model in equation (5.2) using the scalings

$$\tilde{i} = \frac{i}{N}, \quad \tilde{r} = \frac{r}{aN}, \quad \tilde{t} = \frac{\alpha t}{N^2} = \frac{kt}{\eta N^2},$$

so that $\tilde{i} \in \{0, 1/N, 2/N, \dots, 1\}$ and $\tilde{r} = \mathcal{O}(1)$. The model equations are then

$$\begin{aligned}\tilde{r}(0, \tilde{t}) &= 0, \\ \frac{\partial \tilde{r}(\tilde{i}, \tilde{t})}{\partial \tilde{t}} &= N^2 \left[\tilde{r}(\tilde{i} - 1/N, \tilde{t}) - 2\tilde{r}(\tilde{i}, \tilde{t}) + \tilde{r}(\tilde{i} + 1/N, \tilde{t}) \right], \\ \frac{\partial \tilde{r}(1, \tilde{t})}{\partial \tilde{t}} &= N^2 \left[\tilde{r}(1 - 1/N, \tilde{t}) - 2\tilde{r}(1, \tilde{t}) + 1/N \right],\end{aligned}\tag{5.4}$$

where equation (5.4) holds for $\tilde{i} = 1/N, \dots, 1 - 1/N$.

Now we are able to consider the Taylor expansion about $\tilde{i}$, after substitution of the scalings in equations (5.2) and (5.3), to arrive at

$$\begin{aligned}\tilde{r}(0, \tilde{t}) &= 0, \\ \frac{\partial \tilde{r}(\tilde{i}, \tilde{t})}{\partial \tilde{t}} &= N^2 \left[\tilde{r}(\tilde{i}, \tilde{t}) - \frac{1}{N} \frac{\partial \tilde{r}}{\partial \tilde{i}} + \frac{1}{2N^2} \frac{\partial^2 \tilde{r}}{\partial \tilde{i}^2} - \frac{1}{6N^3} \frac{\partial^3 \tilde{r}}{\partial \tilde{i}^3} + \frac{1}{24N^4} \frac{\partial^4 \tilde{r}}{\partial \tilde{i}^4} + \dots \right. \\ &\quad \left. + \tilde{r}(\tilde{i}, \tilde{t}) + \frac{1}{N} \frac{\partial \tilde{r}}{\partial \tilde{i}} + \frac{1}{2N^2} \frac{\partial^2 \tilde{r}}{\partial \tilde{i}^2} + \frac{1}{6N^3} \frac{\partial^3 \tilde{r}}{\partial \tilde{i}^3} + \frac{1}{24N^4} \frac{\partial^4 \tilde{r}}{\partial \tilde{i}^4} + \dots \right. \\ &\quad \left. - 2\tilde{r}(\tilde{i}, \tilde{t}) \right] \\ &= \frac{\partial^2 \tilde{r}}{\partial \tilde{i}^2} + \mathcal{O}\left(\frac{1}{N^2}\right), \quad \tilde{i} \in (0, 1) \\ \frac{1}{N} \frac{\partial \tilde{r}(1, \tilde{t})}{\partial \tilde{t}} &= N \left[\tilde{r}(1, \tilde{t}) - \frac{1}{N} \frac{\partial \tilde{r}}{\partial \tilde{i}} + \frac{1}{2N^2} \frac{\partial^2 \tilde{r}}{\partial \tilde{i}^2} + \dots \right. \\ &\quad \left. - \tilde{r}(1, \tilde{t}) + \frac{1}{N} \right], \\ &= 1 - \frac{\partial \tilde{r}}{\partial \tilde{i}} + \frac{1}{2N} \frac{\partial^2 \tilde{r}}{\partial \tilde{i}^2} + \mathcal{O}\left(\frac{1}{N^2}\right).\end{aligned}$$

Neglecting terms which are $\mathcal{O}(1/N^2)$, in the limit when N is large, we obtain a PDE for the scaled positions of nodes, $\tilde{r}$, in terms of the continuous index variable $\tilde{i}$. In terms of the original variables, (using the above continuum approximation for i) we

obtain the diffusion equation,

$$\frac{\partial r}{\partial t} = \alpha \frac{\partial^2 r}{\partial i^2}, \text{ for } i \in (0, N), \quad (5.5)$$

along with the boundary conditions

$$r = 0, \text{ for } i = 0 \quad \text{and} \quad \frac{\partial r}{\partial t} = \alpha \left[a - \frac{\partial r}{\partial i} + \frac{1}{2} \frac{\partial^2 r}{\partial i^2} \right], \text{ for } i = N. \quad (5.6)$$

Initial conditions are specified by extending the discrete initial conditions, $r_i(0) = r_i^0$ for $i = 1, \dots, N$, to a continuous function $r(i, 0)$ such that $r(i, 0) = r_i^0$ for $i = 1, \dots, N$. For simplicity, throughout this work extend the discrete initial condition to a piecewise linear continuous function.

5.2.3 Derivation of cell density model

We convert equations (5.5) and (5.6) to be in terms of cell density by performing a change of variables from (i, t) to (r, τ) , and we define the cell density, $q(r, \tau)$, using the node index, $i = i(r, \tau)$, as follows,

$$i(r, \tau) = \int_0^r q(r', \tau) dr', \quad t = \tau, \quad (5.7)$$

which is more intuitively expressed as $q(r, \tau) = \partial i(r, \tau) / \partial r$. Note that this definition ensures $i = 0$ at the boundary $r = 0$. We perform the change of variables from (i, t) to (r, τ) by noting that,

$$\begin{pmatrix} \frac{\partial r}{\partial i} \Big|_t & \frac{\partial r}{\partial t} \Big|_i \\ \frac{\partial \tau}{\partial i} \Big|_t & \frac{\partial \tau}{\partial t} \Big|_i \end{pmatrix} = \begin{pmatrix} \frac{\partial i}{\partial r} \Big|_\tau & \frac{\partial i}{\partial \tau} \Big|_r \\ \frac{\partial t}{\partial r} \Big|_\tau & \frac{\partial t}{\partial \tau} \Big|_r \end{pmatrix}^{-1} = \frac{1}{\frac{\partial i}{\partial r} \Big|_\tau \frac{\partial t}{\partial \tau} \Big|_r - \frac{\partial i}{\partial \tau} \Big|_r \frac{\partial t}{\partial r} \Big|_\tau} \begin{pmatrix} \frac{\partial t}{\partial \tau} \Big|_r & -\frac{\partial i}{\partial \tau} \Big|_r \\ -\frac{\partial t}{\partial r} \Big|_\tau & \frac{\partial i}{\partial r} \Big|_\tau \end{pmatrix},$$

and therefore, $t = \tau$ yields,

$$\frac{\partial r}{\partial i} = \frac{1}{\frac{\partial i}{\partial r}} = \frac{1}{q} \quad \text{and} \quad \frac{\partial r}{\partial t} = -\frac{1}{\frac{\partial i}{\partial r}} \frac{\partial i}{\partial \tau} = -\frac{1}{q} \frac{\partial i}{\partial \tau}. \quad (5.8)$$

Another interpretation of the second equation in (5.8) is that as we follow the frame of reference of node i while it moves, we know that i does not change, and hence the material derivative of i is zero, so that we must have

$$\frac{Di}{Dt} = \frac{\partial i}{\partial t} + \frac{\partial i}{\partial r} \frac{\partial r}{\partial t} = 0.$$

If we now substitute equation (5.8) into the right-hand side of equation (5.5), we obtain

$$\alpha \frac{\partial^2 r}{\partial i^2} = \alpha \frac{1}{\frac{\partial i}{\partial r}} \frac{\partial}{\partial r} \left(\frac{1}{\frac{\partial i}{\partial r}} \right) = \alpha \frac{1}{q} \frac{\partial}{\partial r} \left(\frac{1}{q} \right) = -\frac{\alpha}{q^3} \frac{\partial q}{\partial r}. \quad (5.9)$$

Therefore equation (5.5) can be rewritten as

$$\frac{\partial r}{\partial t} = -\frac{\alpha}{q^3} \frac{\partial q}{\partial r}, \quad \text{for } i \in (0, N), \quad (5.10)$$

which represents how the domain, r , evolves along characteristics with constant index, i . After a simple rearrangement using equation (5.8), by multiplying by $-q$ and differentiating with respect to r , we obtain a PDE for the cell density,

$$\frac{\partial q}{\partial \tau} = \frac{\partial}{\partial r} \left[\frac{\alpha}{q^2} \frac{\partial q}{\partial r} \right], \quad \text{for } r \in (0, L(\tau)), \quad (5.11)$$

which is a free boundary problem where the position of the boundary is given by $L(\tau) = r(N, t)$.

The same change of variables applied to the boundary conditions in equation (5.6)

yields

$$\frac{\partial q}{\partial r} = 0, \text{ for } r = 0 \quad \text{and} \quad \alpha \left[a - \frac{1}{q} - \frac{1}{2q^3} \frac{\partial q}{\partial r} \right] = 0, \text{ for } r = L(\tau). \quad (5.12)$$

To check the validity of the established boundary conditions, we consider whether these conditions conserve mass. The mass is given by the integral of the cell density, q , so consider the differentiation of the mass with respect to τ ,

$$\begin{aligned} \frac{d}{dt} \int_0^{L(\tau)} q(r', \tau) dr' &= \frac{dL(\tau)}{d\tau} q(L(\tau), \tau) + \int_0^{L(\tau)} \frac{\partial q}{\partial \tau}(r', \tau) dr', \\ &= \frac{dL(\tau)}{d\tau} q(L(\tau), \tau) + \int_0^{L(\tau)} \frac{\partial}{\partial r} \left[\frac{\alpha}{q^2} \frac{\partial q}{\partial r} \right] dr', \\ &= \left[\alpha q \left(a - \frac{1}{q} - \frac{1}{2q^3} \frac{\partial q}{\partial r} \right) + \frac{\alpha}{q^2} \frac{\partial q}{\partial r} \right] \Big|_{r=L(\tau)} - \frac{\alpha}{q^2} \frac{\partial q}{\partial r} \Big|_{r=0}, \\ &= 0, \end{aligned}$$

where the final line comes from using equations (5.14). Therefore mass is conserved under these boundary conditions. Note that the boundary condition applied at the right-hand boundary $r = L(\tau)$, is slightly different to that derived in [57], where the boundary condition was derived by neglecting terms that are $\mathcal{O}(1/N)$ and is of the form $q = 1/a$ for $r = L(\tau)$.

To establish an initial density, $q(r, 0) \equiv q_0(r)$, for initial node positions, r_i^0 , equation (5.7) is used with a finite difference approximation to write,

$$2 = (i+1) - (i-1) = \int_{r_{i-1}^0}^{r_{i+1}^0} q(r, 0) dr \approx (r_{i+1}^0 - r_{i-1}^0) q_0(r_i^0), \quad i = 1, \dots, N-1$$

which can be rearranged to give

$$q_0(r_i^0) = \frac{2}{(r_{i+1}^0 - r_{i-1}^0)}, \text{ for } i = 1, \dots, N-1.$$

A similar finite difference approximation on the left- and right-hand boundaries gives,

$$q_0(0) = \frac{1}{r_1^0}, \quad \text{and} \quad q_0(r_N^0) = \frac{1}{r_N^0 - r_{N-1}^0},$$

and we take $q_0(r)$ to be piecewise linear between node positions.

5.2.3.1 Numerical solution

In summary we obtain a PDE for the cell density in the domain,

$$\frac{\partial q}{\partial \tau} = \frac{\partial}{\partial r} \left[\frac{\alpha}{q^2} \frac{\partial q}{\partial r} \right], \text{ for } r \in (0, L(\tau)), \quad (5.13)$$

along with the boundary conditions

$$\left. \frac{\partial q}{\partial r} \right|_{r=0} = 0 \quad \text{and} \quad \left[\frac{1}{2q^2} \frac{\partial q}{\partial r} + aq - 1 \right] \Big|_{r=L(\tau)} = 0. \quad (5.14)$$

and an initial condition that is computed from initial node positions as described previously,

$$q(r, 0) = q_0(r) \quad \text{for} \quad r \in (0, L(0)). \quad (5.15)$$

We are therefore left with an equation for the density on an evolving domain. We numerically solve these equations by using a Lagrangian transformation to convert this to a PDE for the density on a fixed domain, and a PDE for the evolving domain. We outline the full method for the more general model with proliferation in Appendix D.2, where an implicit finite difference scheme is used, with Picard iteration.

5.2.4 Comparing the accuracy of the continuum model with different boundary conditions.

To arrive at the boundary conditions, we neglected quadratically small terms, $\mathcal{O}(1/N^2)$, however if we neglected the $\mathcal{O}(1/N)$ terms in the derivation of equation (5.6), we could instead arrive at the boundary condition derived in Murray et al. [57], $q = 1/a$ on $r = L(\tau)$. Intuitively, we expect the approximation that maintains the higher order terms to be more accurate, so to examine the importance of keeping these higher order terms, we consider an example problem where we have N cells with equal lengths initialised in $R \in (0, 30)$, and we systematically vary N . We alter the parameters, $\alpha = 15(N/45)^2$ and $a = 45/N$, so that the scalings for r and t do not change with increasing N . This means that we consider $aN = 45$ and $N^2/\alpha = 135$ as being constants. In each case, cells are initially compressed but will expand to fill the domain $r \in (0, 45)$.

In Figure 5.1 we compare predictions of the PDE model, equation (5.11) and the forward Euler numerical solution of the ODE model, equation (5.2), for both the time evolution of the position of the leading edge, $L(\tau)$, and the density profile, $q(r, \tau)$. This comparison reveals that the continuum limit description with the new boundary condition is considerably more accurate than the simpler boundary condition. The improved accuracy of this boundary condition can also be observed for different force laws, different initial conditions and different parameter values (results not shown). It is also worth noting that the continuum approximation appears to approximate the discrete model even with low cell numbers ($N = 15$ is used in the left-most plots), this may be aided by the simple nature of the initial condition. The deviation in the discrete and continuum models are clearly seen in plot of the leading edge deviation, where increasing cell number has a large impact on the accuracy of the $q = 1/a$ boundary condition, but relatively little effect on solutions with the new boundary condition.

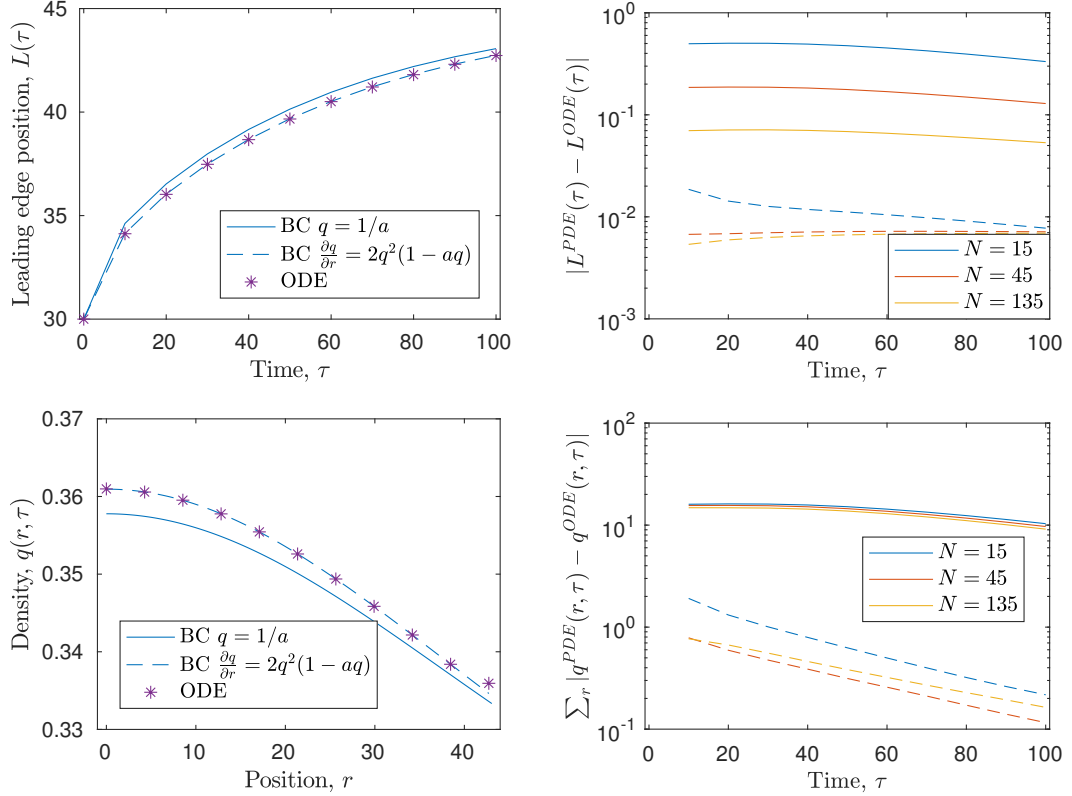


Figure 5.1: Comparison of the leading edge positions, L , and cell density, q , in the solutions of the cell density PDE, equation (5.11), varying the number of cells, N , and the boundary condition at $L(\tau)$. We initialise $N = 15$ (blue), $N = 45$ (red) or $N = 135$ (yellow) cells uniformly in $R \in (0, 30)$, while keeping $aN = 45$ and $N^2/\alpha = 135$ fixed. PDE solutions with the boundary condition $q = 1/a$ are solid lines, PDE solutions with the new boundary condition $\partial q/\partial r = 2q^2(1 - aq)$ are dashed lines, and the ODE solution is given as discrete stars. In the left plots, we display the PDE and ODE solutions for $N = 15$, and in the right plots we compute the difference in the PDE solutions, L^{PDE} and q^{PDE} , and the discrete ODE solutions, L^{ODE} and q^{ODE} over time. The details of the numerical discretisation for the cell density model is given in Appendix D.2.

dx/dt	0.01	0.005	0.004	0.002	0.001
0.01	19.7644	28.9119	34.8034	67.1836	132.9991
0.005	42.2553	-	-	-	-
0.004	54.9275	-	-	-	-
0.002	136.9538	-	-	-	-
0.001	382.3898	-	-	-	1174.966357

Table 5.1: The time taken in seconds to solve the PDE system in Section 5.4.6, using the algorithms outlined in Appendix D.2, for varying space (dx) and time (dt) resolutions. The solver uses $N = 45$ cells with node positions uniformly distributed in $[0, 30]$, with $a = 1$, $\alpha = 15$, a linear force model, and no proliferation. Dashes indicate spatial/temporal discretisations not considered.

The bottom right plot of Figure 5.1 implies a less accurate density in the case of $N = 45$ than $N = 135$, and this suggests a potential issue with the convergence of the PDE solver. To investigate this further, parameters are fixed as $N = 45$, $a = 1$, $\alpha = 15$, and the spatial discretisation and temporal discretisations of the numerical solver are varied. Table 5.1 shows the run times and Table 5.2 shows the leading edge position at $\tau = 100$ for each discretisation. It appears that the solver is converging, but the solving times increase prohibitively if spatial or temporal resolutions are decreased further. As the leading edge position varies by as much as 0.02 depending on the discretisation, it is likely that the deviation seen in the bottom right plot of Figure 5.1 arises due to the solver not having converged, however this could be addressed in future work by using a smaller time step, for example.

5.3 General force law model

In this section we generalise the mechanistic one-dimensional cell-based model to account for a more general force law, $F_{i,i\pm 1}^i$, between neighbouring nodes i and $i \pm 1$.

dx/dt	0.01	0.005	0.004	0.002	0.001
0.01	42.9669	42.9580	42.9561	42.9522	42.9501
0.005	42.9629	-	-	-	-
0.004	42.9618	-	-	-	-
0.002	42.9596	-	-	-	-
0.001	42.9594	-	-	-	42.9462

Table 5.2: The leading edge position of the solved PDE system in Section 5.4.6 at $\tau = 100$, using the algorithms outlined in Appendix D.2, for varying space (dx) and time (dt) resolutions. The ODE solution has a leading edge position of 42.9429. The solver uses $N = 45$ cells with node positions uniformly distributed in $[0, 30]$, with $a = 1$, $\alpha = 15$, a linear force model, and no proliferation. Dashes indicate spatial/temporal discretisations not considered.

Force model	Force law (F_i)	Diffusion term ($D(q)$)
Linear	$k(a - L)$	$\frac{\alpha}{q^2}$
Cubic	$k(a - L)^3$	$\frac{3\alpha}{q^2}(a - \frac{1}{q})^2$
Hertz	$k(a - L)^{3/2}$	$\frac{3\alpha}{2q^2}(a - \frac{1}{q})^{1/2}$

Table 5.3: The force laws we consider and the corresponding diffusion term in the cell density model. These force laws, scaled appropriately, are shown in Figure 5.2, and cover a wide range of potential interaction phenotypes. Diffusion terms relate to discussions in Chapter 5.

In our examples, we will work with a force law of the form

$$F_{i,j}^{int} = F(r_i - r_j) = k(a - |r_i - r_j|)^n \frac{r_i - r_j}{|r_i - r_j|},$$

for some real valued exponent, n . The force laws considered have $n = 1$ (for a linear force law), $n = 3$ (for a cubic force law) or $n = 3/2$ (for a Hertz force law). These are shown in Table 5.3 and visualised in Figure 5.2 and are chosen to cover a wide range of potential cell interactions.

We maintain generality in the force law in the derivation following equation (5.1), where node positions, r_i are allowed to move according to the forces on them from

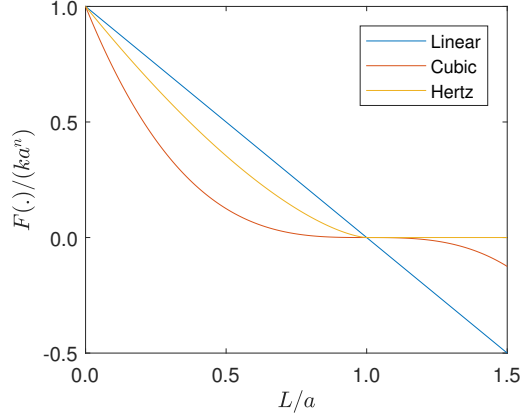


Figure 5.2: The force laws we consider, whose formulas are given in Table 5.3. They are shown with appropriate scaling to allow ease of comparison. These force laws cover a wide range of potential interaction phenotypes.

neighbouring cells, as shown in Figure 4.2, to give,

$$r_0 = 0, \quad (5.16)$$

$$\eta \frac{dr_i}{dt} = F(r_i - r_{i-1}) - F(r_{i+1} - r_i), \quad i = 1, \dots, N-1, \quad (5.17)$$

$$\eta \frac{dr_N}{dt} = F(r_N - r_{N-1}). \quad (5.18)$$

where $F = k(a - r)^n$. The initial conditions are $r_i(0) = r_i^0$ for $i = 1, \dots, N$. As for the linear force law, we solve for the position of each node numerically using a simple forward Euler method with time step $dt = 0.001$.

5.3.1 Continuum approximation

We again extend the positions, $r_i(t)$, to a smooth function $r(i, t)$, for $i \in [0, N]$, before non-dimensionalising equations (5.16)-(5.18) using similar scalings to the simple linear case,

$$\tilde{i} = \frac{i}{N}, \quad \tilde{r} = \frac{r}{aN}, \quad \tilde{t} = \frac{\alpha a^{n-1} t}{N^2} = \frac{ka^{n-1} t}{\eta N^2},$$

so that $\tilde{i} = \mathcal{O}(1)$ and $\tilde{r} = \mathcal{O}(1)$. We also non-dimensionalise the force law to be

$$\tilde{F}(\cdot) = (1 - \cdot)^n. \quad (5.19)$$

To derive an equivalent coarse-grained approximation we proceed as the derivation in the previous section, performing a Taylor expansion about $\tilde{i}$ to give, on neglecting terms that are $\mathcal{O}(1/N^2)$,

$$\tilde{r}(0, \tilde{t}) = 0, \quad (5.20)$$

$$\frac{\partial \tilde{r}(\tilde{i}, \tilde{t})}{\partial \tilde{t}} = -\frac{\partial^2 \tilde{r}}{\partial \tilde{i}^2} \tilde{F}'\left(\frac{\partial \tilde{r}}{\partial \tilde{i}}\right), \quad \tilde{i} \in (0, 1). \quad (5.21)$$

$$\frac{1}{N} \frac{\partial \tilde{r}(1, \tilde{t})}{\partial \tilde{t}} = \tilde{F}\left(\frac{\partial \tilde{r}}{\partial \tilde{i}}\right) - \frac{1}{2N} \frac{\partial^2 \tilde{r}}{\partial \tilde{i}^2} \tilde{F}'\left(\frac{\partial \tilde{r}}{\partial \tilde{i}}\right). \quad (5.22)$$

Rewriting in terms of dimensional variables (where again we use $\hat{i}$ as a continuum approximation of i and drop the hat), gives the following diffusion equation,

$$\eta \frac{\partial r(i, t)}{\partial t} = -\frac{\partial^2 r}{\partial i^2} F'\left(\frac{\partial r}{\partial i}\right), \quad i \in (0, N). \quad (5.23)$$

along with the boundary conditions,

$$r = 0, \text{ for } i = 0 \quad \text{and} \quad \eta \frac{\partial r}{\partial t} = F\left(\frac{\partial r}{\partial i}\right) - \frac{1}{2} \frac{\partial^2 r}{\partial i^2} F'\left(\frac{\partial r}{\partial i}\right), \text{ for } i = N. \quad (5.24)$$

As a consistency check, note that when $n = 1$, as for a linear force law, equations (5.23) and (5.24) reduce to equations (5.5) and (5.11).

5.3.2 Derivation of cell density model

Using the change of variables from equation (5.8), along with the methodology introduced in Section 5.2.3, we can similarly establish a cell density formulation with a general force law. Following a simple substitution of terms from equations (5.8) and (5.9) into equation (5.23) we obtain the PDE,

$$\eta \frac{\partial r}{\partial t} = \frac{F'\left(\frac{1}{q}\right)}{q^3} \frac{\partial q}{\partial r}, \text{ for } i \in (0, N), \quad (5.25)$$

which represents how the domain, r , evolves along characteristics with constant index. Substitution of equation (5.8), and a simply rearrangement and differentiation, identically to earlier arguments, results in a PDE for the cell density,

$$\frac{\partial q}{\partial \tau} = \frac{\partial}{\partial r} \left[D(q) \frac{\partial q}{\partial r} \right], \text{ for } r \in (0, L(\tau)), \quad (5.26)$$

where we have defined the diffusion term, $D(q)$,

$$D(q) = \frac{-F'\left(\frac{1}{q}\right)}{\eta q^2}. \quad (5.27)$$

The characteristic equation (5.25) can then be written as

$$\frac{\partial r}{\partial \tau} = \frac{1}{q} D(q) \frac{\partial q}{\partial r}, \quad r \in [0, L(\tau)].$$

Upon similar substitution into equation (5.24) we obtain boundary conditions,

$$\frac{\partial q}{\partial r} = 0, \text{ for } i = 0 \quad \text{and} \quad \frac{F\left(\frac{1}{q}\right)}{\eta} + \frac{D(q)}{2q} \frac{\partial q}{\partial r} = 0, \text{ for } r = L(\tau). \quad (5.28)$$

Note that this model also leads to conservation of mass. The mass is given by the

integral of the cell density, q . Considering the differentiation of the mass with respect to τ ,

$$\begin{aligned}
\frac{d}{dt} \int_0^{L(\tau)} q(r', \tau) dr' &= \frac{dL(\tau)}{d\tau} q(L(\tau), \tau) + \int_0^{L(\tau)} \frac{\partial q}{\partial \tau}(r', \tau) dr', \\
&= \frac{dL(\tau)}{d\tau} q(L(\tau), \tau) + \int_0^{L(\tau)} \frac{\partial}{\partial r} \left[D(q) \frac{\partial q}{\partial r} \right] dr', \\
&= \left[\frac{qF\left(\frac{1}{q}\right)}{\eta} - \frac{D(q)}{2} \frac{\partial q}{\partial r} + D(q) \frac{\partial q}{\partial r} \right] \Big|_{r=L(\tau)} - D(q) \frac{\partial q}{\partial r} \Big|_{r=0}, \\
&= 0,
\end{aligned}$$

where the final line comes from using equation (5.28). Therefore mass is conserved under these boundary conditions.

5.3.2.1 Numerical solution

In summary we obtain a PDE for the cell density in the domain,

$$\frac{\partial q}{\partial \tau} = \frac{\partial}{\partial r} \left[D(q) \frac{\partial q}{\partial r} \right], \text{ for } r \in (0, L(\tau)), \quad (5.29)$$

along with the boundary conditions

$$\frac{\partial q}{\partial r} \Big|_{r=0} = 0 \quad \text{and} \quad \left[\frac{F\left(\frac{1}{q}\right)}{\eta} + \frac{D(q)}{2q} \frac{\partial q}{\partial r} \right] \Big|_{r=L(\tau)} = 0. \quad (5.30)$$

and an initial condition that is computed from initial node positions as described previously,

$$q(r, 0) = q_0(r) \quad \text{for} \quad r \in (0, L(0)). \quad (5.31)$$

We numerically solve these equations by using a Lagrangian transformation to convert this to a PDE for the density on a fixed domain, and a PDE for the evolving domain. We outline the full method for the more general model with proliferation in Appendix D.2, where an implicit finite difference scheme is used, with Picard iteration.

5.3.3 Comparing the accuracy of the continuum model with different force laws.

In Figure 5.3 we compare predictions of the cell densities and the leading edge given by the cell density model and the node positions model, considering each of the force laws from Table 5.3.

We find that across all force laws tests, the coarse grained PDE is very accurate in the prediction of both the evolution of the cell density, $q(r, \tau)$, and the free boundary, $r = L(\tau)$. The only minor deviation in the predictions of the models is found at the leading edge position, where the gradient in the cell density is largest. Note that in this region the approximation of the density in the cell-based model is a lower order in N , so this deviation could perhaps be expected.

5.4 Introducing a proliferation mechanism

To apply this model to cell spreading data, it is important to account for cell proliferation events. We introduce a proliferation mechanism that assumes a proliferation event introduces a new node mid-way between two nodes (proliferation from spring centre). After this mechanism is established we derive the corresponding continuum formulation followed by the cell density model, by using techniques similar to those used in the previous section.

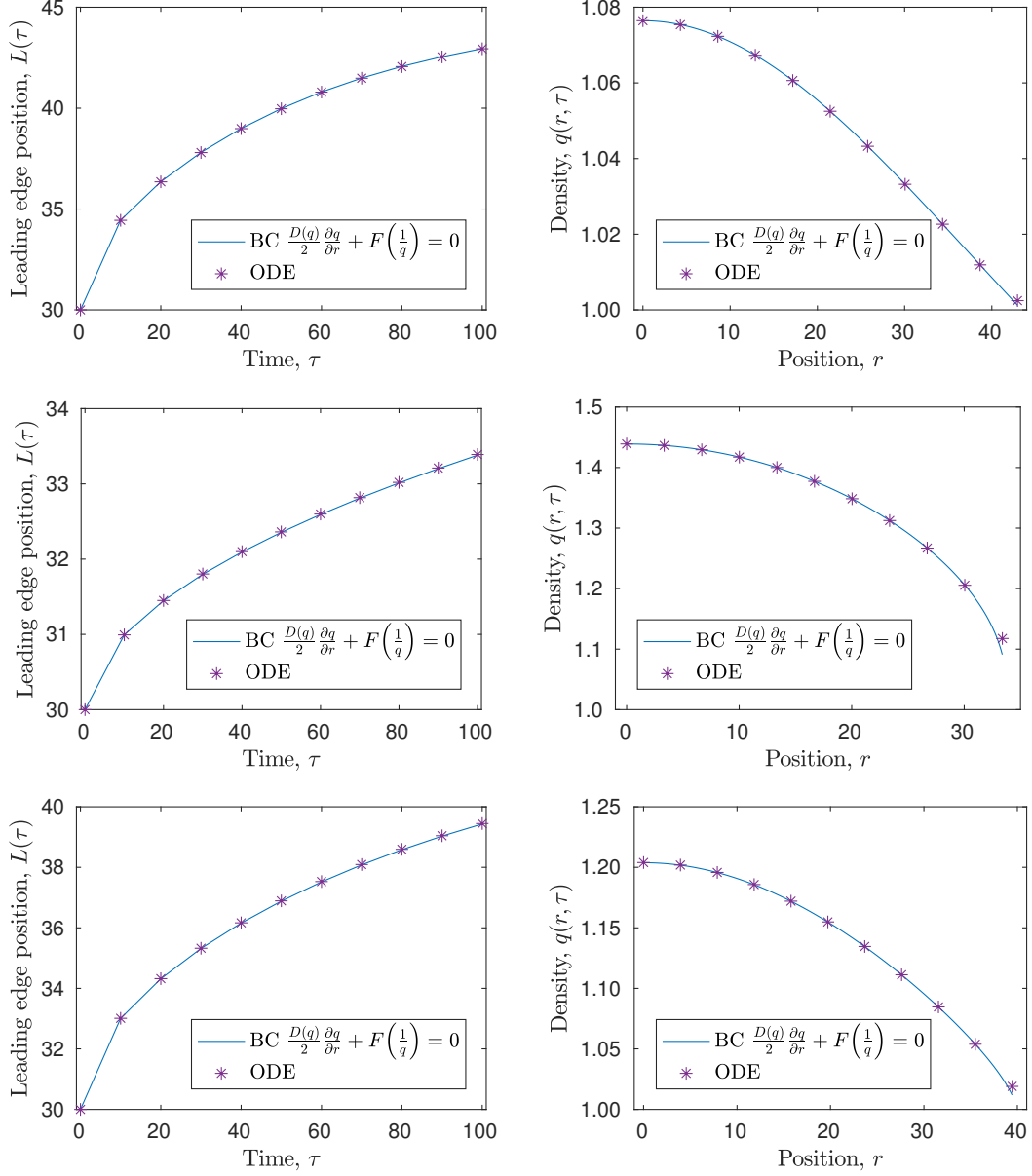


Figure 5.3: Comparison of the leading edge positions (left) and the cell density (right) in the solutions of the cell density PDE in equation (5.29) and the system of ODEs for node positions in equations (5.16) - (5.18). We consider how varying the force law affects the agreement of the models, the top row uses the linear force law, the middle row uses the cubic force law, and the bottom row uses the Hertz force law (see Table 5.3). In each case we $N = 15$ cells uniformly in $R \in (0, 30)$ and keep $aN = 45$ and $\alpha/a^{1-n}N^2 = 135$ fixed (i.e. the time-scale, and length-scale respectively). PDE solutions with the boundary condition $D(q)/2\partial q/\partial r + F(1/q) = 0$ are solid blue lines, and the ODE model solution is shown as purple stars.

5.4.1 Proliferation mechanism

We begin by recalling the proliferation mechanism we adopt. Each cell will proliferate according to a function of its length, i.e. cell i proliferates with rate $G_i = G(r_i - r_{i-1})$. This mechanism is intended to account for the fact that older cells may have grown larger, or that the compression of surrounding cells discourages a cell from proliferating [19].

When a cell proliferates, a new node is introduced at its centre, and we relabel the node indices to ensure their order, that is, $r_i(t) < r_{i+1}(t)$ for $i = 0, \dots, N(t) - 1$, and $t \geq 0$. Subsequently, when a new node is introduced due to the proliferation of the j^{th} cell, we relabel the nodes with index $j + 1, \dots, N$ to indices $j + 2, \dots, N + 1$, respectively, as shown in Figure 5.4.

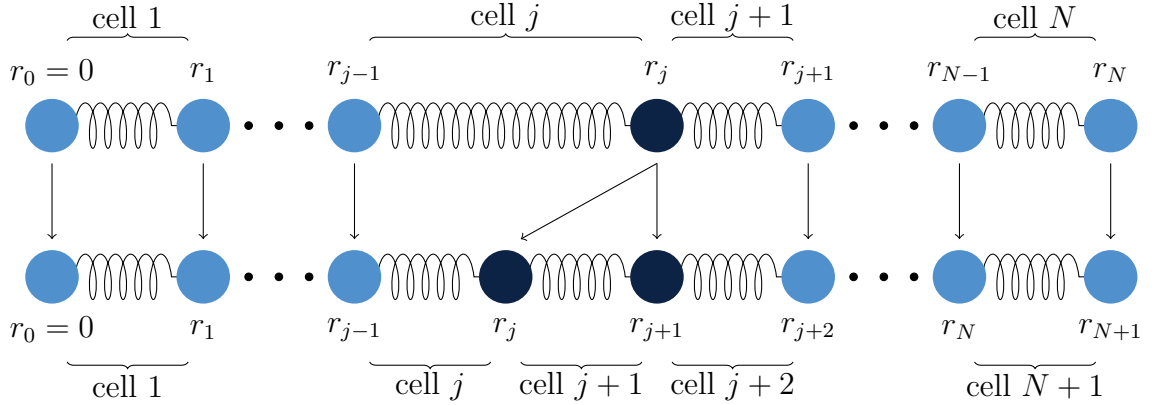


Figure 5.4: A proliferation of cell j causes a new node to be introduced at its centre, and cells are then relabelled to maintain their order ($r_i < r_{i+1}$ for $i \in \{0, \dots, N\}$).

In this work we explore the dynamics introduced by three types of proliferation mechanism: (i) cells proliferate at a constant rate; (ii) cells proliferate at a rate proportional to their length; and (iii) cells proliferate once they have reached a target length. Functional forms of the growth rates are shown in Table 5.4 and visualised in Figure 5.5.

To simulate from the stochastic discrete model, we solve the system of equations (5.16) - (5.18) using a simple forward Euler method, where each time-step $dt = 0.001$,

Proliferation	Proliferation law (G_i)
Constant	β
Linear	$\beta(r_i - r_{i-1})/a$
Hyperbolic	$\beta(\tanh(8((r_i - r_{i-1}) - 3a/4))/2 + 1/2)$

Table 5.4: Definitions of the proliferation laws we consider. A constant proliferation law is considered in the main text, but one could easily apply a law that is a function of the cell lengths. The hyperbolic proliferation law is adopted in supplementary Section C.2.2. These laws are visualised in Figure 5.5.

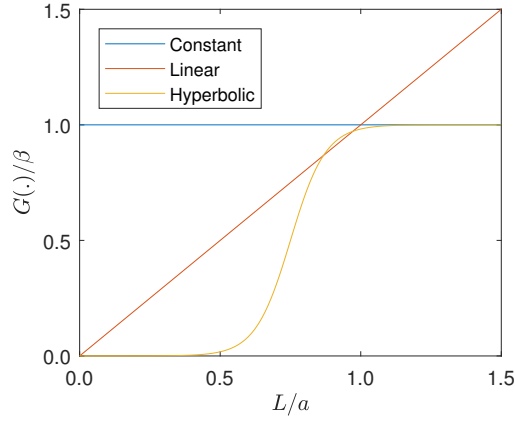


Figure 5.5: The proliferation laws we consider, whose formulas are given in Table 5.4. Here they are shown with appropriate scaling. We use proliferation laws that are representative of cells that either proliferate: independently of cell size; linearly dependent on their size; or only proliferate after reaching a fraction of their resting size (we use $3/4$), approximated using the tanh function.

a proliferation will occur with probability $P_p dt$, where $P_p = \sum_{i=1}^N G_i$. If a proliferation occurs, a new node is introduced at the centre of the proliferating cell, and N is updated ready for the following time-step. Note that this algorithm enforces the condition that at most one cell can proliferate per time-step; which is a reasonable assumption for the parameters used in this work.

5.4.2 Derivation of cell density model with proliferation

As the proliferation mechanism we have introduced is stochastic, we now consider the evolution of the possible locations of the nodes over time. We make progress by considering an infinitesimally small time interval $[t, t + dt)$ and condition on whether cell proliferation takes place during that time interval. The position of each node after a small time interval will depend on if and where cell proliferation occurs and the resulting forces exerted by neighbouring cells. We write, for $i \in \{1, \dots, N - 1\}$,

$$\begin{aligned}
r_i^N(t + dt) = & \left(r_i^N(t) + dt \frac{F_{i,i-1}^N - F_{i,i+1}^N}{\eta} \right) \\
& \times \mathbb{1} \{ \text{no proliferation in } [t, t + dt) \} \\
& \left(r_i^{N-1}(t) + dt \frac{F_{i,i-1}^{N-1} - F_{i,i+1}^{N-1}}{\eta} \right) \\
& \times \mathbb{1} \{ \text{proliferation to the right of cell } i \text{ in } [t, t + dt) \} \\
& + \frac{1}{2} \left[\left(r_i^{N-1}(t) + dt \frac{F_{i,i-1}^{N-1} - F_{i,i+1}^{N-1}}{\eta} \right) + \left(r_{i-1}^{N-1}(t) + dt \frac{F_{i-1,i-2}^{N-1} - F_{i-1,i}^{N-1}}{\eta} \right) \right] \\
& \times \mathbb{1} \{ \text{proliferation of cell } i \text{ in } [t, t + dt) \} \\
& + \left(r_{i-1}^{N-1}(t) + dt \frac{F_{i-1,i-2}^{N-1} - F_{i-1,i}^{N-1}}{\eta} \right) \\
& \times \mathbb{1} \{ \text{proliferation to the left of cell } i \text{ in } [t, t + dt) \} \\
& + (\text{position of cell } i \text{ if more than one proliferation event in } [t, t + dt)) \\
& \times \mathbb{1} \{ \text{more than one cell proliferation event in } [t, t + dt) \}, \quad (5.32)
\end{aligned}$$

and

$$\begin{aligned}
r_N^N(t + dt) = & \left(r_N^N(t) + dt \frac{F_{N,N-1}^N}{\eta} \right) \\
& \times \mathbb{1} \{ \text{no proliferation in } [t, t + dt) \} \\
& \left(r_N^{N-1}(t) + dt \frac{F_{N,N-1}^{N-1}}{\eta} \right) \\
& \times \mathbb{1} \{ \text{proliferation to the right of cell } N \text{ in } [t, t + dt) \} \\
& + \frac{1}{2} \left[\left(r_N^{N-1}(t) + dt \frac{F_{N,N-1}^{N-1}}{\eta} \right) + \left(r_{N-1}^{N-1}(t) + dt \frac{F_{N-1,N-2}^{N-1} - F_{N-1,N}^{N-1}}{\eta} \right) \right] \\
& \times \mathbb{1} \{ \text{proliferation of cell } N \text{ in } [t, t + dt) \} \\
& + \left(r_{N-1}^{N-1}(t) + dt \frac{F_{N-1,N-2}^{N-1} - F_{N-1,N}^{N-1}}{\eta} \right) \\
& \times \mathbb{1} \{ \text{proliferation to the left of cell } N \text{ in } [t, t + dt) \} \\
& + (\text{position of cell } N \text{ if more than one proliferation event in } [t, t + dt)) \\
& \times \mathbb{1} \{ \text{more than one cell proliferation event in } [t, t + dt) \}. \quad (5.33)
\end{aligned}$$

In equations (5.32) and (5.33) and $\mathbb{1}$ is the indicator function and $r_i^N(t)$ is the position of cell i at time t when there are N cells. The superscript N on the force terms, $F_{i,j}$, indicate that they are evaluated using the positions of cells i and j when there are N cells.

The required probabilities to specify the indicator functions are

$$\mathbb{P}(\text{no proliferation in } [t, t + dt)) = 1 - dt \sum_{j=1}^N G(r_j - r_{j-1}), \quad (5.34)$$

$$\mathbb{P}(\text{proliferation to the right of cell } i \text{ in } [t, t + dt)) = \begin{cases} dt \sum_{j=i+1}^N G(r_j - r_{j-1}) & j \neq N, \\ 0 & j = N, \end{cases} \quad (5.35)$$

$$(5.36)$$

$$\mathbb{P}(\text{proliferation of cell } i \text{ in } [t, t + dt)) = dt G(r_i - r_{i-1}), \quad (5.37)$$

$$\mathbb{P}(\text{proliferation to the left of cell } i \text{ in } [t, t + dt)) = \begin{cases} dt \sum_{j=1}^{i-1} G(r_j - r_{j-1}) & j \neq 1, \\ 0 & j = 1, \end{cases} \quad (5.38)$$

$$(5.39)$$

and

$$\mathbb{P}(\text{more than one cell proliferation event in } [t, t + dt)) = \mathcal{O}(dt^2). \quad (5.40)$$

We now take expectations on both sides of equations (5.32) and (5.33), denoting by $\langle r_i^N(t) \rangle$ the expected position of node i at time t when there are N nodes. We then make two simplifying assumptions: (i) that $\langle r_i^N(t) \rangle$ is a continuous function of time; and (ii) that $\langle F(r_i^N(t)) \rangle = F(\langle r_i^N(t) \rangle)$ and $\langle G(r_i^N(t)) \rangle = G(\langle r_i^N(t) \rangle)$. The former assumes that the cell size is small relative to the tissue, and allows us to rearrange and take the limit as $dt \rightarrow 0$ and, together with the latter (the standard

mean-field approximation), we have

$$\langle r_0^N(t) \rangle = 0, \quad (5.41)$$

$$\begin{aligned} \eta \frac{d}{dt} \langle r_i^N(t) \rangle &= F(\langle r_i^N(t) \rangle - \langle r_{i-1}^N(t) \rangle) - F(\langle r_{i+1}^N(t) \rangle - \langle r_i^N(t) \rangle) \\ &\quad - \eta \langle r_i^{N-1}(t) \rangle \sum_{j=1}^i G(\langle r_j^{N-1}(t) \rangle - \langle r_{j-1}^{N-1}(t) \rangle) \\ &\quad + \frac{\eta}{2} (\langle r_i^{N-1}(t) \rangle + \langle r_{i-1}^{N-1}(t) \rangle) G(\langle r_i^{N-1}(t) \rangle - \langle r_{i-1}^{N-1}(t) \rangle) \\ &\quad + \eta \langle r_{i-1}^{N-1}(t) \rangle \sum_{j=1}^{i-1} G(\langle r_j^{N-1}(t) \rangle - \langle r_{j-1}^{N-1}(t) \rangle), \quad i = 1, \dots, N-1, \end{aligned} \quad (5.42)$$

$$\begin{aligned} \eta \frac{d}{dt} \langle r_N^N(t) \rangle &= F(\langle r_N^N(t) \rangle - \langle r_{N-1}^N(t) \rangle) \\ &\quad - \eta \langle r_N^{N-1}(t) \rangle \sum_{j=1}^N G(\langle r_j^{N-1}(t) \rangle - \langle r_{j-1}^{N-1}(t) \rangle) \\ &\quad + \frac{\eta}{2} (\langle r_N^{N-1}(t) \rangle + \langle r_{N-1}^{N-1}(t) \rangle) G(\langle r_N^{N-1}(t) \rangle - \langle r_{N-1}^{N-1}(t) \rangle) \\ &\quad + \eta \langle r_{N-1}^{N-1}(t) \rangle \sum_{j=1}^{N-1} G(\langle r_j^{N-1}(t) \rangle - \langle r_{j-1}^{N-1}(t) \rangle). \end{aligned} \quad (5.43)$$

5.4.3 Continuum approximation

To enable a continuum approximation to be formulated we make the further approximation $\langle r_i^N(t) \rangle = \langle r_i^{N-1}(t) \rangle$, and to simplify exposition going forward, we will drop

use of the angle brackets. After algebraic simplification, we have

$$r_0(t) = 0, \quad (5.44)$$

$$\begin{aligned} \eta \frac{dr_i}{dt} = & F(r_i - r_{i-1}) - F(r_{i+1} - r_i) \\ & - \eta(r_i - r_{i-1}) \left[\sum_{j=1}^{i-1} G(r_j - r_{j-1}) + \frac{1}{2} G(r_i - r_{i-1}) \right], \quad i = 1, \dots, N-1, \end{aligned} \quad (5.45)$$

$$\begin{aligned} \eta \frac{dr_N}{dt} = & F(r_N - r_{N-1}) \\ & - \eta(r_N - r_{N-1}) \left[\sum_{j=1}^{N-1} G(r_j - r_{j-1}) + \frac{1}{2} G(r_N - r_{N-1}) \right]. \end{aligned} \quad (5.46)$$

To make progress in deriving an equivalent continuum, coarse-grained model, we proceed as before: we extend node position, $r_i(t)$, to a smooth function $r(i, t)$ for $i \in [0, N(t)]$, and non-dimensionalise using the scalings

$$\tilde{i} = \frac{i}{N_0}, \quad \tilde{r} = \frac{r}{aN_0}, \quad \tilde{N} = \frac{N}{N_0}, \quad \tilde{t} = \frac{tka^{n-1}}{\eta N_0^2},$$

where $N_0 = N(0)$, the number of nodes at $t = 0$. We also define the non-dimensional proliferation function such that

$$\tilde{G}(\cdot) = \frac{\eta N_0}{ka^{n-1}} G(a \cdot), \quad (5.47)$$

and the non-dimensional force function is as specified in equation (5.19).

We then work in the same manner as before, using Taylor expansion together with quadrature approximations of the form

$$\sum_{j=1}^{i-1} G(r_j - r_{j-1}) + \frac{1}{2} G(r_i - r_{i-1}) \approx \int_{j=0}^i G(r_j - r_{j-1}) dj, \quad (5.48)$$

to give, upon neglecting terms that are $\mathcal{O}(1/N_0^2)$,

$$\frac{\partial \tilde{r}}{\partial t} + \frac{\partial \tilde{r}}{\partial \tilde{i}} \int_0^{\tilde{i}} \tilde{G} \left(\frac{\partial \tilde{r}}{\partial \tilde{j}} \right) d\tilde{j} = -\tilde{F}' \left(\frac{\partial \tilde{r}}{\partial \tilde{i}} \right) \frac{\partial^2 \tilde{r}}{\partial \tilde{i}^2} + \frac{1}{2N_0} \frac{\partial^2 \tilde{r}}{\partial \tilde{i}^2} \int_0^{\tilde{i}} \tilde{G} \left(\frac{\partial \tilde{r}}{\partial \tilde{j}} \right) d\tilde{j}, \quad \tilde{i} \in [0, \tilde{N}(\tilde{t})]. \quad (5.49)$$

The left-hand boundary condition remains as $\tilde{r}(0, \tilde{t}) = 0$ and, once again, we derive the right-hand boundary condition by Taylor expanding and neglecting terms that are $\mathcal{O}(1/N^2)$ to give, at $\tilde{i} = \tilde{N}(\tilde{t})$,

$$\frac{1}{N_0} \frac{\partial \tilde{r}}{\partial \tilde{t}} = \tilde{F} \left(\frac{\partial \tilde{r}}{\partial \tilde{i}} \right) - \frac{1}{2N_0} \tilde{F}' \left(\frac{\partial \tilde{r}}{\partial \tilde{i}} \right) \frac{\partial^2 \tilde{r}}{\partial \tilde{i}^2} - \frac{1}{N_0} \frac{\partial \tilde{r}}{\partial \tilde{i}} \int_0^{\tilde{N}} \tilde{G} \left(\frac{\partial \tilde{r}}{\partial \tilde{j}} \right) d\tilde{j}. \quad (5.50)$$

Rewriting equation (5.49) in terms of the dimensional variables we have the following PDE for $r(i, t)$:

$$\eta \frac{\partial r}{\partial t} + \eta \frac{\partial r}{\partial i} \int_0^i G \left(\frac{\partial r}{\partial j} \right) dj = -F' \left(\frac{\partial r}{\partial i} \right) \frac{\partial^2 r}{\partial i^2} + \frac{\eta}{2} \frac{\partial^2 r}{\partial i^2} \int_0^i G \left(\frac{\partial r}{\partial j} \right) dj \quad i \in (0, N(t)). \quad (5.51)$$

The boundary conditions are

$$r(0, t) = 0 \quad \text{and} \quad \eta \frac{\partial r}{\partial t} \Big|_{i=N(t)} = \left[F \left(\frac{\partial r}{\partial i} \right) - \frac{1}{2} \frac{\partial^2 r}{\partial i^2} F' \left(\frac{\partial r}{\partial i} \right) - \eta \frac{\partial r}{\partial i} \int_0^i G \left(\frac{\partial r}{\partial j} \right) dj \right] \Big|_{i=N(t)}. \quad (5.52)$$

As before, the initial conditions can be specified by extending the discrete initial conditions, $r_i(0) = r_i^0$ for $i = 1, \dots, N(0)$, to a continuous function $r(i, 0)$ such that $r(i, 0) = r_i^0$ for $i = 1, \dots, N(0)$.

5.4.4 Derivation of the corresponding cell density model

We now establish a PDE describing the evolution of cell density with position, r , and time, t , for a proliferative cell population with general force and proliferation laws. Changing variables from (i, t) to (r, τ) , as before, with a simple substitution of terms from equations (5.8) into equation (5.51) we obtain the PDE

$$\eta \frac{\partial r}{\partial \tau} + \frac{\eta}{q} \int_0^r q G\left(\frac{1}{q}\right) dy = \frac{1}{q^3} F'\left(\frac{1}{q}\right) \frac{\partial q}{\partial r} - \frac{\eta}{2q^3} \frac{\partial q}{\partial r} \int_0^r q G\left(\frac{1}{q}\right) dy, \quad r \in [0, L(\tau)], \quad (5.53)$$

which represents how the domain evolves along the characteristics. Note that, due to proliferation, this is no longer equivalent to following constant cell index, i .

After further rearrangement, as before, we have

$$\frac{\partial q}{\partial \tau} = \frac{\partial}{\partial r} \left([D(q) + E(q)] \frac{\partial q}{\partial r} \right) + q G\left(\frac{1}{q}\right), \quad r \in (0, L(\tau)), \quad (5.54)$$

where

$$D(q) = -\frac{1}{\eta q^2} F'\left(\frac{1}{q}\right) \quad \text{and} \quad E(q) = \frac{1}{2q^2} \int_0^r q G\left(\frac{1}{q}\right) dy. \quad (5.55)$$

Under the same change of variables, the boundary conditions become

$$\left. \frac{\partial q}{\partial r} \right|_{r=0} = 0 \quad \text{and} \quad \left[\frac{1}{\eta} F\left(\frac{1}{q}\right) + \left(\frac{D(q)}{2q} + \frac{E(q)}{q} \right) \frac{\partial q}{\partial r} \right] \Big|_{r=L(\tau)} = 0, \quad (5.56)$$

and the initial conditions, computed as in Section 5.2.2, are

$$q(r, 0) = q_0(r), \quad r \in (0, L(0)). \quad (5.57)$$

Using equation (5.55), characteristic equation (5.53) can be re-written as

$$\frac{Dr}{D\tau} = \frac{\partial r}{\partial \tau} + \frac{1}{q} \int_0^r q G\left(\frac{1}{q}\right) dy = -\frac{1}{q} [D(q) + E(q)] \frac{\partial q}{\partial r}, \quad r \in [0, L(\tau)], \quad (5.58)$$

That the left-hand side of equation (5.58) constitutes a material derivative can be seen by following a small “tissue element” as the cell population grows and divides.

We have $r = r(i(t), t)$ with

$$\frac{Dr}{D\tau} = \frac{\partial r}{\partial \tau} + \frac{\partial r}{\partial i} \frac{\partial i}{\partial \tau} = \frac{\partial r}{\partial \tau} + \frac{1}{q} \int_0^r q G\left(\frac{1}{q}\right) dy, \quad (5.59)$$

where we have used the fact that the rate of change of cell index is equal to the rate of cell proliferation in the region to the left of the cell *i.e.*

$$\frac{\partial i}{\partial \tau} = \int_0^r q G\left(\frac{1}{q}\right) dy. \quad (5.60)$$

Finally, using equation (5.58) we can specify the rate of growth of the domain over time as

$$\frac{dL(\tau)}{d\tau} = \left(-\frac{1}{q} [D(q) + E(q)] \frac{\partial q}{\partial r} \right) \Big|_{L(\tau)}. \quad (5.61)$$

5.4.5 Evolution of cell number

Note that, since cell proliferation is now present in the model, the cell number changes over time and the system does not conserve mass. The cell number at time τ is specified by equation (5.7) as

$$N(\tau) = \int_0^{L(\tau)} q(r, \tau) dr. \quad (5.62)$$

Differentiating with respect to τ and using the left-hand boundary condition (5.66) gives

$$\begin{aligned}
\frac{dN}{d\tau} &= \frac{dL}{d\tau} q(L(\tau), \tau) + \int_0^{L(\tau)} \frac{\partial q}{\partial \tau}(r, \tau) dr, \\
&= \frac{dL}{d\tau} q(L(\tau), \tau) + \int_0^{L(\tau)} \left\{ \frac{\partial}{\partial r} \left([D(q) + E(q)] \frac{\partial q}{\partial r} \right) + q G\left(\frac{1}{q}\right) \right\} dr, \\
&= \left[\frac{dL}{d\tau} q + [D(q) + E(q)] \frac{\partial q}{\partial r} \right] \Big|_{r=L(\tau)} + \int_0^{L(\tau)} q G\left(\frac{1}{q}\right) dr, \tag{5.63}
\end{aligned}$$

Substituting equation (5.61) into equation (5.63) gives the rate of change of cell number over time as

$$\frac{dN}{d\tau} = \int_0^{L(\tau)} q G\left(\frac{1}{q}\right) dy. \tag{5.64}$$

This equation states simply and intuitively that the rate of change of cell number is simply equal to the sum of the proliferation rates of each cell.

5.4.6 Numerical solution

In summary, the coarse-grained model consists of a PDE for the evolution of cell density

$$\frac{\partial q}{\partial \tau} = \frac{\partial}{\partial r} \left([D(q) + E(q)] \frac{\partial q}{\partial r} \right) + q G\left(\frac{1}{q}\right), \quad r \in (0, L(\tau)), \tag{5.65}$$

together with boundary conditions

$$\left. \frac{\partial q}{\partial r} \right|_{r=0} = 0 \quad \text{and} \quad \left[\frac{1}{\eta} F\left(\frac{1}{q}\right) + \left(\frac{D(q)}{2q} + \frac{E(q)}{q} \right) \frac{\partial q}{\partial r} \right] \Big|_{r=L(\tau)} = 0, \tag{5.66}$$

and initial condition

$$q(r, 0) = q_0(r), \quad r \in (0, L(0)). \quad (5.67)$$

The characteristic equation is

$$\frac{Dr}{D\tau} = \frac{\partial r}{\partial \tau} + \frac{1}{q} \int_0^r q G\left(\frac{1}{q}\right) dy = -\frac{1}{q} [D(q) + E(q)] \frac{\partial q}{\partial r}, \quad r \in [0, L(\tau)], \quad (5.68)$$

and we have

$$\frac{dL(\tau)}{d\tau} = \left(-\frac{1}{q} [D(q) + E(q)] \frac{\partial q}{\partial r} \right) \Big|_{L(\tau)}. \quad (5.69)$$

As in Section 5.2.3.1 and Section 5.3.2.1, in order to solve the coarse-grained model numerically, we employ a Lagrangian transformation to map the free boundary problem to a fixed domain: we let $\tau = T$ and

$$r = \Gamma(R, T) \quad \text{with} \quad R = \Gamma(R, 0), \quad 0 = \Gamma(0, T), \quad L(T) = \Gamma(L(0), T), \quad (5.70)$$

to give

$$\frac{\partial \Gamma}{\partial T} = -\frac{D(q) + E(q)}{q\Gamma_R} \frac{\partial q}{\partial R}, \quad R \in (0, L(0)), \quad (5.71)$$

$$\frac{\partial q}{\partial T} - \frac{1}{\Gamma_R} \frac{\partial \Gamma}{\partial T} \frac{\partial q}{\partial R} = \frac{1}{\Gamma_R} \frac{\partial}{\partial R} \left(\frac{D(q) + E(q)}{\Gamma_R} \frac{\partial q}{\partial R} \right) + q G\left(\frac{1}{q}\right), \quad R \in (0, L(0)). \quad (5.72)$$

The initial and boundary conditions for $\Gamma(R, T)$ are specified in equation (5.70), and

for $q(R, T)$ we have

$$\left. \frac{\partial q}{\partial R} \right|_{R=0} = 0 \quad \text{and} \quad \left[\frac{1}{\eta} F \left(\frac{1}{q} \right) + \frac{1}{q \Gamma_R} \left(\frac{D(q)}{2} + E(q) \right) \frac{\partial q}{\partial R} \right] \bigg|_{R=L(0)} = 0, \quad (5.73)$$

together with

$$q(R, 0) = q_0(R) \quad R \in (0, L(0)). \quad (5.74)$$

We solve the model numerically using an implicit finite difference method with Picard iteration. Full details are given in D.2.

Note that the second boundary condition is equivalent to requiring the domain PDE holds on the boundary.

A table of force laws, F , diffusion terms, D and proliferation laws, G , along with their visualisation can be seen in Figure 5.2 and Figure 5.5.

5.4.7 Comparing the cell density model with the discrete model

We compare the PDE solution and the average of 100 realisations of the stochastic discrete model, with each force law and proliferation law in Figure 5.2, and although we could keep track of the positions of all cells or the cell density in the domain, we instead track the simpler statistics, leading edge positions (Figure 5.6) and the number of cells over time (Figure 5.7). We initialise with $N = 30$ cells spaced uniformly in $R \in (0, 30)$, with $a = 1$, $\alpha = 15$, and $\beta = 0.001$, so that in the absence of proliferation there would be no change.

By visual inspection, there very good agreement between the two models for the leading edge and number of cells for each force law and proliferation law. The different laws result in varying behaviours in both cases, in particular how quickly the leading

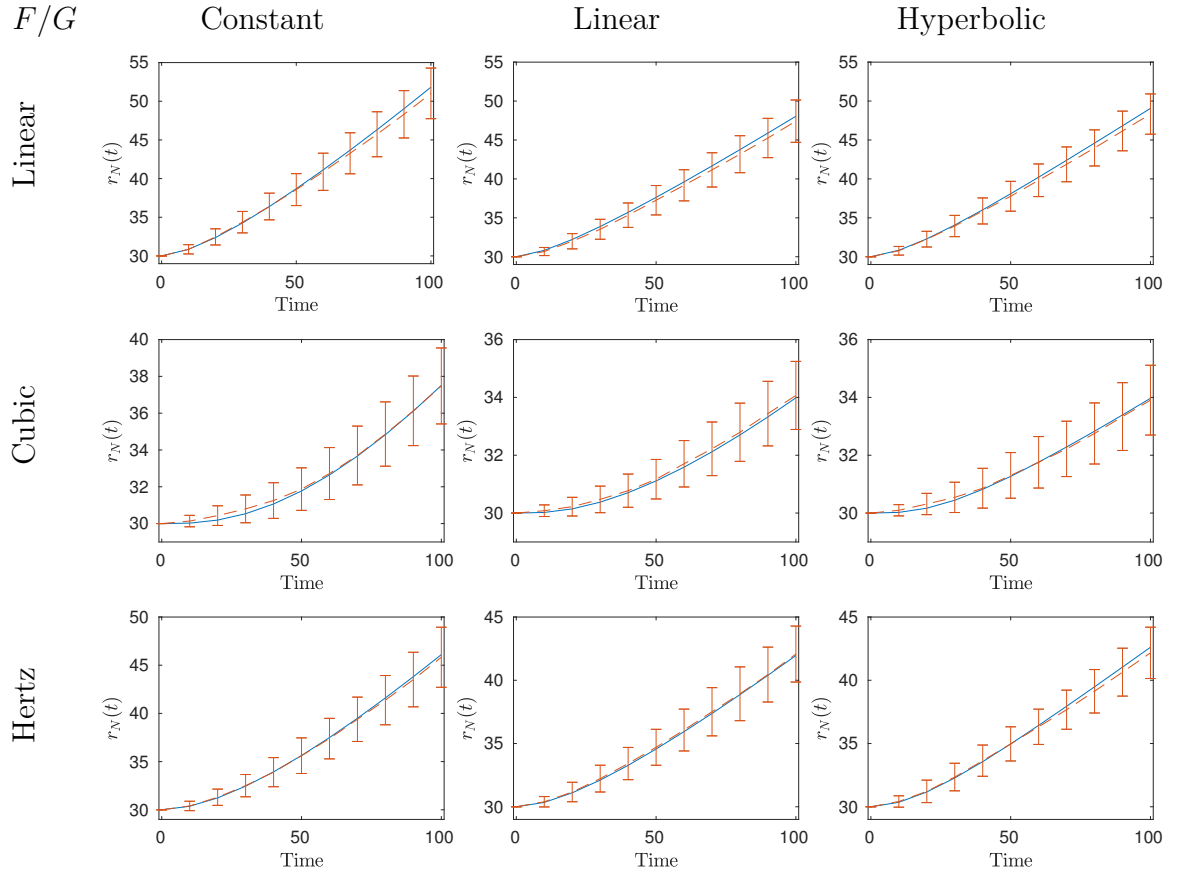


Figure 5.6: The position of the leading edge over time for varying force and proliferation laws. Each row represents a different force law, whereas each column represents a different proliferation law. Each law is visualised in Figures 5.2 and 5.5. The red dashed line is the average of 100 realisations of the stochastic discrete model (and their error bars), whereas the solid blue line is the solution of the full system from Section 5.4.6.

edge expands or how rapidly the number of cells grows. In all circumstances we are left with relatively large error-bars on the data, suggesting that the introduction of the stochastic proliferation has a relatively large impact on the cell density, even though the proliferation parameter is small relative to the motility parameter in these examples. The agreement of the models for varying α/β (doubled and halved) is also very good (results not shown).

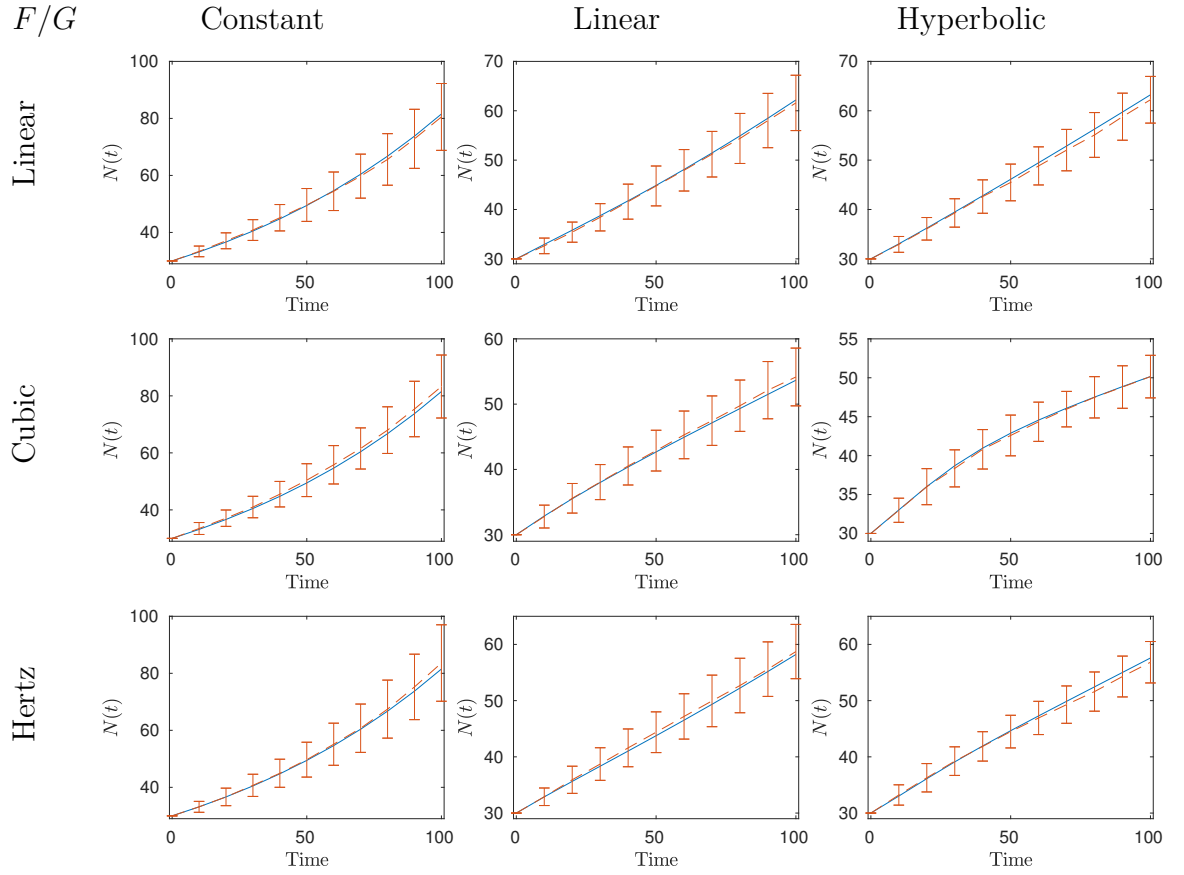


Figure 5.7: Figure of the number of cells over time for varying force and proliferation laws. Each row represents a different force law, whereas each column represents a different proliferation law. Each law is visualised in Figures 5.2 and 5.5. The red dashed line is the average of 100 realisations of the stochastic discrete model (and their error bars), whereas the solid blue line is the solution of the full system from Section 5.4.6.

5.5 Discussion

In this chapter we have derived a continuum model from the discrete off-lattice model introduced by Murray et al. [57], using non-dimensionalisation and an expansion for large N similar to Fozard et al. [29]. With this we have derived a boundary condition that is more accurate than has been previously established by keeping higher order terms from the Taylor expansions. We have also introduced a general proliferation scheme akin to the general force model from Murray et al. [58], and derived a continuum model from this with similar boundary conditions. The accuracy of the continuum model is tested against the stochastic discrete model for a variety of force laws and proliferation laws, and shows very good agreement well within the standard errors.

We note that the continuum model is unfortunately non-trivial to solve numerically, requiring a small time and space step in the numerical solver. In particular, averaging simulations from the IBM, such as in Chapter 4 is quicker than solving the PDE model. As there is a large computational cost in some inference methods such as ABC or particle Markov Chain Monte Carlo [4, 8], it would be useful for future applications of this model with experimental data to be preceded by improving the speed of the PDE solver (for example with an adaptive time-stepping), or solving the discrete model more efficiently with parallelisation of the simulation, for example.

In Chapter 4 we applied ABC-SMC to estimate model parameters of experimental data, however we restricted our attention to a simple linear force law and constant proliferation law. We have seen in Figure 5.7 that different proliferation laws lead to very different profiles of the number of cells in time, so it is possible that different proliferation laws could be inferred using a number of cells statistic. Motivation in using a more complex proliferation law stems from the constant rate proliferation failing to account for cell age or suppression of proliferation due to surrounding mechanical stresses [19].

Similarly for the leading edge statistic, we see in Figure 5.6 that different force laws result in different leading edge trajectories, so it may be possible to infer an appropriate force law for this model framework. Through model selection methods [99, 100], we could hope to understand how different cells choose when to proliferate and what force law best describes cell-cell interactions.

In the following section we recap the results shown in this thesis, and discuss future extensions that are possible.

Chapter 6

Conclusions and discussion

In this thesis we have utilised various ABC approaches in order to understand how user choices impact our ability to accurately infer parameters in different cells spreading models. We have seen that experimental geometry, summary statistics, and the initial cell density all have an impact on the information one can obtain from experimental data. We have considered a multitude of cell spreading models in our investigations, accounting for mechanisms of motility, proliferation, biased motility and cell-cell interaction, and in each case, we have inferred model parameters using both synthetically generated data and experimental data. The experimental data considered has varied from *growth-to-confluence* assays, *scratch assays* and *barrier assays*. The inference methods we have adopted use the ABC methodology, subverting the need to evaluate a likelihood function, with the specific method in each chapter chosen depending on the computational demand of the model under consideration. In addition, we have considered coarse graining of an off-lattice model that includes mechanisms of cell-cell feedback, cell growth and proliferation.

In this final chapter, we review the conclusions from each chapter, before discussing potential future avenues of investigation to consider.

6.1 Conclusions

In Chapter 2 we used a volume excluding lattice-based model, along with both synthetic and experimental data, to determine what experimental geometries help facilitate accurate parameter inference. We found that in growth-to-confluence experiments, cell trajectory information is necessary to accurately determine the motility parameter, whereas for scratch assays, it is possible to infer both model parameters with spatial correlations statistics. Therefore, time-consuming manual cell tracking procedures are not necessary to infer both model parameters, accounting for proliferation and motility, in these experiments. We also described procedures for combining statistics to increase the accuracy of parameter estimates. Finally, we determined that the ABC-DC inference approach is an effective means of approximating MLEs of model parameters with greater efficiency.

In Chapter 3 we considered a similar model, but now accounting for motility bias due to the influence of an electric field. We considered experimental assays in various shapes, and found a large square highly confluent group of cells provided the most informative posterior distributions out of those considered, however, as these results did not keep the cell number fixed that may not be the case in general. Further the position of the leading edge or position of centre of mass statistics, though not previously usable with the *growth-to-confluence* assays of Chapter 2, significantly improved the accuracy of parameter estimates.

In Chapter 4, we applied ABC-SMC to an off-lattice model that accounted for cell growth and cell contact. We used synthetic experiments in this model to consider a plethora of experimental changes that could improve parameter inference, and hence be used to guide future experimental designs. We determined that a combination of experiments, whereby cells lie on a membrane that is uniformly stretched or compressed, is necessary to determine all model parameters. We also inferred parameters of the model for experimental data, and found that the posterior distribution from

one experiment can be used to predict statistics of different initial configurations.

In Chapter 5, we coarse grained the model introduced in Chapter 4, to derive a continuum formulation for the cell density for a general force law and a general proliferation law. Through non-dimensionalisation of the model equations, and Taylor expansions in the limit of large N , we establish cell density equations that can provide accurate estimates of the average behaviour of the discrete model for every force and proliferation law considered.

6.2 Extensions

The work in this thesis goes some way towards investigating how parameter inference can be applied to gain a deeper understanding of cell biology, by considering methods that establish whether a model is appropriate to capture and predict experimental data. However, in each chapter, there are still multiple avenues that can be considered to extend our work.

The two-dimensional lattice-based model from Chapter 2 was adopted for its simplicity and its ease of application to the relatively computationally intensive ABC rejection. Future work could consider a more complex model, however more efficient Bayesian methodologies should also be adopted if the model becomes computationally intensive, for example ABC-SMC or ABC-DC. The simplicity of this model allows the extension to include mechanisms such as the interaction of different cell populations, which is particularly relevant in the invasion of cancer cells, for example [3].

In Chapter 3, we demonstrated how parameter inference can be attempted with an extension of the two-dimensional lattice-based model. We incorporated a mechanism to account for motility bias in experiments with an electric field. Similarly, Ross et al. [77] consider an adhesion mechanism in this model. One could also consider including mechanisms of cell pushing [110], interacting cell populations, or cell cycle dynamics.

When adding complexity to a model, however, care must be taken to ensure model parameters are still (structurally) identifiable. We have shown in Chapter 3, and Chapter 4, that through the addition of certain statistics, different experimental data or different experimental protocols, it is possible to obtain more information about model parameters.

In Chapter 4 and 5, we have established a one-dimensional off-lattice model that assumes cells are in constant contact, so that cell growth and cell-cell interactions are influencing cell spreading. To apply this model more generally to consider epithelial dynamics, it will be important to account for cells moving in two dimensions, and to allow the freedom for cells to move away from each other. Another potential extension of the off-lattice model is to account for cells being in particular phases of a cell cycle, and adjusting their motility or proliferation accordingly [111]. The adjustment of proliferation rates according to lengths can be seen as a means of limiting their proliferation when being compressed by neighbours, but extending these methods to cell age is a logical next step. Another way to account for cell cycle dynamics is to include a transition between a motile state or a proliferating state, potentially resulting in variations in resting lengths or spring constants of individual cells due to variations in their position in the cell cycle.

6.2.1 Final remarks

In this thesis we have considered a wide variety of experimental data, computational models, and analysis methods. Along with standard methodologies, such as ABC rejection and regression adjustment, we have also demonstrated novel approaches, for example through ABC-DC and in the coarse graining in Chapter 5. Throughout this work, we have strived to demonstrate the applicability of inference methods to help investigate the best way to link between experimental data and computational and mathematical models. Our inference investigations into optimal experimental

design have shown without doubt the potential of applying these methods to help guide experimental choices in the future, to help further our understanding of cell spreading processes.

Appendix A

Consistency of results from Chapter 2

In this Appendix, we present additional evidence that demonstrate consistency of the results in Chapter 2 across different parameter regimes, different experimental conditions, and an extended simulation domain.

A.1 Results incorporating a larger spatial domain

Instead of simulating our model on a domain corresponding to that of the experimental image, we can more accurately replicate the experimental domain by initialising cells and simulating in a larger $R = 72 \times C = 96$ grid, and compute statistics only by considering cells in the central portion of that domain, of size $R = 24 \times C = 32$, as performed previously. By considering a snapshot of a larger domain, we are accounting for situations when cells may move into or out of the part of the domain under observation, which may influence the resulting statistics. The domain configurations are demonstrated in Figure A.1 (a), and the difference in information obtained from ABC parameter inference for each statistic when initialising cells over different number of rows in the domain are shown in Figure A.1 (b). The information gain from

each statistic is very much consistent with the information gains shown in Chapter 2.

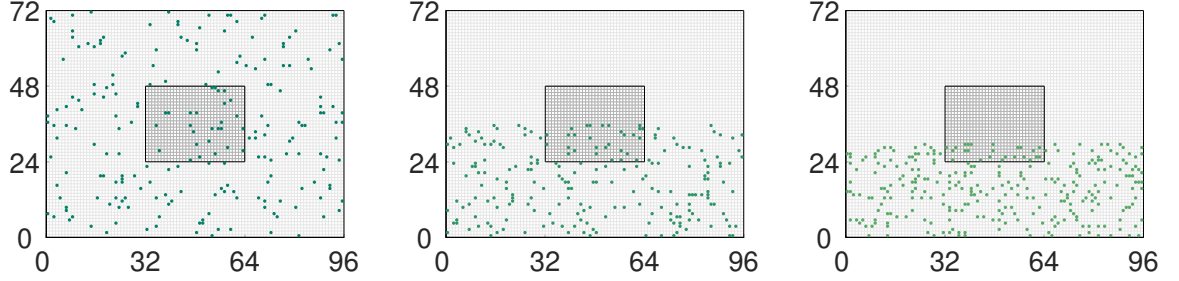
A.2 Results for longer experiments

In the Chapter 2, simulations were performed over 12 h, as most experimental data of cell spreading is observed for at least this amount of time. Here we investigate how our results change if we have data from a longer experiment. In Figure A.2 we consider the change in the information obtained from ABC parameter inference in the case where simulations and experiments are performed over 24 or 36 h, using the same parameter values as in Chapter 2 ($P_m = 0.25$, $P_p = 0.0025$). For each statistic considered, the information gained is either the same or slightly higher when simulating for 24 h rather than 12 h. For some statistics, more information still is gained when simulating for 36 h instead of 12 h (e.g. the connectivity statistics, κ_4 and κ_8), however for others there is a drop in the information gained (e.g. the quadrat statistics, Q_2 and Q_4).

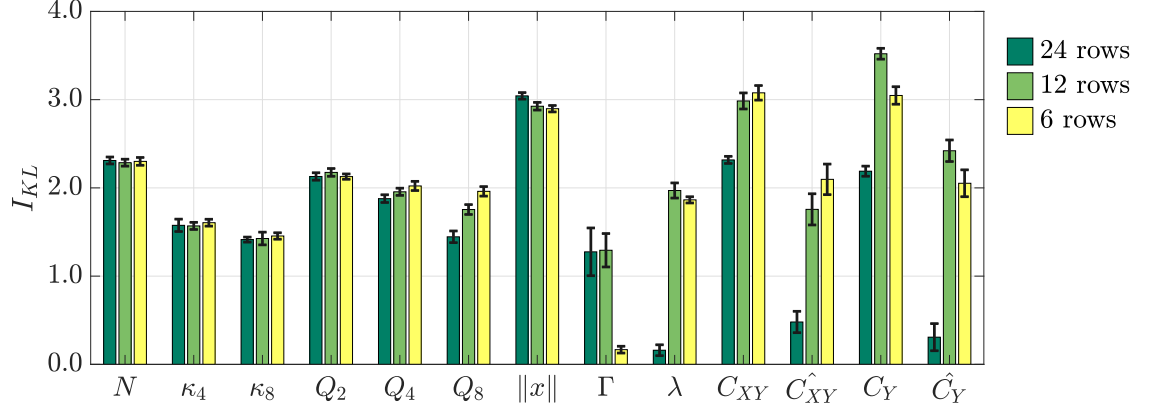
A.3 Results for different parameter regimes

In Chapter 2, synthetic data are generated with $P_m = 0.25$ and $P_p = 0.0025$, giving $Ratio := P_m/P_p = 100$. To test the consistency of our results across a range of cell types, we present additional results here from identical simulations as in Chapter 2 with $Ratio = 10$ or $Ratio = 1$. As the 3T3 fibroblast and MDA MB 231 cell lines we consider in Chapter 2 have similar proliferation rates we keep P_p fixed and vary P_m to $P_m = 0.025$ or $P_m = 0.25$, corresponding to $Ratio = 10$ or $Ratio = 1$. Our results for these new parameter regimes are shown in Supplementary Figures A.3, A.4, A.5 and A.6. The trends in these results are consistent with those in Chapter 2. Note that, as the parameter ratio, $Ratio$, decreases (through decreasing P_m), less information is obtained from the trajectory-based statistics, $\|x\|$ and Γ , as we might

(a) Examples initialising into 24 rows, 12 rows or six rows of the central domain



(b) Initialisation in a larger domain, parameter ratio $Ratio = 100$



(c) Initialisation in the domain used in Chapter 2, parameter ratio $Ratio = 100$

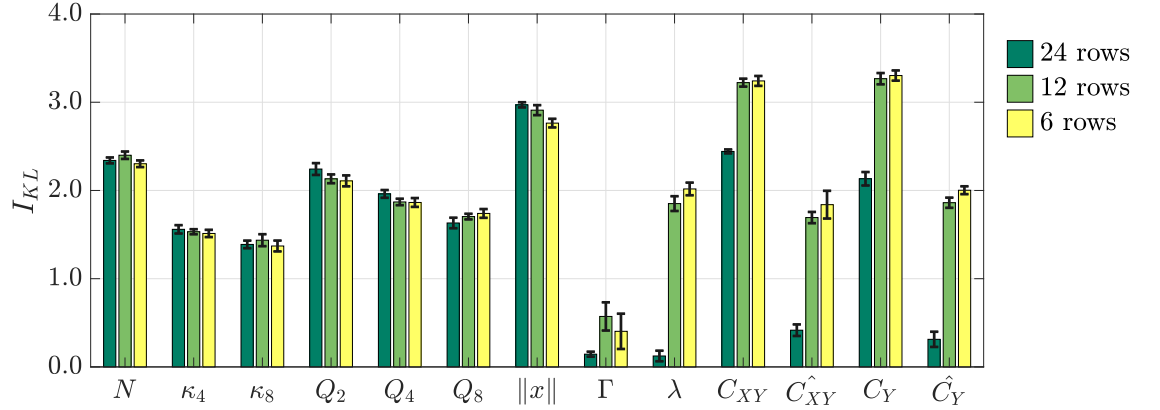


Figure A.1: Information gain in moving from the prior to the posterior for each summary statistic when initialised with a larger domain ($R = 72$, $C = 96$), and computing statistics using only the central portion of the domain. (a) Example experimental domains, corresponding to a growth-to-confluence assays (24 rows) or scratch assays over either half the central domain (12 rows) or one quarter of the central domain (6 rows). (b) Experimental statistics are computed with $Ratio = 100$, corresponding to $P_m = 0.25$, $P_p = 0.0025$. (c) The version of (b) from Chapter 2.

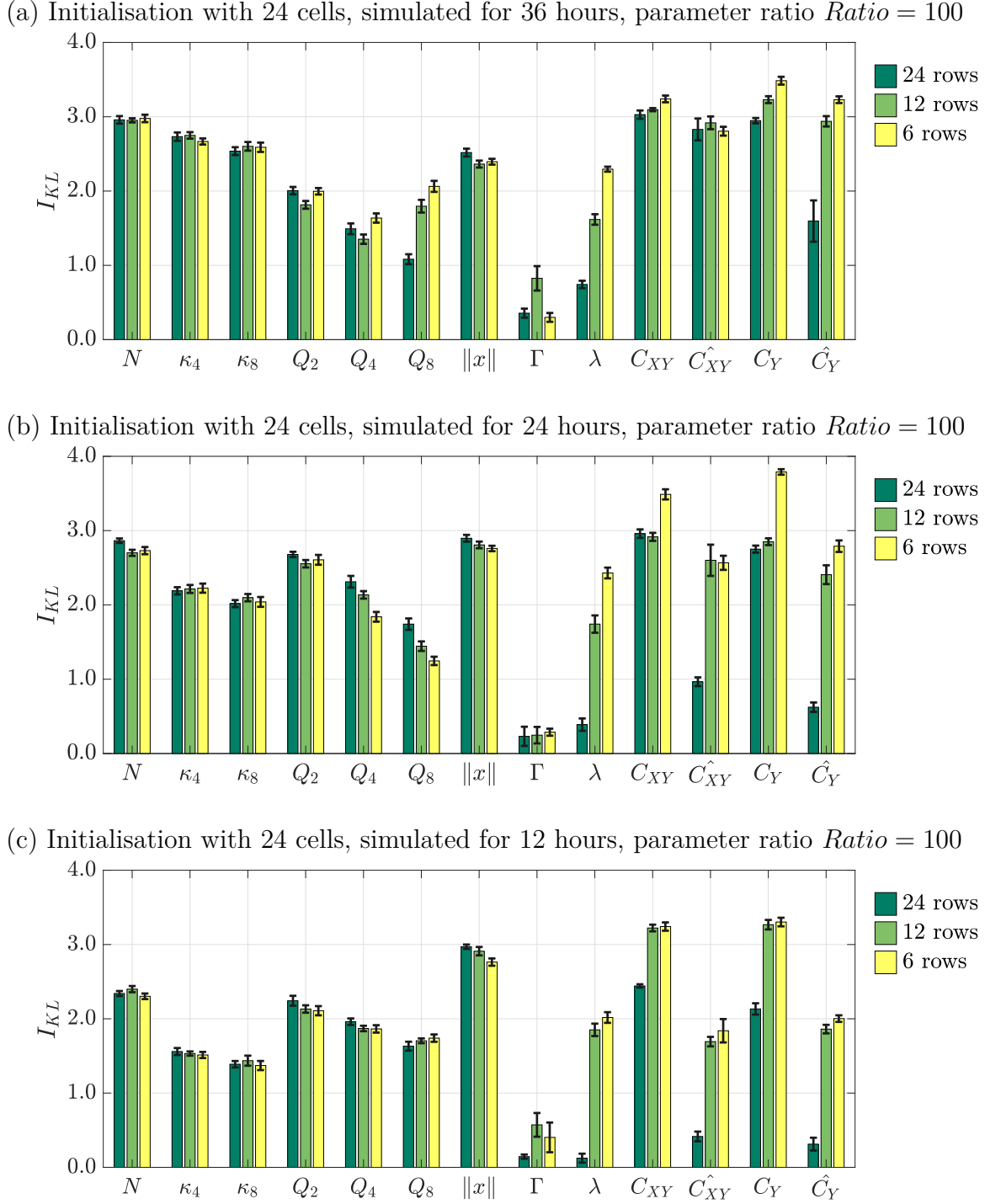


Figure A.2: Information gain in moving from the prior to the posterior for each summary statistic and when initialised with 24 cells and the length of the simulations are varied. Experimental statistics are computed with $Ratio = 100$, corresponding to $P_m = 0.25$, $P_p = 0.0025$. (a) Simulations are run for 36 hours, corresponding to three times the length of the simulations in Chapter 2. (b) Simulations are run for 24 hours. (c) Simulations are run for 12 hours.

anticipate. Further, the information gained from the pairwise correlations statistics, C_{XY} and C_Y , also decreases as *Ratio* is reduced, but only when cell are initialised over six rows.

A.4 Extracting statistics from experimental images

Experimental images are quantified either through the manual or automated extraction of cell positions. In this section these two approaches are outlined. In the manual extraction, the software `imageJ` is used to select cell centres, whereas in the automated extraction, an image segmentation algorithm using the `MATLAB` image processing functions is adopted that automatically finds cell centres.

A.4.1 Manual extraction

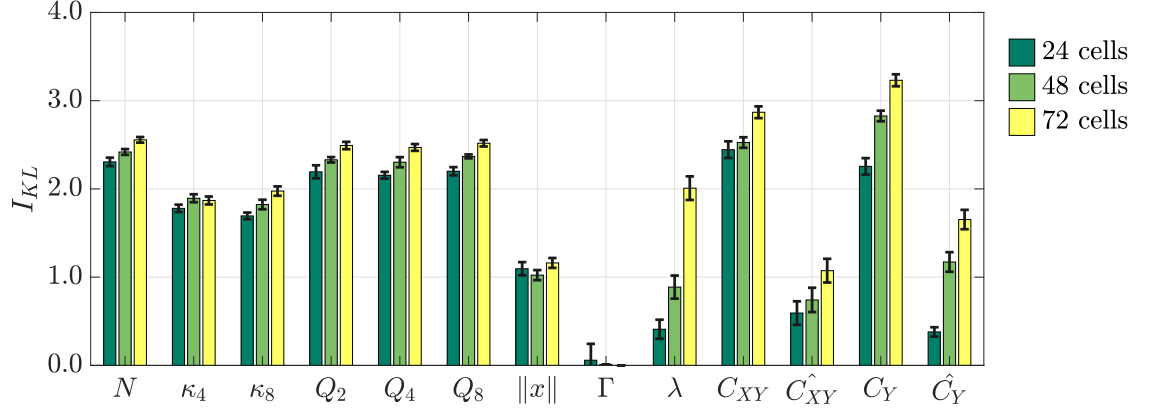
In the manual extraction of cell positions, cell centres are manually tracked using the `MtrackJ` plugin for `imageJ` [70] at the start and end of an experiment, and these cell positions (in pixels) are converted to lattice-site positions. With the lattice-based model used in Chapter 2, experimental images are taken at a resolution of $696\text{px} \times 520\text{px}$, and the lattice has dimensions $R \times C$. The following transformation is used to convert between pixels and lattice site position,

$$X = \left\lceil \frac{C \times x}{696} \right\rceil, \quad (\text{A.1})$$

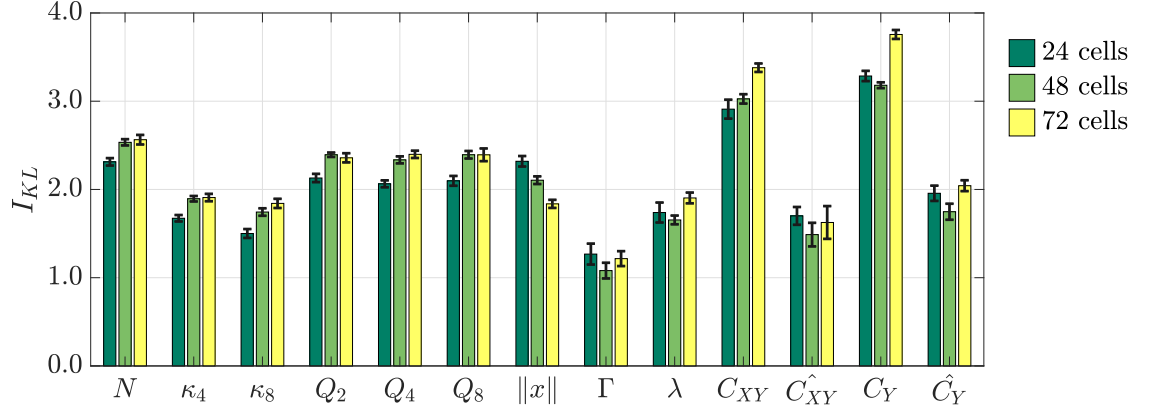
$$Y = \left\lceil \frac{R \times y}{520} \right\rceil, \quad (\text{A.2})$$

where (x, y) are the coordinates of the cell centroids in pixels, (X, Y) are the coordinates on the lattice, and $\lceil \cdot \rceil$ is the ceiling function. If two or more cells are placed in the same lattice site, the excess cells are redistributed into free adjacent lattice sites

(a) Initialisation over six rows, parameter ratio $Ratio = 1$



(b) Initialisation over six rows, parameter ratio $Ratio = 10$



(c) Initialisation over six rows, parameter ratio $Ratio = 100$

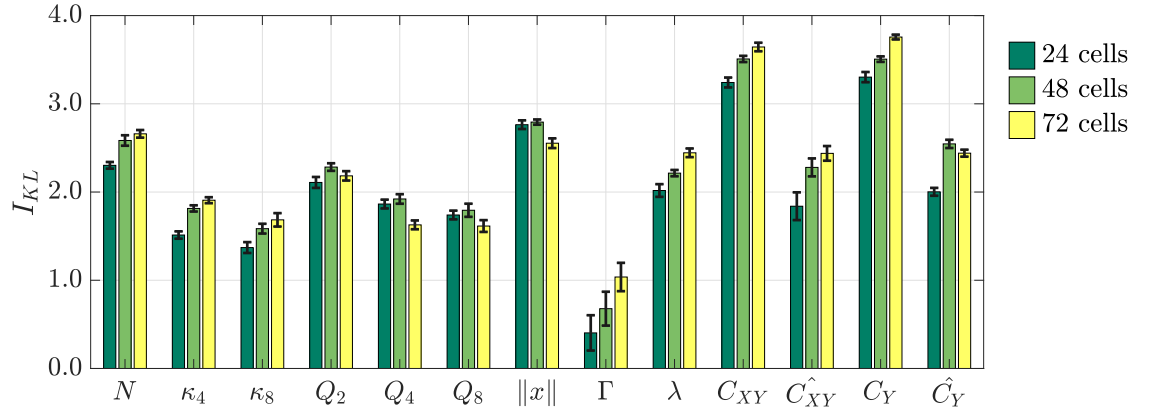
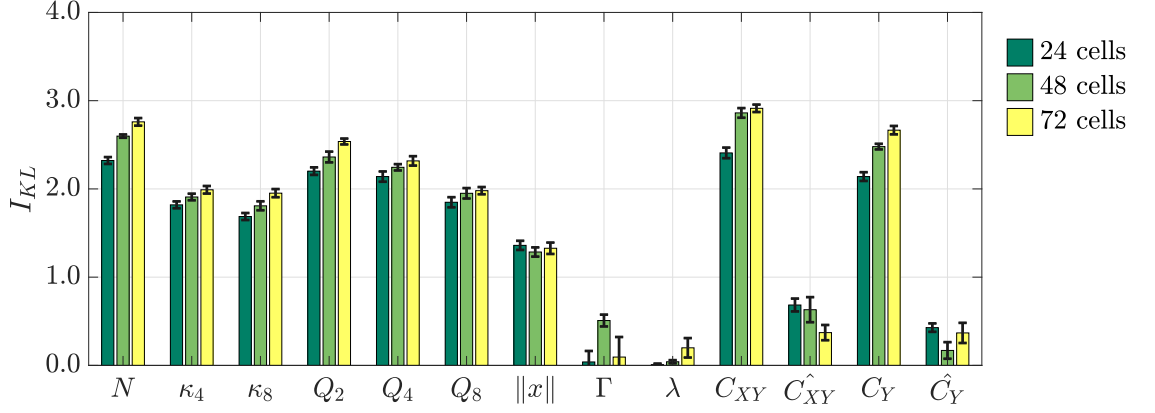
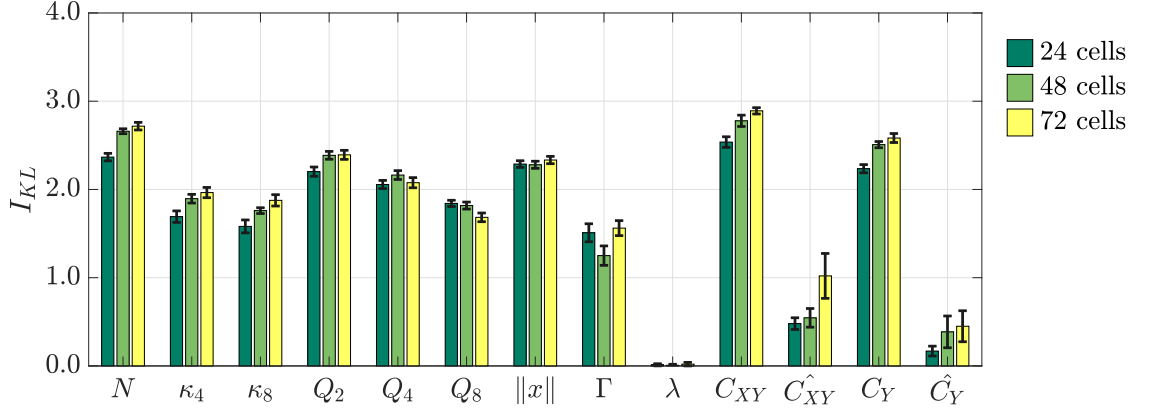


Figure A.3: Information gain in moving from the prior to the posterior for each summary statistic when initialising a scratch assay with 24, 48, or 72 cells. (a) Experimental statistics are computed with $Ratio = 1$ corresponding to $P_m = 0.0025$, $P_p = 0.0025$. (b) Experimental statistics are computed with $Ratio = 10$, corresponding to $P_m = 0.025$, $P_p = 0.0025$. (c) Experimental statistics are computed with $Ratio = 100$, corresponding to $P_m = 0.25$, $P_p = 0.0025$.

(a) Initialisation over 24 rows, parameter ratio $Ratio = 1$



(b) Initialisation over 24 rows, parameter ratio $Ratio = 10$



(c) Initialisation over 24 rows, parameter ratio $Ratio = 100$

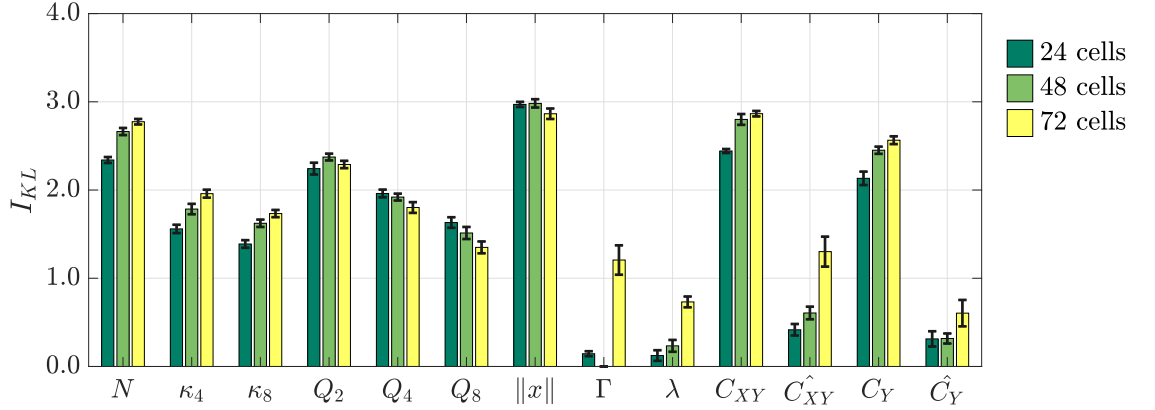
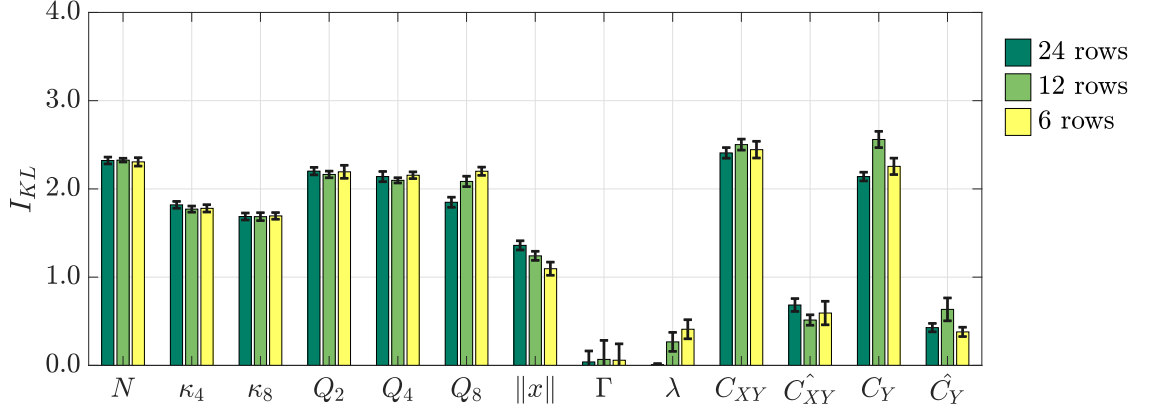
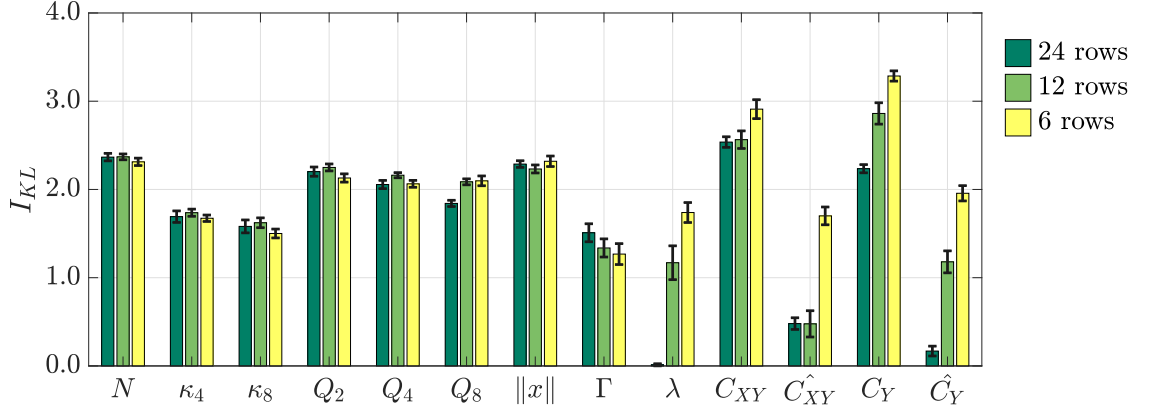


Figure A.4: Information gain in moving from the prior to the posterior for each summary statistic when initialising a growth-to-confluence assay with 24, 48, or 72 cells. (a) Experimental statistics are computed with $Ratio = 1$ corresponding to $P_m = 0.0025$, $P_p = 0.0025$. (b) Experimental statistics are computed with $Ratio = 10$, corresponding to $P_m = 0.025$, $P_p = 0.0025$. (c) Experimental statistics are computed with $Ratio = 100$, corresponding to $P_m = 0.25$, $P_p = 0.0025$.

(a) Initialisation with 24 cells, parameter ratio $Ratio = 1$



(b) Initialisation with 24 cells, parameter ratio $Ratio = 10$



(c) Initialisation with 24 cells, parameter ratio $Ratio = 100$

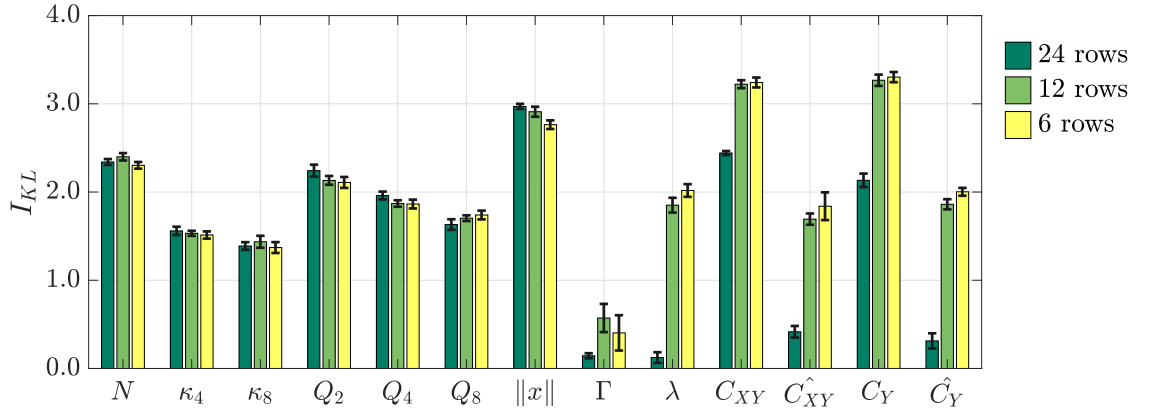
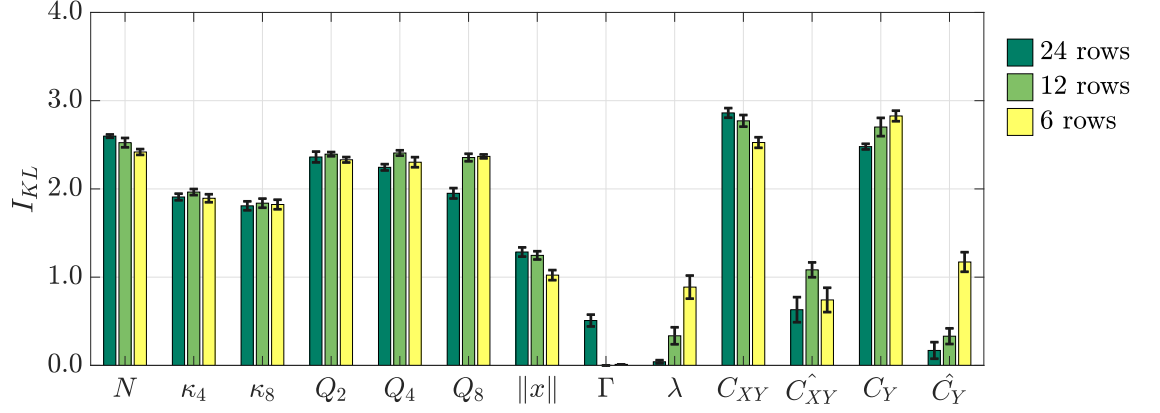
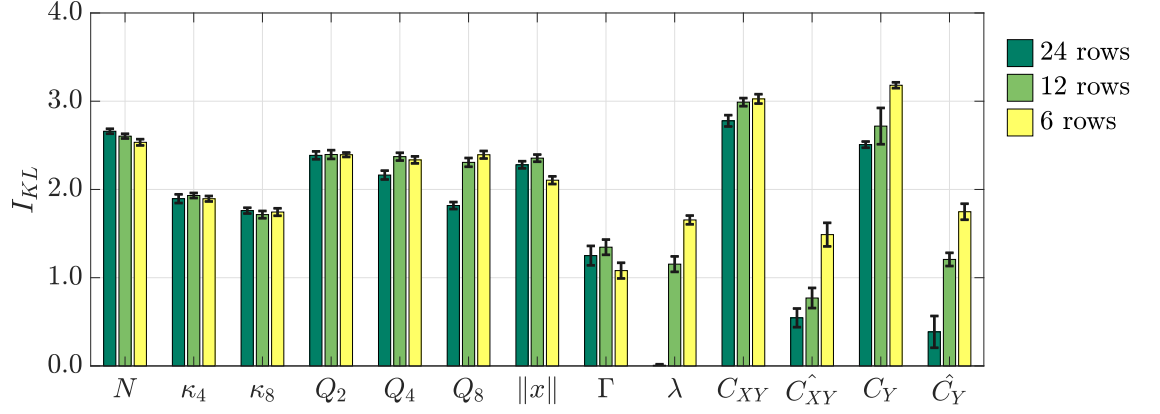


Figure A.5: Information gain in moving from the prior to the posterior for each summary statistic and for each experimental design when initialised with 24 cells. (a) Experimental statistics are computed with $Ratio = 1$ corresponding to $P_m = 0.0025$, $P_p = 0.0025$. (b) Experimental statistics are computed with $Ratio = 10$, corresponding to $P_m = 0.025$, $P_p = 0.0025$. (c) Experimental statistics are computed with $Ratio = 100$, corresponding to $P_m = 0.25$, $P_p = 0.0025$.

(a) Initialisation with 48 cells, parameter ratio $Ratio = 1$



(b) Initialisation with 48 cells, parameter ratio $Ratio = 10$



(c) Initialisation with 48 cells, parameter ratio $Ratio = 100$

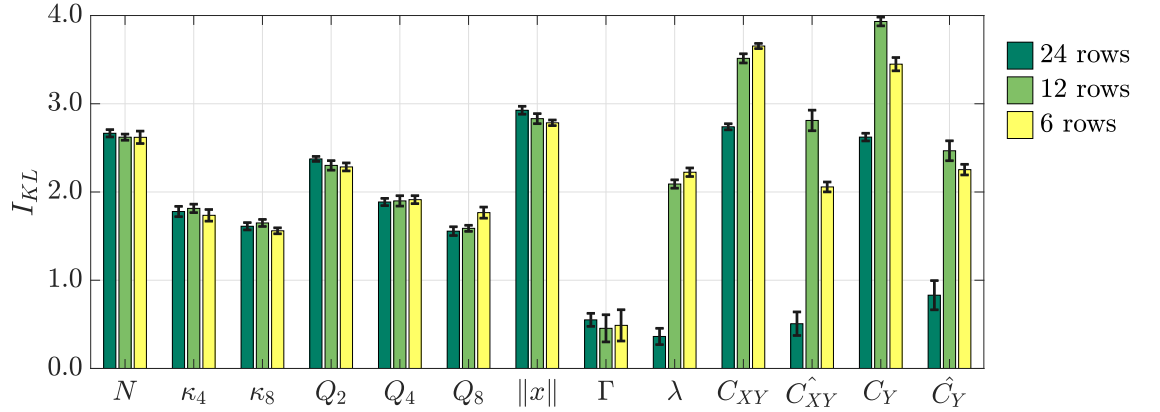


Figure A.6: Information gain in moving from the prior to the posterior for each summary statistic and for each experimental design when initialised with 48 cells. (a) Experimental statistics are computed with $Ratio = 1$ corresponding to $P_m = 0.0025$, $P_p = 0.0025$. (b) Experimental statistics are computed with $Ratio = 10$, corresponding to $P_m = 0.025$, $P_p = 0.0025$. (c) Experimental statistics are computed with $Ratio = 100$, corresponding to $P_m = 0.25$, $P_p = 0.0025$.

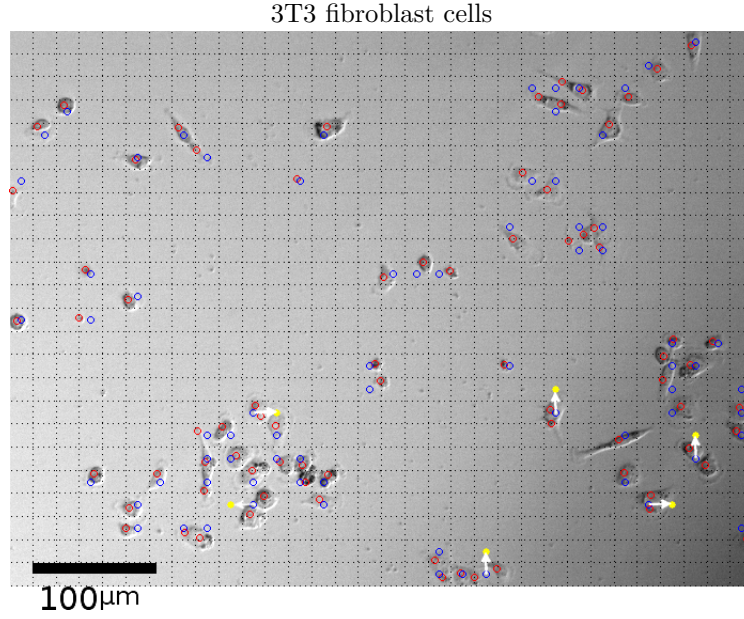


Figure A.7: Cell positions are manually tracked using the **MtrackJ** plugin for **imageJ** [70]. Cell positions are approximated to an $R \times C$ lattice using the simple transformations in equations (A.1) and (A.2). In red are cell centroid positions, in blue is the lattice approximation of cell centre positions. Any cells that cannot be placed into their nearest lattice site are randomly placed into the next closest available lattice site.

until no two cells occupy the same site. This process is illustrated in Figure A.7.

In order to extract cell trajectories, cells are manually tracked using the **MtrackJ** plugin for **imageJ** for every frame of an experiment, and these cell positions (in pixels) are converted to lattice-site positions using the equations (A.1) and (A.2).

A.4.2 Automated extraction

To quantify experimental images automatically, a cell segmentation algorithm is developed that identifies “objects” in experimental images. These “objects” connected over multiple frames using a cell tracking algorithm allows cell tracks to be obtained. The segmentation approach is shown schematically in Figure A.8, and the full procedure is reasoned below.

1. Perform a 2-level Haar wavelet decomposition with the **MATLAB** function **wdencmp**.

This leads to a smoother image, which reduces the amount of small objects we

expect in step 5.

2. Erode the image from step 1 with a disk structural element of radius 20 px, then reconstruct in order to remove bright spots around cells.
3. Dilate the image from step 2 with a disk structural element of radius 20 px, then reconstruct the image from the images from step 2 and 3. This image is the background gradient without any cells.
4. Take the absolute difference of the image from step 2 and the background from step 3, and remove any pixels of intensity value 5 or lower, to remove noise.
5. Convert the image into a binary image and remove further noise by removing connected regions containing 30 pixels or less.
6. Find the distance between each pixel in a connected region and the background. Compute the regional maxima of this distance to locate cell centres.
7. Morphologically close the image using a disk structural element of radius 3 px, and then dilate the image using a disk structural element of radius 1 px.
8. Watershed the image from step 7 based on the centres found in the image from step 6.

These steps were motivated from previous work in [62, 70, 74] along with trial and error using the experimental images.

Cell tracking is attempted in an intuitive way, by connecting cells between frames which are within a radius, rad , of each another. Although more complicated methods exist, (for example PHD filtering [78, 108], or transformation into an assignment problem [15, 62]), this method is intuitive and quick to implement. The following procedure is applied, and is illustrated in Figure A.9.

1. Find all the object locations in two consecutive frames, i and $i + 1$;

2. For each object in frame $i + 1$, list the objects in frame i that are within radius rad ;
3. For each object in frame i , list the objects in frame $i + 1$ that are within radius rad ;
4. If any two objects in consecutive frames have only each other in their lists, create a track between these two objects.

To find the optimal search radius, rad , four independent (randomly chosen) sets of experimental images are considered. In each case, rad is varied between one and fifteen pixels (inclusive), to find the value of rad resulting in the largest number of tracks with a length of 36 frames or longer. This method finds $rad = 7$ to be optimal for all four data sets. The number of tracks and length of tracks against rad is shown for one set of experimental images in Figure A.10.

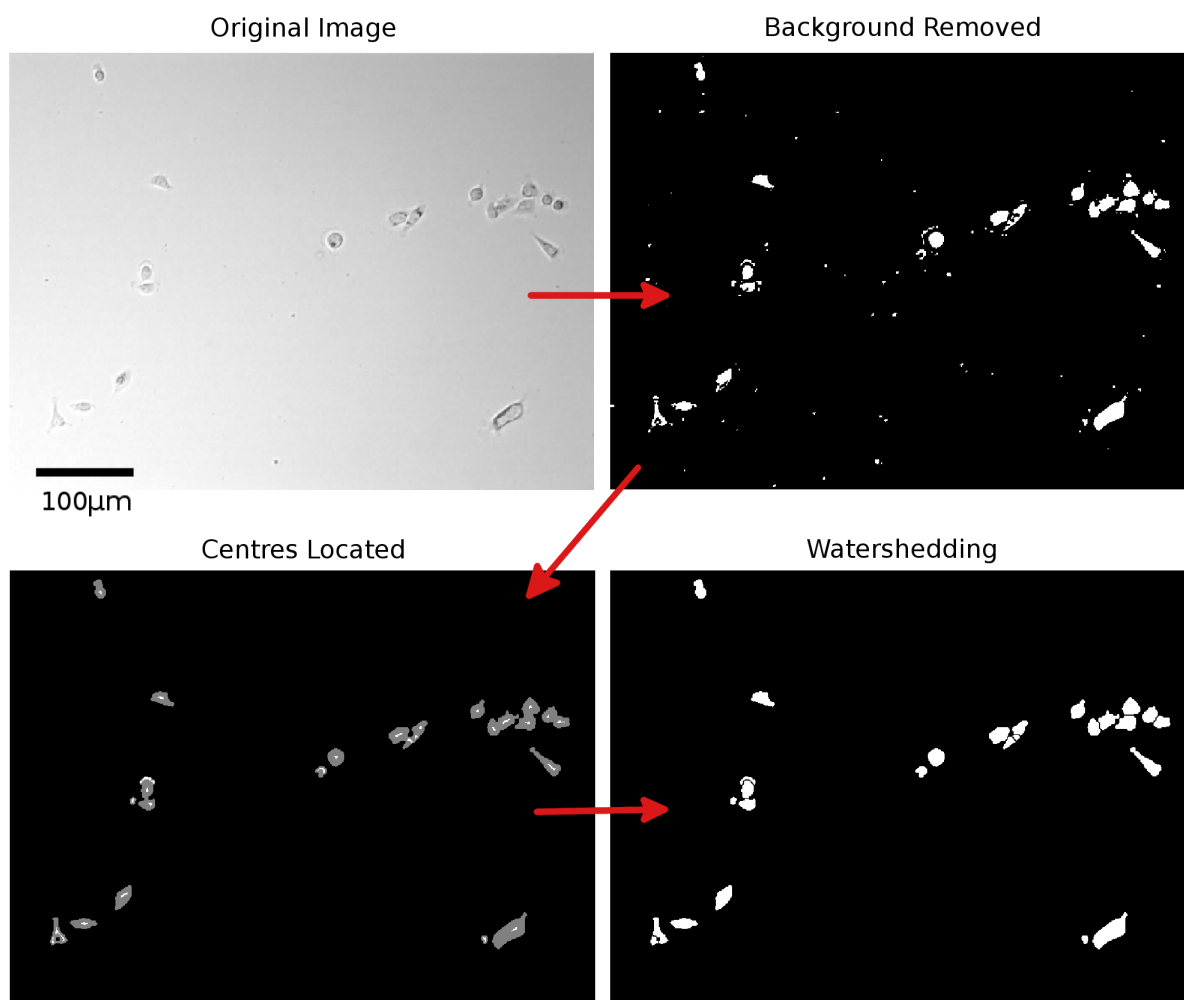


Figure A.8: Overview of the cell segmentation procedure used on MDA MB 231 cells to identify centroids of cells. The original image is shown in the top-left, which is denoised, eroded and dilated using disk structural elements of radius 20 px to obtain the background image. The background is subtracted from the smoothed original image and converted into binary to obtain the image in the top-right. The small objects in the top-right image are removed, and the centres of objects remaining are obtained by finding the regional maxima of the distance between each pixel and the nearest empty pixel, to obtain the image in the bottom-left. A closure and dilation is applied to the image in the bottom-left with disk structural elements, then the resulting image is watershedded based on the cell centres found previously, to obtain the segmented image in the bottom-right.

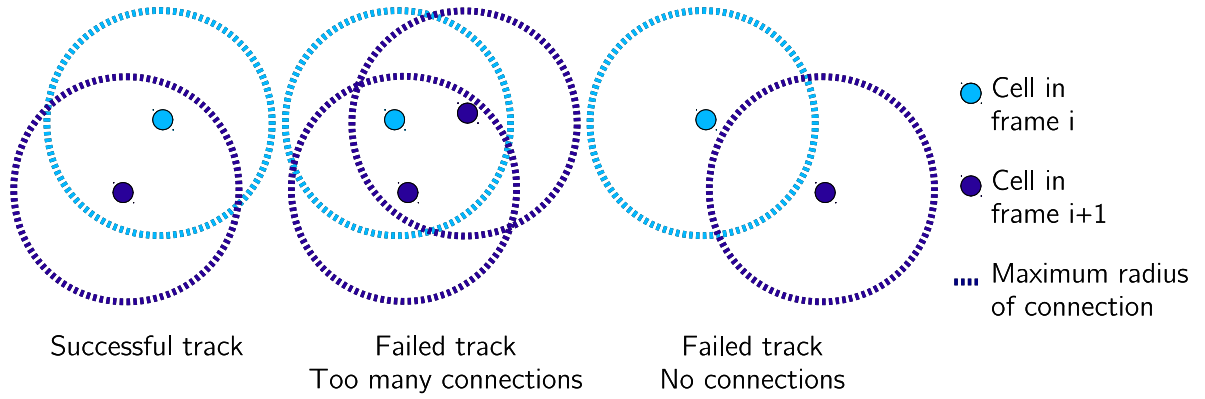


Figure A.9: Overview of the cell tracking procedure used on MDA MB 231 cells. Cells centres in frame i are shown in light blue, whereas those in frame $i + 1$ are shown in dark blue. Cells in consecutive frames continue a track if they are within each others radius of connection, however if two or more possible tracks could occur, no connection will be created.

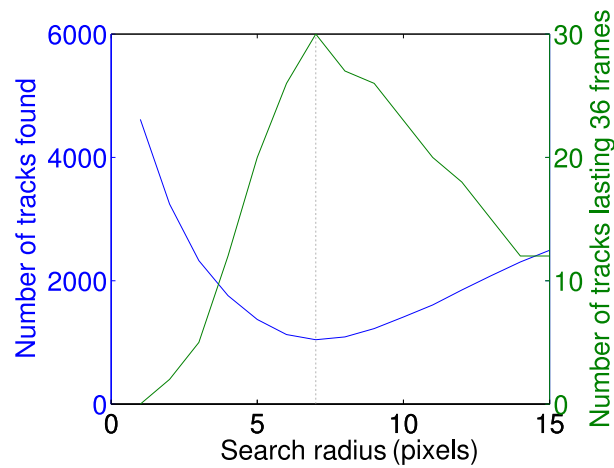


Figure A.10: Finding the search radius which results in the largest number of long (36 frames or longer) tracks. Blue: total number of tracks found; green: total number of long tracks found for a given search radius, rad , in pixels. The minimum total number of tracks and maximum total number of long tracks occurs at $rad = 7$ pixels.

Appendix B

Additional methods and results for Chapter 3

In this Appendix, we present the algorithms used to emulate initial experimental domains, along with additional results from the coarse graining in Chapter 3. We consider how regression adjustment affects the posterior distributions for synthetic or experimental data, and how different experimental domains or data generative from different parameter regimes affects inference investigations.

B.1 Replication of initial domains

In order to create each shape; square, circle or triangle; in **MATLAB**, first create a matrix of zeroes, **Mtemp**, of size $(2r + 1) \times (2r + 1)$, where r is the radius of the shape, and occupy sites of the matrix with the code below¹.

```
1 %% Compute circle :  
2 for x = 0:2*r  
3     for y = 0:2*r
```

¹Initialisation as a square is a straightforward usage of the **ones** function, and therefore not shown.


```

4         if ((x-r)^2+(y-r)^2) <= r^2
5             Mtemp(x+1, y+1) = 1;
6         end
7     end
8 end

1 %% Compute triangle:
2 for x = 0:r
3     for y = 0:2*r
4         if y <= sqrt(3)*x
5             Mtemp(x+1, y+1) = 1;
6         end
7     end
8 end

9 for x = r:2*r
10    for y = 0:2*r
11        if y <= sqrt(3)*(r-x)
12            Mtemp(x+1, y+1) = 1;
13        end
14    end
15 end

```

These small matrices are inserted into the centre of the $R \times C$ lattice, as demonstrated in Figure 3.3 in Chapter 3. To emulate different degrees of initial confluence, we randomly choose occupied sites to retain in the initial lattice (using the `rand` and `find` MATLAB functions).

B.2 Regression adjustment

In this section we investigate how regression adjustment affects the posterior distributions resulting from the ABC-rejection inference procedure in Chapter 3. Regression adjustment accounts for non-zero distances between simulation and experimental statistics by fitting a regression line of best fit through the statistics, and adjusting parameters to fall on this line of best fit.

B.2.1 Inference without regression adjustment on synthetic data

In Chapter 3 we used regression adjustment to account for non-zero distances between simulated statistics and experimental statistics (that are typical when using ABC-rejection due to the approximate nature of the procedure). We expect that if regression adjustment is not used, the information gain in moving from prior to posterior, particularly when considering more statistics, is lower. To demonstrate this, no regression adjustment is used to generate the plots in Figure B.1, and these can be directly compared with plots in Figure 3.4. We observe a marked decrease in the one-dimensional information gain for the pairwise correlation statistics, and for combinations of statistics.

B.2.2 Inference with regression adjustment on experimental data

In Figure B.2 regression adjustment is used on the accepted parameters from the inference procedure applied to experimental data. These plots are analogous to those in Figure 3.9, except parameter values are shifted via regression adjustment to account for non-zero distances between simulation statistics and experimental statistics.

When inferring model parameters with regression adjustment on experimental

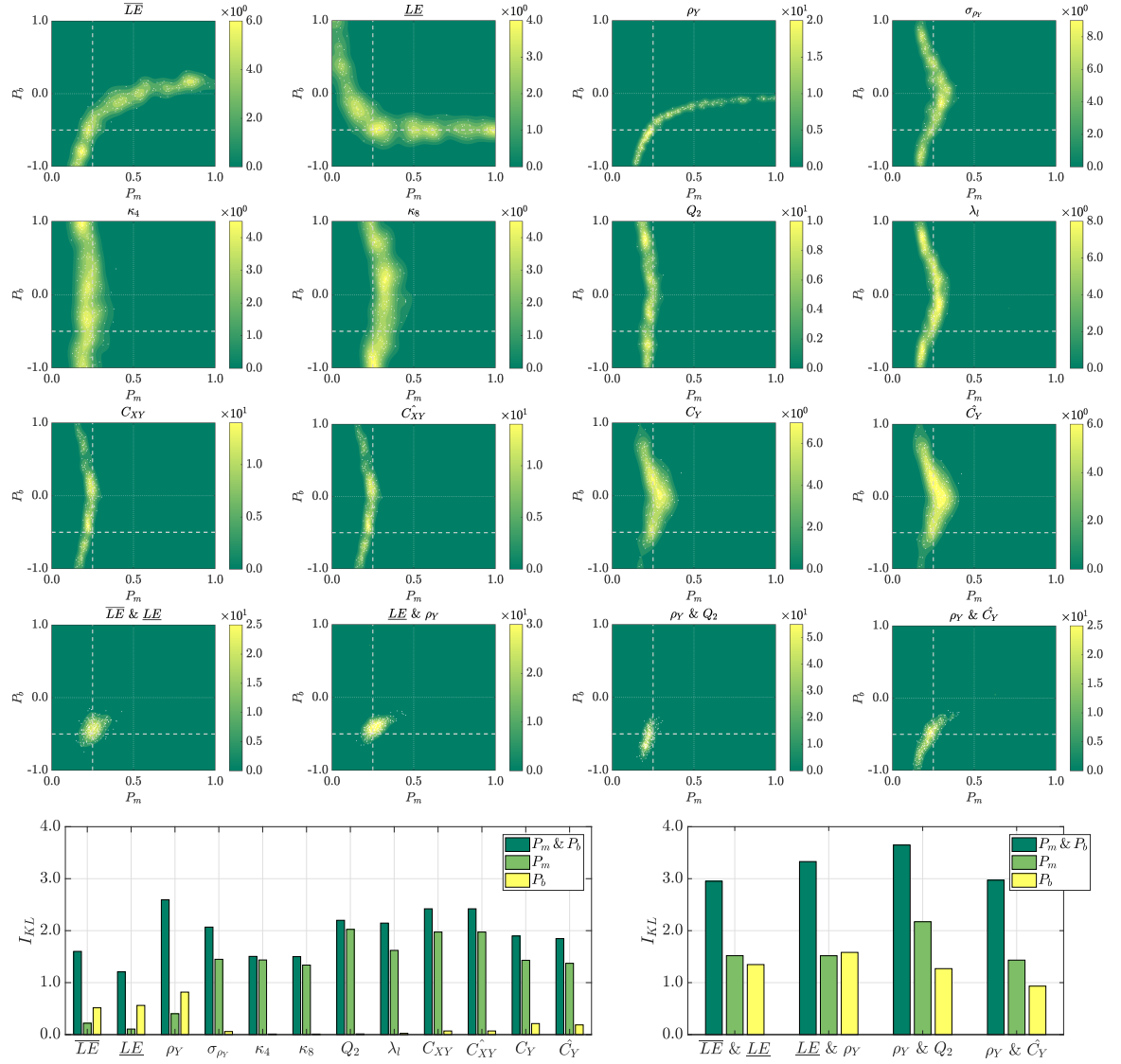


Figure B.1: The details of this figure are identical to Figure 3.4, except here regression adjustment is not used to adjust parameters. Comparison to Figure 3.4 clearly shows a decrease in information gain without the use of regression adjustment.

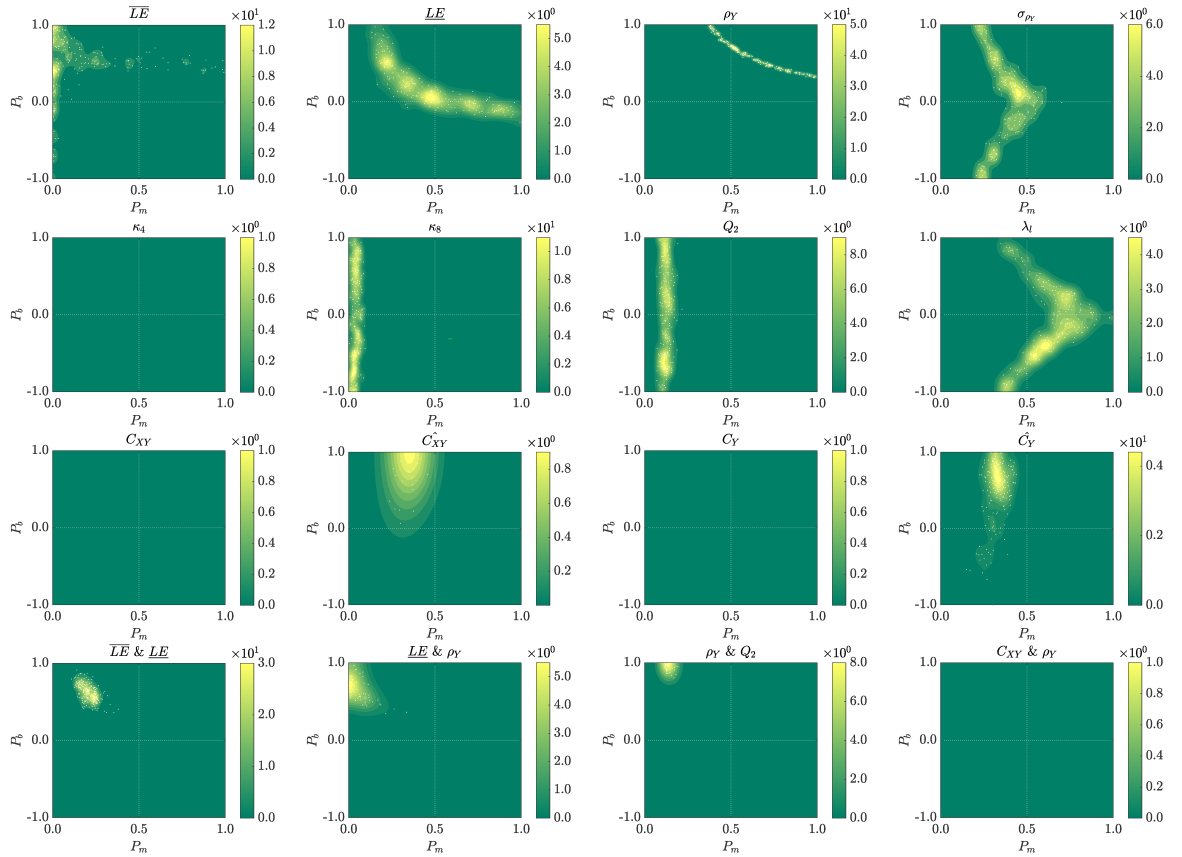


Figure B.2: Displayed in the first three rows are the posterior distributions generated by the samples from the prior distributions that resulted in the lowest distance between simulated and experimental statistics. The posterior distributions correspond to the statistic titled, where the statistic shorthand is used (see Table 3.1). The fourth row posterior distributions correspond to those generated when using a combination of statistics. In each posterior distribution, larger posterior densities are more yellow (according to the colourbar), the parameters used to generate the posterior distribution are shown as white dots. The initial domain is taken from experiments before the introduction of an upwards bias, and the final domain is taken from the experiment after two hours of the bias, these domains are shown in Figure 3.1. Regression adjustment is used, so some parameters used to approximate a posterior distribution fall outside the prior domain.

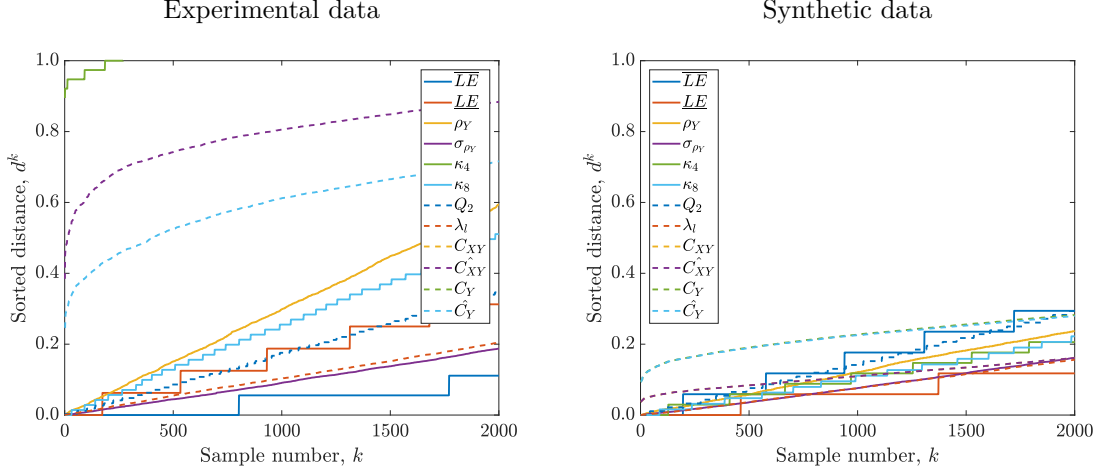


Figure B.3: The resulting distances (the smallest 10% are sorted for each statistic) resulting from experimental data inference investigations (left, posteriors shown in Figure 3.9) or the synthetic data inference investigations (right, posteriors shown in Figure 3.4). When using experimental data, even the smallest distances for C_{XY} , C_Y and κ_4 fall outside the axis limit ($d^k < 1$), whereas when using synthetic data this is not the case (the line for C_{XY} lies underneath $\hat{C}_{XY}$).

data some statistics result in parameters that fall outside the range of the prior distributions, $P_m \in (0, 1)$ or $P_b \in (-1, 1)$. To understand why this is the case, in Figure B.3 we plot the distances, d_j^k , between simulation statistics and experimental statistics, for each sample, k , and each statistic, j . For each statistic distances are sorted from smallest to largest, and the smallest 2000 (10%) distances are displayed (note that posterior distributions are generated using the smallest 200 distances) for both the experimental data or the synthetic data from the investigations in Section 3.3.1. This figure demonstrates that distances resulting from the experimental data are much larger than those when using synthetic data, suggesting that the model and statistics considered do not fully capturing the behaviour of cells in experiments. Using regression adjustment on parameters inferred from experimental data therefore alters parameters considerably more than in the case of synthetic data, so regression adjustment is not recommended when using experimental data.

For reference, when using regression adjustment on parameters inferred using experimental data, the estimates of P_m and P_b from the posterior distribution generated

with statistic combination ρ_Y & Q_2 are:

- $P_m = 0.0976 \pm 0.0341$ and $P_b = 1.1032 \pm 0.1014$.

B.3 Different experimental domains or parameter regimes

In this section, different parameter regimes for synthetic data inference investigations are considered using the model in Chapter 3, in order to demonstrate consistency of results. We consider whether the bias parameter is still identifiable in experiments with different degrees of bias, or with the inclusion of a proliferation mechanism akin to the model in Chapter 2.

B.3.1 Are parameters inferable with less or no bias?

We investigate whether synthetic data generated with different values of P_b (different from the $P_b = 0.5$ used in Chapter 3) also result in inferable parameters. In Figure B.4 we show the posteriors generated from the combinations of statistics used in the main text for experimental $P_b = 0.00, 0.25, 0.50$ and 0.75 .

B.3.2 If proliferation is included how do results change?

We investigate how results change if we introduce a proliferation mechanism in the model. Define P_p as the probability of a proliferation event occurring in time-step τ . The proliferation mechanism is the same as in Chapter 2.

Note that the corneal epithelium cells considered in Chapter 3 only proliferate a small amount over the seven hour period of the experiment. Estimating P_p for these cells can be attempted using the number of cells statistic, running ABC-rejection with the model from Chapter 2, using data from the first and last frame of the

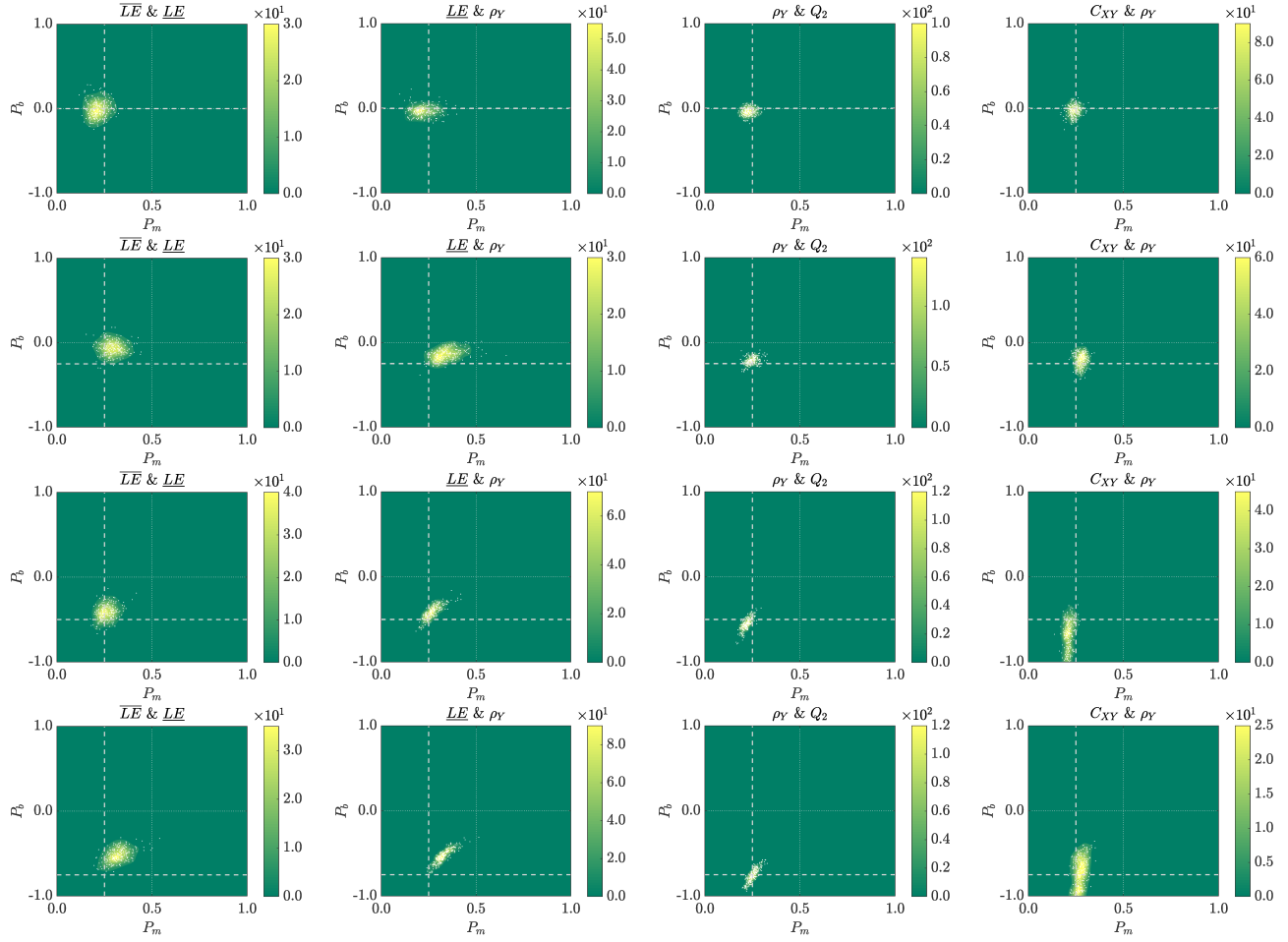


Figure B.4: Posterior distributions for the most informative two-statistic combinations established in Chapter 3. The bias parameter used to generate synthetic experimental data are varied, $P_b = 0.0$ (first row), 0.25 (second row), 0.5 (third row), or 0.75 (bottom row).

7 hour control experiment (shown in Figure 3.1), and specifying P_m as the mean value determined from the inference in section 3.3.4. Averaging accepted P_p values results in $P_p = 0.0025$ (results not shown), therefore a prior distribution for P_p of $\pi(P_p) = \mathcal{U}(0, 0.01)$ is suitable.

Now consider inference of the three model parameters simultaneously, P_m , P_b and P_p , using the statistics outlined in Chapter 3 along with the number of cells statistic, N . Note that the prior for P_p is the same as used in Chapter 2 for MDA MB 231 cells or 231 fibroblast cells, and in that case, in order to satisfy $P_m + P_p < 1$, the prior for P_m is restricted to $\pi(P_m) = \mathcal{U}(0, 0.99)$.

Synthetic data is generated using $P_m = 0.25$, $P_b = -0.5$ and $P_p = 0.0025$ and results in the two-dimensional posterior distributions and information gains shown in Figure B.5. We display only a small number of statistic combinations considered; N and C_{XY} , C_{XY} and Q_2 , and $\underline{LE}$ and ρ_Y . When proliferation is introduced we observe that certain summary statistic combinations still provide enough information to infer the three model parameters. For example, N and C_{XY} can be used to infer P_m and P_p ; while $\underline{LE}$ and ρ_Y can be used to infer P_m and P_b . Information gains are comparable with those in Figure B.1.

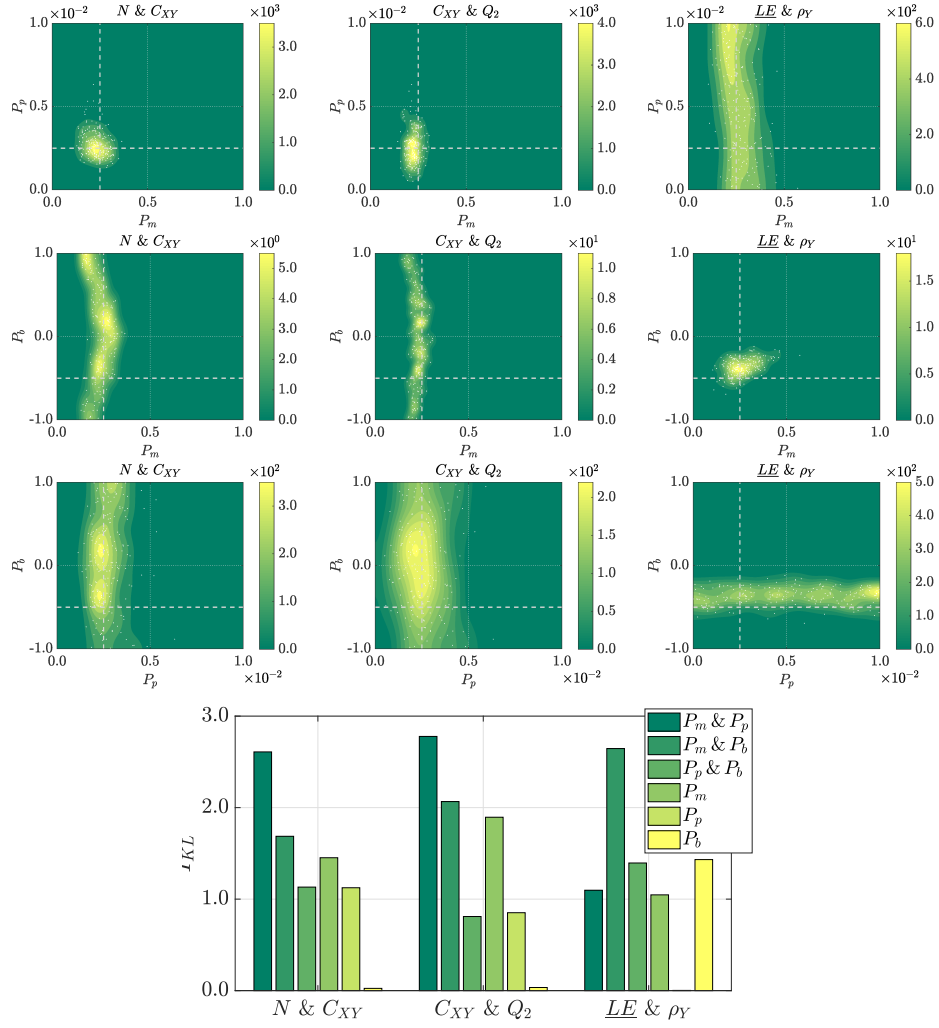


Figure B.5: The posterior distributions and corresponding marginal two-dimensional and one-dimensional information gains from three chosen combinations of statistics are shown. The combinations are; N and C_{XY} (left), C_{XY} and Q_2 (centre), or $\underline{LE}$ and ρ_Y (right). In each posterior distribution, larger posterior densities are more yellow (according to the colourbar), the parameters used to generate the posterior distribution are shown as white dots. Grey dashed lines represent the parameter values used to generate the synthetic experimental data ($P_m = 0.25$, $P_p = 0.0025$, $P_b = -0.5$). Note that these posterior distributions and information gains are generated without using regression adjustment.

Appendix C

Extraction of experimental data and further results from Chapter 4

In this Appendix, we present additional methods and results from the inference of the off-lattice model in Chapter 4. The method used to extract experimental data is presented, along with other inference investigations considering combinations of statistics or a different proliferation law.

C.1 Extracting data from experiments or simulations

In order to compare the model introduced in Chapter 4 with experimental data, we outline a means of comparison between node positions and experimental cell densities. Algorithms are developed that convert from simulated node positions to cell densities, and from experimental cell densities to node positions, allowing us to use experimental data to establish initial conditions in the model.

C.1.1 Conversion from positions to density

Firstly consider computing statistics from the model. It is simple to extract the number of cells at the final time, $N = N(T)$, as this is continually updated when solving. The leading edge, LE , is simply the position of the $N(T)^{\text{th}}$ node, or $LE = r_{N(T)}$. To convert cell positions into cell density, we use the formula established in Murray et al. [57],

$$q(r, t) = \frac{\partial i}{\partial r}, \quad (\text{C.1})$$

where $q(r, t)$ is the cell density at position r and time t , and i and r are continuous versions of the discrete positions r_i and discrete indices i . (See Chapter 5 or [57] for clarification). Using equation (C.1), estimating the cell density, q , at the node positions, $r_i(t)$, is done using,

$$\begin{aligned} q(r_0, t) &= \frac{1}{r_1(t)}, \\ q(r_i, t) &= \frac{2}{r_{i+1} - r_{i-1}} \text{ for } i = 1, \dots, N(t) - 1, \\ q(r_{N(t)}, t) &= \frac{1}{r_{N(t)} - r_{N(t)-1}}. \end{aligned} \quad (\text{C.2})$$

The value of q at any intermediary position between nodes can be computed by interpolation.

C.1.2 Conversion from density to positions

As the IBM simulates positions of cells, and experimental data are recordings of the cell density, we require a method of converting from density data to position data. By simply rearranging equation (C.2), the following equations for node positions are

established,

$$\begin{aligned}
r_1(t) &= \frac{1}{q(r_0, t)}, \\
r_{i+1}(t) &= \frac{2}{q(r_i, t)} - r_{i-1}(t) \text{ for } i = 1, \dots, N(t) - 1, \\
r_{N(t)}(t) &= \frac{1}{q(r_{N(t)}, t)} + r_{N(t)-1}(t).
\end{aligned} \tag{C.3}$$

C.1.3 Extracting statistics from experimental data

In order to satisfy the model assumption that cells are in contact at their edges, we perform the following steps; for each of the average cell density data sets considered, first compute node positions using the algorithm in Section C.1.2, then restrict attention to positions where there is a large cell density. As the experimental data has a two separate expanding cell populations (from the left and right), consider only the portion of the domain between 0 and LE , where LE denotes the position when the distance between node points is larger than two cell widths. For the data considered, Jin et al. [36] computes the expected carrying capacity as $q^{max} = 1.7 \times 10^{-3}$ cells/ μm^2 , which can be converted to an expected cell diameter at confluence of $a = 1/\sqrt{1.7 \times 10^{-3}} = 24.25\mu\text{m}$.

The IBM is simply a single line of cells, so the density of *cells*/ μm^2 must be converted to *cells*/ μm as follows:

1. convert densities to one-dimensional (per column) densities by multiplying by $1430\mu\text{m}$
2. divide the resulting one-dimensional cell density with the expected number of cells that would fit in $1430\mu\text{m}$, computed as 59.¹

These methods are shown in Figure 4.6.

¹Using the expected cell diameter at confluence, calculate $1430/24.25 \approx 59$ as the expected number of cell-width rows at confluence.

C.2 Other inference experiments

In this section we describe the results of other experimental design investigations.

C.2.1 Combining statistics

In much the same way as experiments are combined in the investigation from Section 4.3.1.5, statistics can be combined using the distance formula in Section 4.2.4.2. The hope is that combinations of statistics may provide more information gain than from considering a single statistic. However, when considering more than two time points, note that the increase in data may result in a decrease in information gain, as we are adding ‘non-useful’ data and trying to match all statistics with equal weight (increasing dimensionality).

Results are shown in Figure C.1. Through comparison to Figure 4.7, it is observed that combinations of statistics, while able to pick up more information about the spring length and proliferation parameters a and β simultaneously (in the two-dimensional posteriors), the combination of statistics provide only slightly more information about the spring length a than the cell density statistic q provides on its own. Further, less information is obtained about the proliferation parameter β with a statistic combination than the number of cells statistic N , provides on its own. In summary, it is not necessary to combine statistics in order to maximise information gain about each parameter.

C.2.2 Results using a hyperbolic proliferation law

Here we consider the effect on the information gain of the model parameters if we instead used a hyperbolic proliferation law instead of a constant proliferation law. Proliferation laws are outlined in Table 5.4, and a hyperbolic proliferation law, though slightly harder to implement, provides a more realistic representation of cells, which

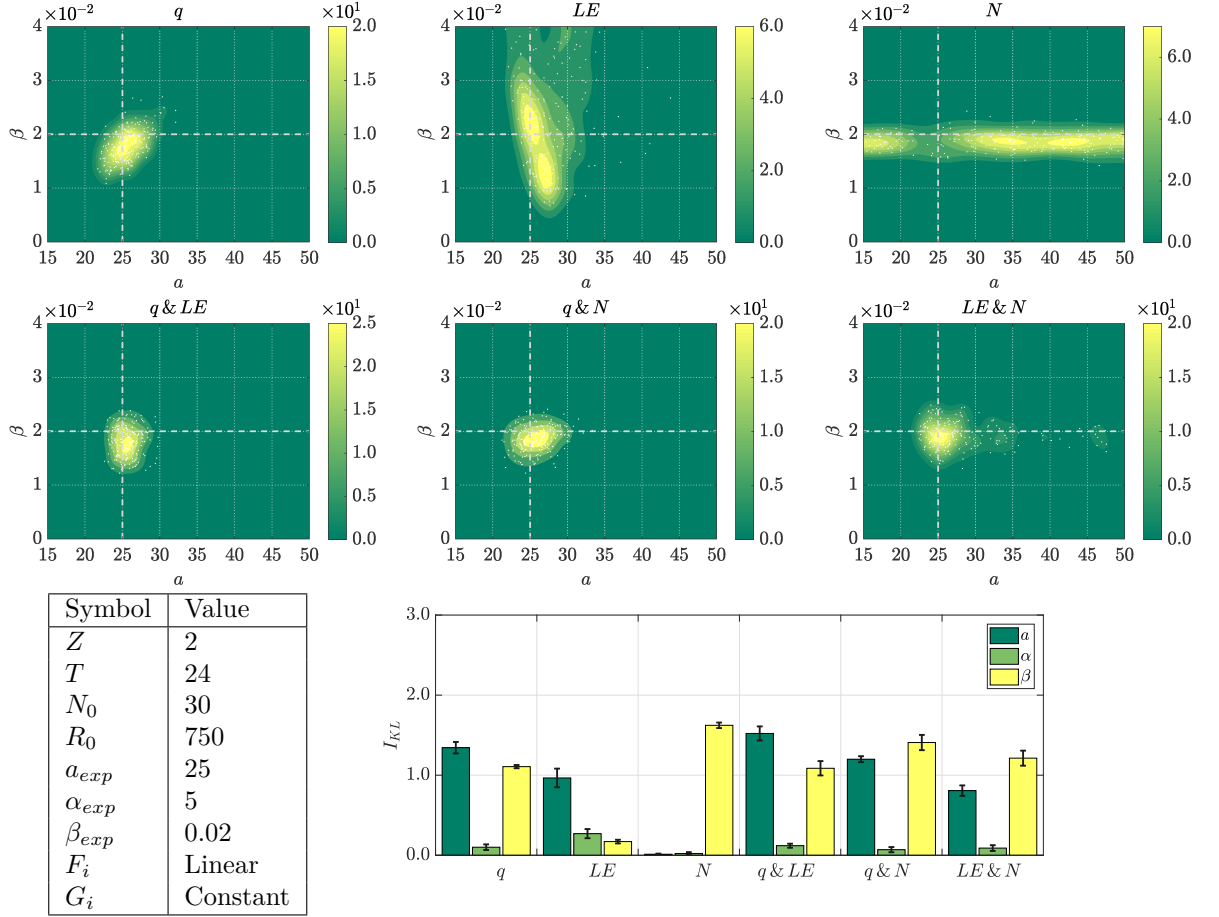


Figure C.1: We perform ABC-SMC with synthetic data computed using the experimental parameters shown. Prior distributions used and other inference choices are outlined in Table 4.1. The experiments performed here use the same inference choices as in Figure 4.7, but we combine the statistics used, according to the distance computations outlined in Section 4.2.4.2. The marginal two-dimensional posterior distributions shown in the top row show that each combination of statistics is able to provide information of parameters a and β simultaneously. The information gain from the marginal one-dimensional posteriors for each parameter and each statistic combination are shown in the bottom row. We have included the information gains from the one-statistic inference in Figure 4.7 for ease of comparison.

generally reach a particular size before proliferating. A hyperbolic law limits the amount of proliferation in the domain until individual cells reach a particular size, and so in this case the proliferation parameter β is more closely tied to the motility parameter, α and the resting spring length, a .

In Figure C.2, we can see that posteriors for the number of cells statistic, N , have a strong correlation with a and α , a relationship that was not previously seen in the case of constant proliferation. We can also note that the information gain for the proliferation parameter is much lower here than in the constant proliferation case, therefore by adopting a more realistic proliferation law, we obtain less precise estimates for the proliferation parameter β .

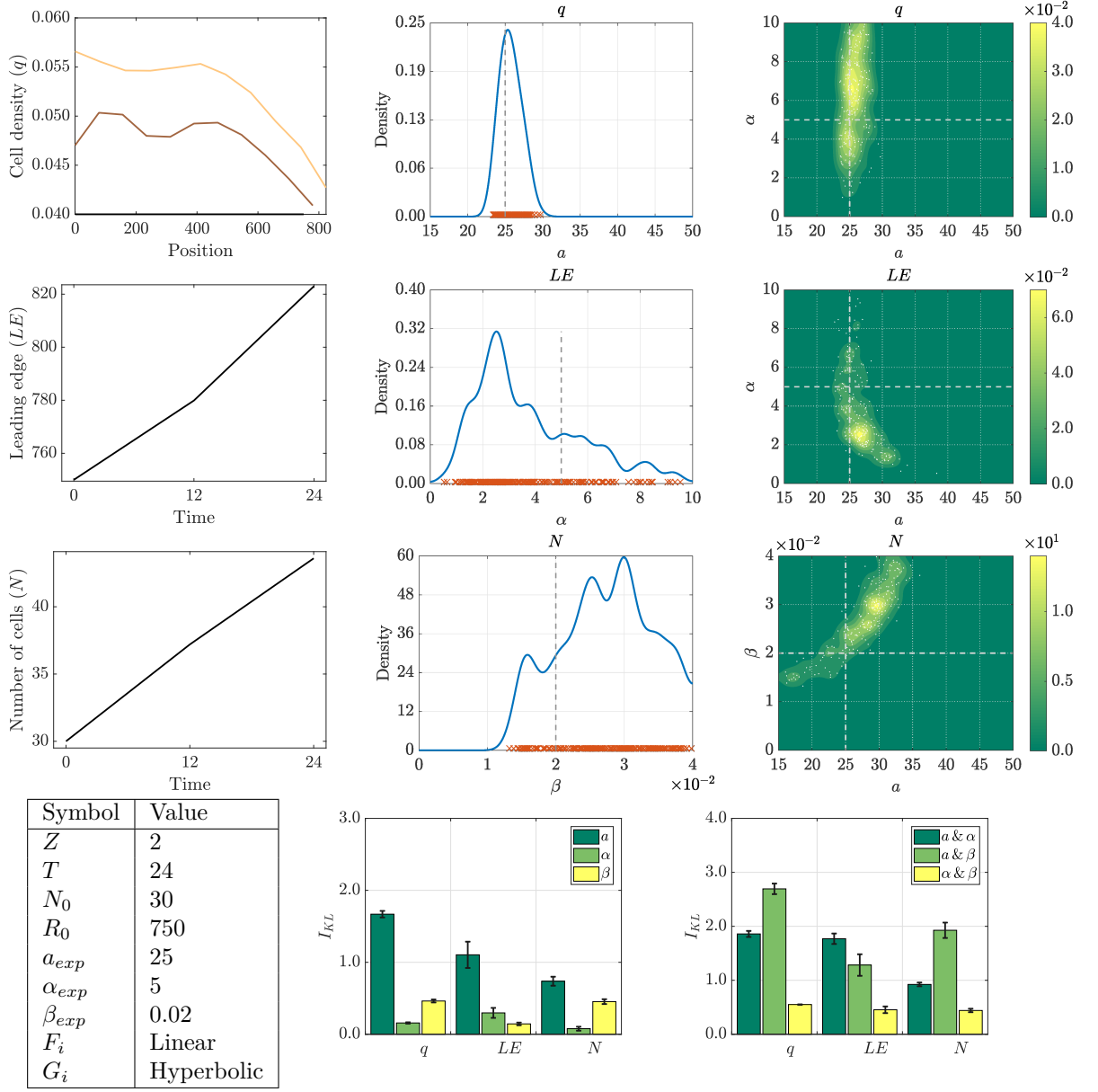


Figure C.2: We perform ABC-SMC with synthetic data computed using the experimental parameters shown. Prior distributions used and other inference choices are outlined in Table 4.1. The experiments performed here use the same inference choices as in Figure 4.7, but we use a proliferation law that is hyperbolic, as defined in Table 5.4. The synthetic experimental data is displayed in the left-most plots, for either the cell density, the leading edge, or the number of cells. The central plots show the marginal one-dimensional posterior distributions for parameters, a , α and β computed using statistics q , LE or N , which generally provide the most one-dimensional information gain. The right-most plots show marginal two-dimensional posteriors for a subset of the two-dimensional parameter combinations. The information gains computed from each marginal one-dimensional posteriors of marginal two-dimensional posteriors are shown for each statistic in the bar plots, where each information gain is shown with error-bars computed from three independent runs of ABC-SMC.

Appendix D

Numerical solution of the model in Chapter 5

In this Appendix, we present additional results from the coarse graining in Chapter 2, investigating higher order terms in the coarse graining, the method employed to numerically solve the full PDE system.

D.1 Investigating the higher order term

In this section we consider the impact of neglecting the $\mathcal{O}(1/N)$ term in equation (5.49). In practice, to ignore higher order terms in the numerical solver, set $E \equiv 0$ in the discretisations outlined in Appendix D.2.

Results are shown in Figure D.1, where the leading edge, the number of cells over time, and the cell density resulting from the model are compared with and without higher order terms. Plots show that these statistics are very similar with and without higher order terms, therefore it is advised that future extensions of this work should neglect these higher order terms to allow for ease of solving.

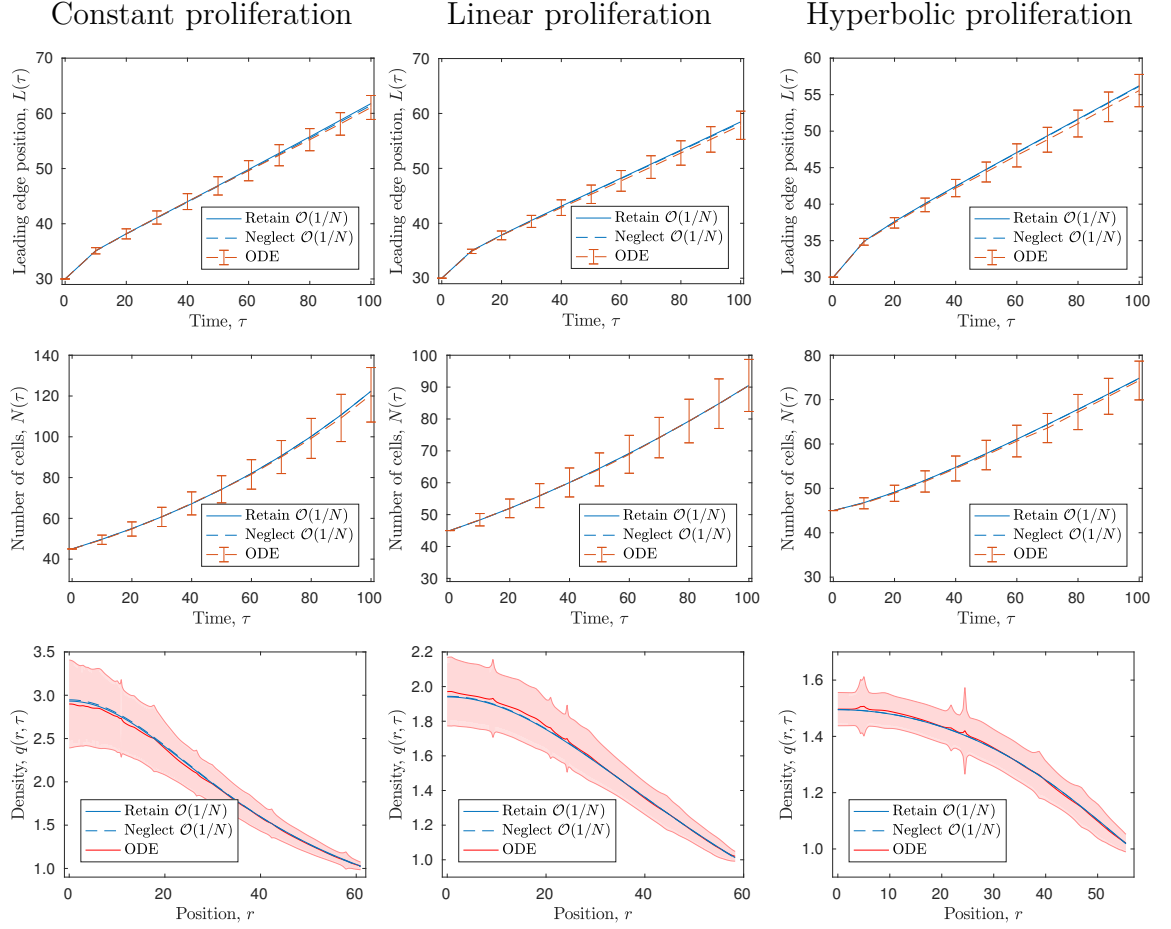


Figure D.1: Figure demonstrating the impact of neglecting the higher order proliferation terms in the continuum derivation leading to equation (5.51). Each column displays a different proliferation law, which are described in Table 5.4. Each row represents a different output of the system, either the leading edge position over time, the number of cells over time, or cell density recorded $\tau = 100$. A linear force law is used in all cases, initial node positions and parameters are as in Figure 5.3.

D.2 Numerical solution of the full system

The system of equations (5.71) - (5.74) in Section 5.4.6 is solved using a finite difference scheme with Picard iteration, which iteratively refines the solution of the PDE system at the next time-step, until two consecutive iterations are within a tolerance, ϵ , of each other.

In the following algorithm, k denotes the iteration number, j is the space-step being considered, $q^{(prev)}$ or $\Gamma^{(prev)}$ is the solution for q or Γ at the previous time-step, and q_j^k is the solution at iteration, k and space-step j . We solve for (Γ_j^k, q_j^k) at every iteration. We proceed to the next time-step when $d(q^k, q^{k+1}) < \epsilon$, where d is the sum of squared Euclidean distances function.

The space and time resolutions used are, $dx = 0.01$, $dt = 0.001$, with a tolerance $\epsilon = 10^{-4}$. The time taken for the PDE solver to complete for varying dx and dt are shown in Table D.1.

Algorithm 4 Numerical solution of the next iterate of the PDE system in Section 5.4.6.

```

t = 0
while t < T do
  for j = 1 to j = R do
    if k = 1 then
       $q_j^k = q_j^{(prev)}$ ,  $\Gamma_j^k = \Gamma_j^{(prev)}$ 
    end if
    Solve for  $\Gamma_j^{k+1}$  using  $q_{j-1}^k, q_j^k, q_{j+1}^k$  and  $\Gamma_{j-1}^k, \Gamma_j^k, \Gamma_{j+1}^k$ 
    Solve for  $q_j^{k+1}$  using  $q_{j-1}^k, q_j^k, q_{j+1}^k$  and  $\Gamma_j^{k+1}, \Gamma_{j-1}^{k+1}, \Gamma_{j+1}^{k+1}$ 
  end for
  if  $d(q^k, q^{k+1}) < \epsilon$  then
     $q^{(prev)} = q^{k+1}$ ,  $\Gamma^{(prev)} = \Gamma^{k+1}$ 
    k = 1
    t = t + dt
  else
    k = k + 1
  end if
end while

```

Equation (5.71) can be discretised:

$$\Gamma_j^{k+1} = \Gamma_j^{(prev)} - \mathrm{d}t \frac{D(q_j^k) + E(q_j^k)}{q_j^k} \frac{q_{j+1}^k - q_{j-1}^k}{\Gamma_{j+1}^k - \Gamma_{j-1}^k}.$$

Equation (5.72) can be discretised:

$$\begin{aligned}
q_j^{k+1} &= q_j^{(prev)} + dt \left[\frac{\Gamma_j^{k+1} - \Gamma_j^{(prev)}}{dt} \frac{(q_{j+1}^{k+1} - q_{j-1}^{k+1})}{(\Gamma_{j+1}^{k+1} - \Gamma_{j-1}^{k+1})^2} \right. \\
&\quad + \frac{2dR}{\Gamma_{j+1}^{k+1} - \Gamma_{j-1}^{k+1}} \frac{\partial}{\partial R} \left(\left(\frac{(D(q_{j+1/2}^k) + D(q_{j-1/2}^k))}{2} \right. \right. \\
&\quad \left. \left. \frac{(E(q_{j+1/2}^k) + E(q_{j-1/2}^k))}{2} \right) \frac{q_{j+1/2}^{k+1} - q_{j-1/2}^{k+1}}{\Gamma_{j+1/2}^{k+1} - \Gamma_{j-1/2}^{k+1}} \right) \\
&\quad \left. + q_j^k G(q_j^k) \right] \\
&= q_j^{(prev)} + dt \left[\frac{\Gamma_j^{k+1} - \Gamma_j^{(prev)}}{dt} \frac{(q_{j+1}^{k+1} - q_{j-1}^{k+1})}{(\Gamma_{j+1}^{k+1} - \Gamma_{j-1}^{k+1})^2} \right. \\
&\quad + \frac{2dR}{\Gamma_{j+1}^{k+1} - \Gamma_{j-1}^{k+1}} \frac{1}{dR} \left(\left(\frac{(D(q_{j+1}^k) + D(q_j^k))}{2} \right. \right. \\
&\quad \left. \left. + \frac{(E(q_{j+1}^k) + E(q_j^k))}{2} \right) \frac{q_{j+1}^{k+1} - q_j^{k+1}}{\Gamma_{j+1}^{k+1} - \Gamma_j^{k+1}} \right) \\
&\quad - \frac{2dR}{\Gamma_{j+1}^{k+1} - \Gamma_{j-1}^{k+1}} \frac{1}{dR} \left(\left(\frac{(D(q_j^k) + D(q_{j-1}^k))}{2} \right. \right. \\
&\quad \left. \left. + \frac{(E(q_j^k) + E(q_{j-1}^k))}{2} \right) \frac{q_j^{k+1} - q_{j-1}^{k+1}}{\Gamma_j^{k+1} - \Gamma_{j-1}^{k+1}} \right) \\
&\quad \left. + q_j^k G(q_j^k) \right] \\
&= q_j^k + dt q_j^k G(q_j^k) \\
&\quad + dt \left[- \frac{\Gamma_j^{k+1} - \Gamma_j^{(prev)}}{dt(\Gamma_{j+1}^{k+1} - \Gamma_{j-1}^{k+1})^2} + \frac{D(q_j^k) + D(q_{j-1}^k) + E(q_j^k) + E(q_{j-1}^k)}{(\Gamma_j^{k+1} - \Gamma_{j-1}^{k+1})(\Gamma_{j+1}^{k+1} - \Gamma_{j-1}^{k+1})} \right] q_{j-1}^{k+1} \\
&\quad + dt \left[- \frac{D(q_{j+1}^k) + D(q_j^k) + E(q_{j+1}^k) + E(q_j^k)}{(\Gamma_{j+1}^{k+1} - \Gamma_j^{k+1})(\Gamma_{j+1}^{k+1} - \Gamma_{j-1}^{k+1})} \right. \\
&\quad \left. - \frac{D(q_j^k) + D(q_{j-1}^k) + E(q_j^k) + E(q_{j-1}^k)}{(\Gamma_j^{k+1} - \Gamma_{j-1}^{k+1})(\Gamma_{j+1}^{k+1} - \Gamma_{j-1}^{k+1})} \right] q_j^{k+1} \\
&\quad + dt \left[\frac{\Gamma_j^{k+1} - \Gamma_j^{(prev)}}{dt(\Gamma_{j+1}^{k+1} - \Gamma_{j-1}^{k+1})^2} + \frac{D(q_{j+1}^k) + D(q_j^k) + E(q_{j+1}^k) + E(q_j^k)}{(\Gamma_{j+1}^{k+1} - \Gamma_j^{k+1})(\Gamma_{j+1}^{k+1} - \Gamma_{j-1}^{k+1})} \right] q_{j+1}^{k+1}
\end{aligned}$$

dx/dt	0.1	0.01	0.001
0.1	1.58	3.95	38.21
0.01	10.80	16.28	158.38
0.001	169.17	176.67	2378.36

Table D.1: The time taken in seconds to solve the PDE system in Section 5.4.6, using the algorithms outlined in this section, for varying space (dx) and time (dt) resolutions. The solver uses $N = 45$ cells with node positions uniformly distributed in $[0, 30]$, with $a = 1$, $\alpha = 15$, a linear force model, and no proliferation.

Defining the coefficients of q_{j-1}^{k+1} , q_j^{k+1} , and q_{j+1}^{k+1} in the above as $-a(q^k)$, $-b(q^k) + 1$, and $-c(q^k)$, respectively, allows us to write

$$\begin{aligned}
a &= -dt \left[-\frac{\Gamma_j^{k+1} - \Gamma_j^{(prev)}}{dt(\Gamma_{j+1}^{k+1} - \Gamma_{j-1}^{k+1})^2} + \frac{D(q_j^k) + D(q_{j-1}^k) + E(q_j^k) + E(q_{j-1}^k)}{(\Gamma_j^{k+1} - \Gamma_{j-1}^{k+1})(\Gamma_{j+1}^{k+1} - \Gamma_{j-1}^{k+1})} \right], \\
b &= 1 - dt \left[-\frac{D(q_{j+1}^k) + D(q_j^k) + E(q_{j+1}^k) + E(q_j^k)}{(\Gamma_{j+1}^{k+1} - \Gamma_j^{k+1})(\Gamma_{j+1}^{k+1} - \Gamma_{j-1}^{k+1})} - \frac{D(q_j^k) + D(q_{j-1}^k) + E(q_j^k) + E(q_{j-1}^k)}{(\Gamma_j^{k+1} - \Gamma_{j-1}^{k+1})(\Gamma_{j+1}^{k+1} - \Gamma_{j-1}^{k+1})} \right], \\
c &= -dt \left[\frac{\Gamma_j^{k+1} - \Gamma_j^{(prev)}}{dt(\Gamma_{j+1}^{k+1} - \Gamma_{j-1}^{k+1})^2} + \frac{D(q_{j+1}^k) + D(q_j^k) + E(q_{j+1}^k) + E(q_j^k)}{(\Gamma_{j+1}^{k+1} - \Gamma_j^{k+1})(\Gamma_{j+1}^{k+1} - \Gamma_{j-1}^{k+1})} \right],
\end{aligned}$$

hence,

$$a(q^k)q_{j-1}^{k+1} + b(q^k)q_j^{k+1} + c(q^k)q_{j+1}^{k+1} = q_j^k + dt q_j^k G(q_j^k),$$

which we can solve for q^{k+1} using the tridiagonal matrix algorithm and Picard iteration. At the end points, we discretise the boundary conditions from equation (5.74) in Section 5.4.6.

Bibliography

- [1] Simon Aeschbacher, Mark A. Beaumont, and Andreas Futschik. A novel approach for choosing summary statistics in approximate Bayesian computation. *Genetics*, 192(3):1027–1047, 2012. ISSN 00166731. doi: 10.1534/genetics.112.143164.
- [2] Oscar Aguilar, Moritz Allmaras, Wolfgang Bangerth, and Luis Tenorio. Statistics of parameter estimates: a concrete example. *SIAM Review*, 57(1):131–149, 2015. ISSN 0036-1445. doi: 10.1137/130929230.
- [3] Vivi Andasari, Alf Gerisch, Georgios Lolas, Andrew P. South, and Mark A. J. Chaplain. Mathematical modeling of cancer cell invasion of tissue: biological insight from mathematical analysis and computational simulation. *Journal of Mathematical Biology*, 63(1):141–171, 2011. ISSN 0303-6812. doi: 10.1007/s00285-010-0369-1.
- [4] Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society - Series B: Statistical Methodology*, 72(3):269–342, 2010. ISSN 13697412. doi: 10.1111/j.1467-9868.2009.00736.x.
- [5] Ruth E. Baker and Matthew J. Simpson. Correcting mean-field approximations for birth-death-movement processes. *Physical Review E - Statistical, Nonlinear,*

- and Soft Matter Physics*, 82(4):41905, 2010. ISSN 15393755. doi: 10.1103/PhysRevE.82.041905.
- [6] Chris P. Barnes, Sarah Filippi, Michael P. H. Stumpf, and Thomas Thorne. Considerate approaches to constructing summary statistics for ABC model selection. *Statistics and Computing*, 22(6):1181–1197, 2012. ISSN 09603174. doi: 10.1007/s11222-012-9335-7.
- [7] Mark A. Beaumont and Bruce Rannala. The Bayesian revolution in genetics. *Nature Reviews Genetics*, 5(4):251–261, 2004. ISSN 14710056. doi: 10.1038/nrg1318.
- [8] Mark A. Beaumont, Wenyang Zhang, and David J. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002. ISSN 0016-6731.
- [9] Benjamin J. Binder and Matthew J. Simpson. Quantifying spatial structure in experimental observations and agent-based simulations using pair-correlation functions. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 88(2):022705, 2013. ISSN 15393755. doi: 10.1103/PhysRevE.88.022705.
- [10] Benjamin J. Binder and Matthew J. Simpson. Cell density and cell size dynamics during in vitro tissue growth experiments: Implications for mathematical models of collective cell behaviour. *Applied Mathematical Modelling*, 40(4):3438–3446, 2016. ISSN 0307904X. doi: 10.1016/j.apm.2015.01.016.
- [11] Michael G. B. Blum. Regression approaches for Approximate Bayesian Computation. In Scott A. Sisson, Yanan Fan, and Mark A. Beaumont, editors, *Handbook of Approximate Bayesian Computation*, chapter Regression. Chapman and Hall/CRC, New York, 1st edition, 2018. ISBN 9781315117195. doi: 10.1201/9781315117195.

- [12] Benjamin M. Bolker. *Ecological Models and Data in R*. Princeton University Press, 2008. ISBN 9781400840908.
- [13] Zdravko I. Botev, Joseph F. Grotowski, and Dirk P. Kroese. Kernel density estimation via diffusion. *The Annals of Statistics*, 38(5):2916–2957, 2010. ISSN 0090-5364. doi: 10.1214/10-AOS799.
- [14] Lucie G. Bowden, Matthew J. Simpson, and Ruth E. Baker. Design and interpretation of cell trajectory assays. *Journal of the Royal Society Interface*, 10(88):20130630, 2013. ISSN 17425662. doi: 10.1098/rsif.2013.0630.
- [15] Rainer Burkard, Mauro Dell’Amico, and Silvano Martello. *Assignment Problems*. Society for Industrial and Applied Mathematics, 2012. ISBN 978-1-61197-222-1. doi: 10.1137/1.9781611972238.
- [16] Erkan O. Buzbas and Noah A. Rosenberg. AABC: Approximate approximate Bayesian computation for inference in population-genetic models. *Theoretical Population Biology*, 99:31–42, 2015. ISSN 0040-5809. doi: 10.1016/J.TPB.2014.09.002.
- [17] Anna Q. Cai, Kerry A. Landman, and Barry D. Hughes. Multi-scale modeling of a wound-healing cell migration assay. *Journal of Theoretical Biology*, 245(3): 576–594, 2007. ISSN 00225193. doi: 10.1016/j.jtbi.2006.10.024.
- [18] Annamaria Carusi, Kevin Burrage, and Blanca Rodríguez. Bridging experiments, models and simulations: an integrative approach to validation in computational cardiac electrophysiology. *AJP: Heart and Circulatory Physiology*, 303(2):H144–H155, 2012. ISSN 0363-6135. doi: 10.1152/ajpheart.01151.2011.
- [19] Gang Cheng, Janet Tse, Rakesh K. Jain, and Lance L. Munn. Micro-environmental mechanical stress controls tumor spheroid size and morphology

- by suppressing proliferation and inducing apoptosis in cancer cells. *PLoS ONE*, 4(2), 2009. ISSN 19326203. doi: 10.1371/journal.pone.0004632.
- [20] Kevin J. Cheung and Andrew J. Ewald. Illuminating breast cancer invasion: Diverse roles for cell-cell interactions. *Current Opinion in Cell Biology*, 30(1): 99–111, 2014. ISSN 18790410. doi: 10.1016/j.ceb.2014.07.003.
- [21] Claudio Cobelli and Joseph J. DiStefano. Parameter and structural identifiability concepts and ambiguities: a critical review and analysis. *American Journal of Physiology*, 239(1):R7–24, 1980. ISSN 0002-9513. doi: 10.1152/ajpregu.1980.239.1.R7.
- [22] Giles Cory. Scratch-Wound Assay. In Claire M. Wells and Maddy Parsons, editors, *Cell Migration: Methods and Protocols*, chapter 2, pages 25–30. Humana Press, 2 edition, 2011. ISBN 9781420008333.
- [23] Eileen Crowley and Alan F. Horwitz. Tyrosine phosphorylation and cytoskeletal tension regulate the release of fibroblast adhesions. *Journal of Cell Biology*, 131(2):525–537, 1995. ISSN 00219525. doi: 10.1083/jcb.131.2.525.
- [24] Gaudenz Danuser, Jun Allard, and Alex Mogilner. Mathematical modeling of Eukaryotic cell migration: Insights beyond experiments. *Annual Review of Cell and Developmental Biology*, 29:501–28, 2013. doi: 10.1146/annurev-cellbio-101512-122308.
- [25] Andreas Deutsch and Sabine Dorman. *Cellular Automaton Modeling of Biological Pattern Formation: Characterization, Applications, and Analysis*. Springer, Birkhauser, Boston, 2004.
- [26] Louise Dyson, Philip K. Maini, and Ruth E. Baker. Macroscopic limits of individual-based models for motile cell populations with volume exclusion.

- Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 86(3): 31903, 2012. ISSN 15393755. doi: 10.1103/PhysRevE.86.031903.
- [27] 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. ISSN 15446115. doi: 10.1515/sagmb-2012-0069.
- [28] Ronald A. Fisher. The wave of advance of advantageous genes. *Annals of Eugenics*, 7(4):355–369, 1937. ISSN 20501420. doi: 10.1111/j.1469-1809.1937.tb02153.x.
- [29] John A. Fozard, Helen M. Byrne, Oliver E. Jensen, and Julie R. King. Continuum approximations of individual-based models for epithelial monolayers. *Mathematical Medicine and Biology*, 27(1):39–74, 2009. ISSN 14778599. doi: 10.1093/imammb/dqp015.
- [30] 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): ra50, 2015. ISSN 19379145. doi: 10.1126/scisignal.aab0562.
- [31] Alan E. Gelfand and Adrian F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990. doi: 10.1080/01621459.1990.10476213.
- [32] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2004. ISBN 978-1-4398-9820-8. doi: 10.1007/s13398-014-0173-7.2.

- [33] Swapna A. Gudipaty, J. Lindblom, Patrick D. Loftus, M. J. Redd, K. Edes, C. F. Davey, V. Krishnegowda, and Jody Rosenblatt. Mechanical stretch triggers rapid epithelial cell division through Piezo1. *Nature*, 543(7643):118–121, 2017. ISSN 14764687. doi: 10.1038/nature21407.
- [34] Keegan E. Hines. A primer on Bayesian inference for biophysical systems. *Biophysical Journal*, 108(9):2103–2113, 2015. ISSN 15420086. doi: 10.1016/j.bpj.2015.03.042.
- [35] Junhwan Jeon, Vito Quaranta, and Peter T. Cummings. An off-lattice hybrid discrete-continuum model of tumor growth and invasion. *Biophysical Journal*, 98(1):37–47, 2010. doi: 10.1016/j.bpj.2009.10.002.
- [36] Wang Jin, Esha T. Shah, Catherine J. Penington, Scott W. McCue, Lisa K. Chopin, and Matthew J. Simpson. Reproducibility of scratch assays is affected by the initial degree of confluence: Experiments, modelling and model selection. *Journal of Theoretical Biology*, 390:136–145, 2016. ISSN 10958541. doi: 10.1016/j.jtbi.2015.10.040.
- [37] 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. ISSN 00092509. doi: 10.1016/j.ces.2018.01.004.
- [38] Stuart T. Johnston, Matthew J. Simpson, and Michael J. Plank. Lattice-free descriptions of collective motion with crowding and adhesion. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 88(6):62720, 2013. ISSN 15393755. doi: 10.1103/PhysRevE.88.062720.
- [39] Stuart T. Johnston, Matthew J. Simpson, and Donald L. S. McElwain. How much information can be obtained from tracking the position of the leading edge

- in a scratch assay? *Journal of the Royal Society Interface*, 11(97):20140325, 2014. ISSN 17425662. doi: 10.1098/rsif.2014.0325.
- [40] Stuart T. Johnston, Matthew J. Simpson, Donald L. S. McElwain, Benjamin J. Binder, and Joshua V. Ross. Interpreting scratch assays using pair density dynamics and approximate Bayesian computation. *Open Biology*, 4(9):140097, 2014. ISSN 20462441. doi: 10.1098/rsob.140097.
- [41] Stuart T. Johnston, Esha T. Shah, Lisa K. Chopin, Donald L. S. McElwain, and Matthew J. Simpson. Estimating cell diffusivity and cell proliferation rate by interpreting IncuCyte ZOOM™ assay data using the Fisher-Kolmogorov model. *BMC Systems Biology*, 9(1):38, 2015. ISSN 17520509. doi: 10.1186/s12918-015-0182-y.
- [42] Phoebe J. M. Jones, Aaron Sim, Harriet B. Taylor, Laurence Bugeon, Margaret J. Dallman, Bernard Pereira, Michael P. H. Stumpf, and Juliane Liepe. Inference of random walk models to describe leukocyte migration. *Physical Biology*, 12(6):066001, 2015. ISSN 14783975. doi: 10.1088/1478-3975/12/6/066001.
- [43] Georgiy P. Karev. Principle of minimum discrimination information and replica dynamics. *Entropy*, 12(7):1673–1695, 2010. ISSN 10994300. doi: 10.3390/e12071673.
- [44] Evgeniy Khain, Casey M. Schneider-Mizell, Michal O. Nowicki, Ennio A. Chiocca, Sean E. Lawler, and Leonard M. Sander. Pattern formation of glioma cells: effects of adhesion. *EPL: Europhysics Letters*, 88(2):6, 2008. doi: 10.1209/0295-5075/88/28006.
- [45] Paul Kirk, Thomas Thorne, and Michael P. H. Stumpf. Model selection in systems and synthetic biology. *Current Opinion in Biotechnology*, 24(4):767–774, 2013. ISSN 09581669. doi: 10.1016/j.copbio.2013.03.012.

- [46] Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. ISSN 0003-4851. doi: 10.1214/aoms/1177729694.
- [47] Jiyeon Lee, Yu L. Wang, Fan Ren, and Tanmay P. Lele. Stamp wound assay for studying coupled cell migration and cell debris clearance. *Langmuir*, 26(22):16672–16676, 2010. ISSN 07437463. doi: 10.1021/la103542y.
- [48] Subhash R. Lele, Brian Dennis, and Frithjof Lutscher. Data cloning: Easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecology Letters*, 10(7):551–563, 2007. ISSN 1461023X. doi: 10.1111/j.1461-0248.2007.01047.x.
- [49] Rong Li and Gregg G. Gundersen. Beyond polymer polarity: how the cytoskeleton builds a polarized cell. *Nature reviews. Molecular cell biology*, 9(11):860–73, 2008. ISSN 1471-0080. doi: 10.1038/nrm2522.
- [50] Juliane Liepe, Chris P. Barnes, Erika Cule, Kamil Erguler, Paul Kirk, Tina Toni, and Michael P. H. Stumpf. ABC-SysBio—approximate Bayesian computation in python with GPU support. *Bioinformatics*, 26(14):1797–1799, 2010. ISSN 13674803. doi: 10.1093/bioinformatics/btq278.
- [51] Juliane Liepe, Paul Kirk, Sarah Filippi, Tina Toni, Chris P. Barnes, and Michael P. H. Stumpf. A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. *Nature Protocols*, 9(2):439–456, 2014. ISSN 17542189. doi: 10.1038/nprot.2014.025.
- [52] Philip K. Maini, Donald L. S. McElwain, and David I. Leavesley. Traveling wave model to interpret a wound-healing cell migration assay for human peritoneal

- mesothelial cells. *Tissue Engineering*, 10(3-4):475–482, 2004. ISSN 1076-3279. doi: 10.1089/107632704323061834.
- [53] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003. ISSN 0027-8424. doi: 10.1073/pnas.0306899100.
- [54] Deborah C. Markham, Matthew J. Simpson, and Ruth E. Baker. Simplified method for including spatial correlations in mean-field approximations. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 87(6):62702, 2013. ISSN 15393755. doi: 10.1103/PhysRevE.87.062702.
- [55] Beatriz Martínez-Poveda, Ana R. Quesada, and Miguel Á. Medina. Hypericin in the dark inhibits key steps of angiogenesis in vitro. *European Journal of Pharmacology*, 516(2):97–103, 2005. ISSN 00142999. doi: 10.1016/j.ejphar.2005.03.047.
- [56] Alistair M. Middleton, Christian Fleck, and Ramon Grima. A continuum approximation to an off-lattice individual-cell based model of cell migration and adhesion. *Journal of Theoretical Biology*, 359:220–232, 2014. ISSN 10958541. doi: 10.1016/j.jtbi.2014.06.011.
- [57] Philip J. Murray, Carina M. Edwards, Marcus J. Tindall, and Philip K. Maini. From a discrete to a continuum model of cell dynamics in one dimension. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 80(3):031912, 2009. ISSN 15393755. doi: 10.1103/PhysRevE.80.031912.
- [58] Philip J. Murray, Carina M. Edwards, Marcus J. Tindall, and Philip K. Maini. Classifying general nonlinear force laws in cell-based models via the continuum

- limit. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 85(2):21921, 2012. ISSN 15393755. doi: 10.1103/PhysRevE.85.021921.
- [59] 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. ISSN 15446115. doi: 10.2202/1544-6115.1576.
- [60] James M. Osborne, Alexander G. Fletcher, Joe M. Pitt-Francis, Philip K. Maini, and David J. Gavaghan. Comparing individual-based approaches to modelling the self-organization of multicellular tissues. *PLoS Computational Biology*, 13(2), 2017. ISSN 15537358. doi: 10.1371/journal.pcbi.1005387.
- [61] Hans G. Othmer, Steven R. Dunbar, and Wolfgang Alt. Models of dispersal in biological systems. *Journal of Mathematical Biology*, 26(3):263–298, 1988. ISSN 03036812. doi: 10.1007/BF00277392.
- [62] Dirk Padfield, Jens Rittscher, and Badrinath Roysam. Coupled minimum-cost flow cell tracking for high-throughput quantitative analysis. *Medical Image Analysis*, 15(4):650–668, 2011. ISSN 13618415. doi: 10.1016/j.media.2010.07.006.
- [63] Andrew Parker, Matthew J. Simpson, and Ruth E. Baker. The impact of experimental design choices on parameter inference for models of growing cell colonies. *Royal Society Open Science*, 5(8), 2018. doi: 10.1098/rsos.180384.
- [64] Umberto Picchini and Rachele Anderson. Approximate maximum likelihood estimation using data-cloning ABC. *Computational Statistics and Data Analysis*, 105:166–183, 2017. ISSN 01679473. doi: 10.1016/j.csda.2016.08.006.
- [65] Michael J. Plank and Matthew J. Simpson. Models of collective cell behaviour with crowding effects: Comparing lattice-based and lattice-free approaches.

- Journal of the Royal Society Interface*, 9(76):2983–2996, 2012. ISSN 17425662. doi: 10.1098/rsif.2012.0319.
- [66] Michael J. Plank and Matthew J. Simpson. Lattice-free models of cell invasion: discrete simulations and travelling waves. *Bulletin of Mathematical Biology*, 75(11):2150–2166, 2013. ISSN 00928240. doi: 10.1007/s11538-013-9885-7.
- [67] Thomas D. Pollard and Gary G. Borisy. Cellular motility driven by assembly and disassembly of actin filaments. *Cell*, 112(4):453–465, 2003. ISSN 00928674. doi: 10.1016/S0092-8674(03)00120-X.
- [68] Nikodem J. Popławski, Maciej Swat, John S. Gens, and James A. Glazier. Adhesion between cells, diffusion of growth factors, and elasticity of the AER produce the paddle shape of the chick limb. *Physica A: Statistical Mechanics and its Applications*, 373:521–532, 2007. ISSN 03784371. doi: 10.1016/j.physa.2006.05.028.
- [69] Dennis Prangle. Summary statistics in approximate Bayesian computation. *bioRxiv*, 2015.
- [70] Dimitar Prodanov and Kris Verstreke. Automated Segmentation and Morphometry of Cell and Tissue Structures. Selected Algorithms in ImageJ. In *Molecular Imaging*. InTech, 2012. ISBN 978-953-51-0359-2. doi: 10.5772/36729.
- [71] Alberto Puliafito, Lars Hufnagel, Neveua Pierre, Sebastian Streichan, Sigal Alex, Deborah K. Fygenson, and Boris I. Shraiman. Collective and single cell behavior in epithelial contact inhibition. *Journal of Urology*, 188(4):1396–1397, 2012. ISSN 00225347. doi: 10.1016/j.juro.2012.06.073.
- [72] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula Klingmüller, and Jens Timmer. Structural and practical iden-

- tifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009. ISSN 13674803. doi: 10.1093/bioinformatics/btp358.
- [73] Brian Reid and Min Zhao. The electrical response to injury: molecular mechanisms and wound healing. *Advances in Wound Care*, 3(2):184–201, 2014. ISSN 2162-1918. doi: 10.1089/wound.2013.0442.
- [74] Aurélien Rizk, Grégory Paul, Pietro Incardona, Milica Bugarski, Maysam Mansouri, Axel Niemann, Urs Ziegler, Philipp Berger, and Ivo F. Sbalzarini. Segmentation and quantification of subcellular structures in fluorescence microscopy images using Squash. *Nature Protocols*, 9(3):586–596, 2014. ISSN 17502799. doi: 10.1038/nprot.2014.037.
- [75] Pere Roca-Cusachs, Vito Conte, and Xavier Trepac. Quantifying forces in cell biology, 2017. ISSN 14764679.
- [76] Robert J. H. Ross, Christian A. Yates, and Ruth E. Baker. Inference of cell-cell interactions from population density characteristics and cell trajectories on static and growing domains. *Mathematical Biosciences*, 264(1):108–118, 2015. ISSN 18793134. doi: 10.1016/j.mbs.2015.04.002.
- [77] Robert J. H. Ross, Ruth E. Baker, Andrew Parker, Matthew J. Ford, Richard L. Mort, and Christian A. Yates. Using approximate Bayesian computation to quantify cell–cell adhesion parameters in a cell migratory process. *npj Systems Biology and Applications*, 3(1):9, 2017. ISSN 2056-7189. doi: 10.1038/s41540-017-0010-7.
- [78] Gabriel A. Rosser, Alexander G. Fletcher, David A. Wilkinson, Jennifer A. de Beyer, Christian A. Yates, Judith P. Armitage, Philip K. Maini, and Ruth E. Baker. Novel methods for analysing bacterial tracks reveal persistence in

- rhodobacter sphaeroides. *PLoS Computational Biology*, 9(10):e1003276, 2013. ISSN 1553734X. doi: 10.1371/journal.pcbi.1003276.
- [79] Joseph Rudnick and George Gaspari. *Elements of the Random Walk: An Introduction for Advanced Students and Researchers*. Cambridge University Press, 2005. ISBN 9780521828918. doi: 10.1007/s10955-004-1994-0.
- [80] Conrad Sanderson and Ryan Curtin. Armadillo: a template-based C++ library for linear algebra. *Journal of Open Source Software*, 1:26, 2016. doi: 10.21105/joss.00026.
- [81] Linda Schneider, Michael Cammer, Jonathan Lehman, Sonja K. Nielsen, Charles F. Guerra, Iben R. Veland, Christian Stock, Else K. Hoffmann, Bradley K. Yoder, Albrecht Schwab, Peter Satir, and Søren T. Christensen. Directional cell migration and chemotaxis in wound healing response to PDGF-AA are coordinated by the primary cilium in fibroblasts. *Cellular Physiology and Biochemistry*, 25(2-3):279–292, 2010. ISSN 10158987. doi: 10.1159/000276562.
- [82] Jonathan A. Sherratt and James D. Murray. Models of epidermal wound healing. *Proceedings of the Royal Society B: Biological sciences*, 241(1300):29–36, 1990. doi: 10.1098/rspb.1990.0061.
- [83] Ivo Siekmann, Larry E Wagner, David Yule, Colin Fox, David Bryant, Edmund J Crampin, James Sneyd, and James Sneyd. MCMC estimation of Markov models for ion channels. *Biophysical journal*, 100(8):1919–29, 2011. ISSN 1542-0086. doi: 10.1016/j.bpj.2011.02.059.
- [84] Daniel Silk, Sarah Filippi, and Michael P. H. Stumpf. Optimizing threshold - schedules for approximate Bayesian computation sequential Monte Carlo samplers: applications to molecular systems. *bioRxiv*, pages 1–18, 2012.

- [85] Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall/CRC, 1988. ISBN 0412246201. doi: 10.2307/2347507.
- [86] Matthew J. Simpson, Alistair Merrifield, Kerry A. Landman, and Barry D. Hughes. Simulating invasion with cellular automata: Connecting cell-scale and population-scale properties. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 76(2):21918, 2007. ISSN 15393755. doi: 10.1103/PhysRevE.76.021918.
- [87] Matthew J. Simpson, Kerry A. Landman, and Barry D. Hughes. Multi-species simple exclusion processes. *Physica A: Statistical Mechanics and its Applications*, 388(4):399–406, 2009. ISSN 03784371. doi: 10.1016/j.physa.2008.10.038.
- [88] Matthew J. Simpson, Kerry A. Landman, and Barry D. Hughes. Cell invasion with proliferation mechanisms motivated by time-lapse data. *Physica A: Statistical Mechanics and its Applications*, 389(18):3779–3790, 2010. ISSN 03784371. doi: 10.1016/j.physa.2010.05.020.
- [89] Matthew J. Simpson, Chris Towne, Donald L. S. McElwain, and Zee Upton. Migration of breast cancer cells: Understanding the roles of volume exclusion and cell-to-cell adhesion. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 82(4):041901, 2010. ISSN 15393755. doi: 10.1103/PhysRevE.82.041901.
- [90] Matthew J. Simpson, Benjamin J. Binder, Parvathi Haridas, Benjamin K. Wood, Katrina K. Treloar, Donald L. S. McElwain, and Ruth E. Baker. Experimental and modelling investigation of monolayer development with clustering. *Bulletin of Mathematical Biology*, 75(5):871–889, 2013. ISSN 00928240. doi: 10.1007/s11538-013-9839-0.
- [91] Matthew J. Simpson, Katrina K. Treloar, Benjamin J. Binder, Parvathi Hari-

- das, Kerry J. Manton, David I. Leavesley, Donald L. S. McElwain, and Ruth E. Baker. Quantifying the roles of cell motility and cell proliferation in a circular barrier assay. *Journal of the Royal Society Interface*, 10(82):20130007, 2013. ISSN 17425662. doi: 10.1098/rsif.2013.0007.
- [92] Scott A. Sisson, Ying Fan, and Mark M. Tanaka. Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007. ISSN 0027-8424. doi: 10.1073/pnas.0607208104.
- [93] Mikael Sunnåker, Alberto G. Busetto, Elina Numminen, Jukka Corander, Matthieu Foll, and Christophe Dessimoz. Approximate Bayesian computation. *PLoS Computational Biology*, 9(1):e1002803, 2013. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002803.
- [94] Kristen F. Swaney, Chuan-Hsiang Huang, and Peter N. Devreotes. Eukaryotic chemotaxis: A network of signaling pathways controls motility, directional sensing, and polarity. *Annual Review of Biophysics*, 39(1):265–289, 2010. ISSN 1936-122X. doi: 10.1146/annurev.biophys.093008.131228.
- [95] Paul R. Taylor, Ruth E. Baker, and Christian A. Yates. Deriving appropriate boundary conditions, and accelerating position-jump simulations, of diffusion using non-local jumping. *Physical Biology*, 12(1):016006, 2014. doi: 10.1088/1478-3975/12/1/016006.
- [96] Eric Theveneau and Roberto Mayor. Neural crest delamination and migration: From epithelium-to-mesenchyme transition to collective cell migration. *Developmental Biology*, 366(1):34–54, 2012. ISSN 1095564X. doi: 10.1016/j.ydbio.2011.12.041.
- [97] Robin N. Thompson, Christian A. Yates, and Ruth E. Baker. Modelling cell

- migration and adhesion during development. *Bulletin of Mathematical Biology*, 74(12):2793–2809, 2012. ISSN 00928240. doi: 10.1007/s11538-012-9779-0.
- [98] William Thompson, Eric C. Rouchka, and Charles E. Lawrence. Gibbs recursive sampler: finding transcription factor binding sites. *Nucleic Acids Research*, 31(13):3580–3585, 2003. ISSN 03051048. doi: 10.1093/nar/gkg608.
- [99] Tina Toni and Michael P. H. Stumpf. Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics*, 26(1):104–110, 2010. doi: 10.1093/bioinformatics/btp619.
- [100] Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael P. H. Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202, 2009. ISSN 17425662. doi: 10.1098/rsif.2008.0172.
- [101] Katrina K. Treloar, Matthew J. Simpson, Donald L. S. McElwain, and Ruth E. Baker. Are in vitro estimates of cell diffusivity and cell proliferation rate sensitive to assay geometry? *Journal of Theoretical Biology*, 356:71–84, 2014. ISSN 10958541. doi: 10.1016/j.jtbi.2014.04.026.
- [102] Alexander Tremel, Anna Q. Cai, Nicolin Tirtaatmadja, Barry D. Hughes, Geoff W. Stevens, Kerry A. Landman, and Andrea J. O’Connor. Cell migration and proliferation during monolayer formation and wound healing. *Chemical Engineering Science*, 64(2):247–253, 2009. ISSN 00092509. doi: 10.1016/j.ces.2008.10.008.
- [103] Xavier Trepats and Erik Sahai. Mesoscale physical principles of collective cell organization. *Nature Physics*, 14(7):671–682, 2018. ISSN 17452481. doi: 10.1038/s41567-018-0194-9.

- [104] Janet M. Tse, Gang Cheng, James A. Tyrrell, Sarah A. Wilcox-Adelman, Yves Boucher, Rakesh K. Jain, and Lance L. Munn. Mechanical compression drives cancer cells toward invasive phenotype. *Proceedings of the National Academy of Sciences of the United States of America*, 109(3):911–916, 2012. ISSN 00278424. doi: 10.1073/pnas.1118910109.
- [105] Alan M. Turing. The chemical basis of morphogenesis. *Bulletin of Mathematical Biology*, 237(641):153–197, 1952. ISSN 00928240. doi: 10.1007/BF02459572.
- [106] Miguel Vicente-Manzanares and Alan R. Horwitz. Cell migration: An overview. In Claire M. Wells and Maddy Parsons, editors, *Cell Migration: Methods and Protocols*, chapter 1, pages 1–24. Humana Press, 2 edition, 2011.
- [107] Brenda N. Vo, Christopher C. Drovandi, Anthony N. Pettitt, and Matthew J. Simpson. Quantifying uncertainty in parameter estimates for stochastic models of collective cell spreading using approximate Bayesian computation. *Mathematical Biosciences*, 263:133–142, 2015. ISSN 18793134. doi: 10.1016/j.mbs.2015.02.010.
- [108] Trevor M. Wood, Christian A. Yates, David A. Wilkinson, and Gabriel A. Rosser. Simplified multitarget tracking using the phd filter for microscopic video data. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(5):702–713, 2012. ISSN 10518215. doi: 10.1109/TCSVT.2011.2177937.
- [109] Michele A. Wozniak and Christopher S. Chen. Mechanotransduction in development: A growing role for contractility. *Nature Reviews Molecular Cell Biology*, 10(1):34–43, 2009. ISSN 14710072. doi: 10.1038/nrm2592.
- [110] Christian A. Yates, Andrew Parker, and Ruth E. Baker. Incorporating pushing in exclusion-process models of cell migration. *Physical Review E*

- *Statistical, Nonlinear, and Soft Matter Physics*, 91(5):052711, 2015. doi: 10.1103/PhysRevE.91.052711.
- [111] Christian A. Yates, Matthew J. Ford, and Richard L. Mort. A multi-stage representation of cell proliferation as a Markov process. *Bulletin of Mathematical Biology*, 79(12):2905–2928, 2017. ISSN 0092-8240. doi: 10.1007/s11538-017-0356-4.
- [112] Patrick Y. K. Yue, Emily P. Y. Leung, N. K. Mak, and Ricky N. S. Wong. A simplified method for quantifying cell migration/wound healing in 96-well plates. *Journal of Biomolecular Screening*, 15(4):427–433, 2010. ISSN 10870571. doi: 10.1177/1087057110361772.
- [113] Yan Zhang, Guoqing Xu, Rachel M. Lee, Zijie Zhu, Jiandong Wu, Simon Liao, Gong Zhang, Yaohui Sun, Alex Mogilner, Wolfgang Losert, Tingrui Pan, Francis Lin, Zhengping Xu, and Min Zhao. Collective cell migration has distinct directionality and speed dynamics. *Cellular and Molecular Life Sciences*, 74(20):3841–3850, 2017. ISSN 14209071. doi: 10.1007/s00018-017-2553-6.
- [114] Min Zhao, Adriana Agius-Fernandez, John V. Forrester, and Colin D. McCaig. Directed migration of corneal epithelial sheets in physiological electric fields. *Investigative Ophthalmology and Visual Science*, 37(13):2548–2558, 1996. ISSN 01460404.
- [115] Juliane Zimmermann, Brian A. Camley, Wouter-Jan Rappel, and Herbert Levine. Contact inhibition of locomotion determines cell-cell and cell-substrate forces in tissues. *Proceedings of the National Academy of Sciences*, 113(10):2660–5, 2016. ISSN 1091-6490. doi: 10.1073/pnas.1522330113.