New Mathematical and Statistical Methods for Tissue Engineering

Daniel John VandenHeuvel
BMaths

Submitted in fulfilment
of the requirement for the degree of
Master of Philosophy

School of Mathematical Sciences
Faculty of Science
Queensland University of Technology

2023

Keywords: Equation learning; Mathematical modelling; Tissue growth; Pore geometry; Moving boundary problem; Parameter estimation.

Abstract

Mathematical models are routinely used in tissue engineering to study tissue growth in various types of experiments. These models can be obtained using a variety of methods, such as taking a phenomenological approach based on differential equations, or constructing discrete cell-based approaches and using coarse-graining to derive approximate differential equations. In this thesis, we begin by interpreting a new set of experiments describing thin tissue growth in 3D-printed scaffolds using a simple partial differential equation. By calibrating the solution of the model to match experimental observations, we explore whether rates of cell migration and cell proliferation depend upon the shape of the 3D-printed pores. Our results suggest that rates of migration and proliferation are apparently independent of pore shape, and this has important implications for how tissue growth experiments are designed and interpreted. The second part of the thesis involves a discrete mechanical cell-based model of epithelial tissues which leads to a nonlinear diffusion equation with a nonlinear source term, and a nonlinear moving boundary condition. While the continuum limit model has been derived and validated previously, here we focus on parameter choices where the solution of the continuum limit model does not give a good approximation to averaged data from the discrete model. We achieve this using equation learning methods, and our approach allows us to derive new tissue-scale partial differential equation models, whose solutions accurately capture averaged data from the corresponding discrete model in a way that was not possible using standard coarse-graining approaches.

Contents

List of Figures

List of Publications

This thesis is comprised of two papers:

- Located in Chapter 2.

[1] VandenHeuvel D, Devlin B, Buenzli P, Woodruff M, Simpson M. 2023 New computational tools and experiments reveal how geometry affects tissue growth in 3D printed scaffolds. *Chemical Engineering Journal* **475**, 145776.

- Located in Chapter 3.

[2] VandenHeuvel DJ, Buenzli PR, Simpson MJ. 2023 Pushing coarse-grained models beyond the continuum limit using equation learning. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 20230619.

Acknowledgements

Firstly, I would like to thank my principal supervisor Prof. Matthew Simpson and associate supervisor Dr Pascal Buenzli. Your support and guidance throughout this degree has been invaluable and has made a profound impact on my learning that I will continually reflect on throughout the rest of my academic journey, and without which this thesis would not be what it is today. I would also like to thank my second associate supervisor, Prof. Mia Woodruff, as well as Ms. Brenna Devlin who were extremely helpful with the experimental component of this work.

This research was supported by QUT Scholarship funding through a QUT Centre for Data Science scholarship, for which I acknowledge and thank the QUT Centre for Data Science. I also acknowledge and thank the additional support provided by the School of Mathematics, Faculty of Science at QUT. I further thank Prof. Scott McCue for many useful discussions about managing research life in general.

Lastly, I thank my family, Mum, Dad, Jonah, and Olivia for being patient with me and encouraging while I complete this thesis, with special gratitude to Adam, Isaac, and Leo. I thank Ms. Kate Skinner for being a significant help in shaping the student I have become, and how I work, today. I also thank my amazing partner Wing for her support which I could not have completed this thesis without.

Chapter 1

Introduction

1.1 Overview

Tissue engineering provides an approach for reconstructing and regenerating tissues and organs [3–5], and has been applied in a variety of contexts such as wound healing [6–9] and bone tissue regeneration [10]. A common experiment in tissue engineering is a tissue growth experiment, depicted in Figure 1.1 [11–13]. In these experiments, cells are initially seeded onto a boundary of scaffold, as shown in day 0 of Figure 1.1. These cells then detach from the scaffold boundary and move into the pore over a short period of time, where they then migrate and proliferate to grow the tissue as in days 1–18 of Figure 1.1, until the pore is eventually completely closed, as in day 28. A key quantity of interest in these experiments is the *bridging time*, namely the time it takes for the tissue to completely cover the pore [13–15]. Recent advancements in 3D printing and melt electrowriting technology for biofabrication [11, 12] have allowed for the efficient investigation of tissue growth in complicated pore geometries, enabling realistic experiments to be performed with precise control over various experimental conditions [10, 14, 16–20].

The influence of pore geometry on tissue growth is well-known, but how cellular-level mechanisms depend on pore geometry is poorly understood [21–23]. An interesting feature of tissue growth experiments, common to many geometries, is the appearance of a circular tissue front as tissue grows, which has been observed previously in square, triangular, hexagonal, and wave-like pore geometries [13, 24–26]. Two examples of a circular front are shown in Figure 1.2, where we show snapshots of images taken from

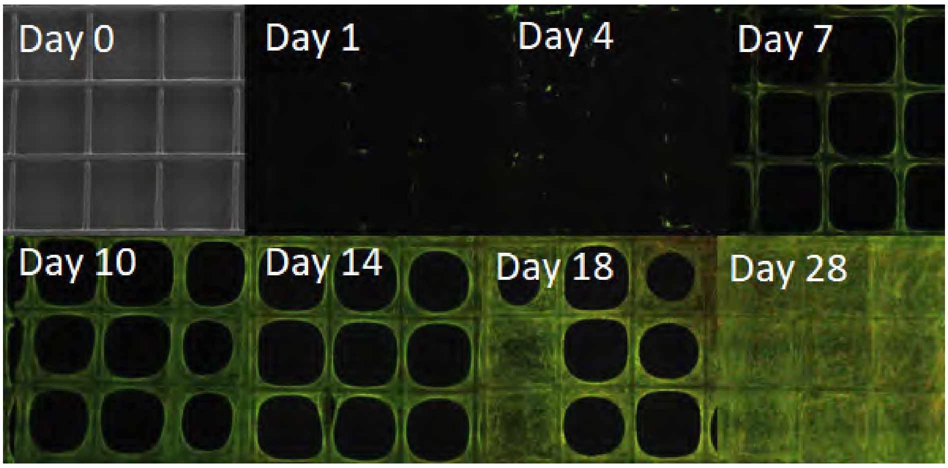

Figure 1.1: Snapshots from a series of tissue growth experiments in 3D printed scaffolds. Each panel shows a different experiment taken at the shown time. The image at day 0 is the initial scaffold of imaged by Scanning Electron Microscopy, where each individual pore is of size 500 × 500 µm. The remaining days show the tissue growth over time, depicting the cell nuclei (red), cytoskeleton (green), and the void (black). Image adapted from Ref. [13] with permission from Elsevier.

a tissue growth experiment on a square pore (left) and a wave-like pore (right); the tissue growth experiments in Figure 1.2 are what we consider in this thesis, where the cells used are murine calvarial osteoblastic cells (MC3T3-E1) [27]. The similarities between the experiments on these geometries, namely that both the square pore and the wave-like pores both show a circular front in Figure 1.2, invokes the question of whether the cellular mechanisms driving tissue growth are independent of pore shape, and whether we can use experimental results from one geometry to make inference about, for example, bridging times on other geometries. Chapter 2 investigates these questions, where we demonstrate that the cellular mechanisms are independent of pore shape and show how to extrapolate results from one geometry onto a new geometry. We demonstrate this independence using mathematical modelling with a likelihood-based uncertainty framework. The likelihood component of this work is crucial as it allows us to not only estimate parameters, denoted by $\boldsymbol{\theta}$, describing the cellular mechanisms using a likelihood function, but it enables us to make predictions with uncertainty so that we can make informed predictions between different geometries. In particular, using the recent work of Simpson and Maclaren [28], a key component of the work in Chapter 2 is the propagation of uncertainty in $\boldsymbol{\theta}$ into some function of the parameters, $q(\boldsymbol{\theta})$, which in this thesis describes either the density profiles themselves, or statistics such as the bridging time.

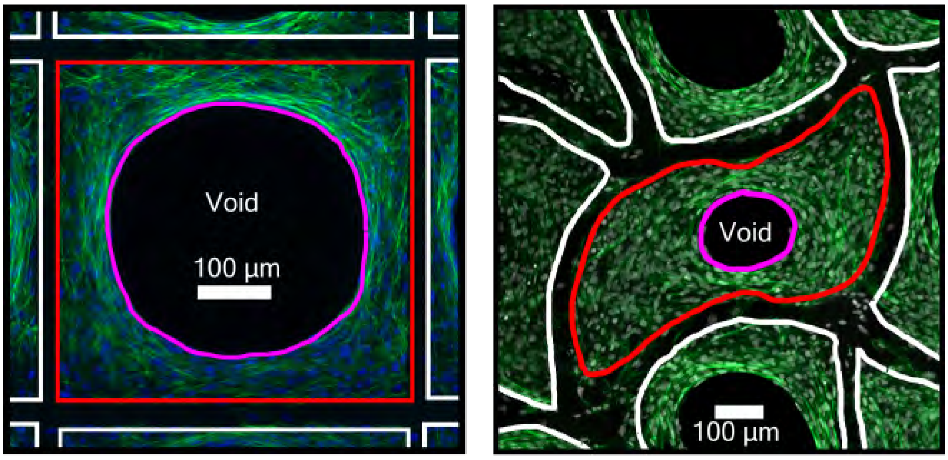

Figure 1.2: Examples of circular fronts arising in tissue growth experiments. The images show the scaffold pore boundary (red boundary), void (black interior), void boundary (magenta boundary), tissue (blue/green region), and fibres (exterior black boundary). The white outlines show the scaffold boundaries. The blue channel in the microscope images shows the cell nuclei (DAPI), and the green channel shows the tissue and cytoskeleton (phalloidin). The DAPI on the right image is shown in grey.

Mathematical modelling is a crucial part of tissue engineering, as experiments by themselves do not provide direct information about observed effects, such as the cellular mechanisms [29]. Many mathematical models have been developed for modelling tissue growth [30–32]. The work in Chapter 2 uses a Porous-Fisher model from Browning et al. [24], given by

$$\begin{aligned}\frac{\partial u(\mathbf{x},t)}{\partial t} &= \overbrace{D\boldsymbol{\nabla}\cdot\left[\frac{u(\mathbf{x},t)}{K}\boldsymbol{\nabla}u(\mathbf{x},t)\right]}^{\text{contact stimulated cell migration}} + \overbrace{\lambda u(\mathbf{x},t)\left[1-\frac{u(\mathbf{x},t)}{K}\right]}^{\text{contact inhibited cell proliferation}}, \quad \mathbf{x}\in\Omega,\\ \frac{\mathrm{d}u(\mathbf{x},t)}{\mathrm{d}t} &= \underbrace{\lambda u(\mathbf{x},t)\left[1-\frac{u(\mathbf{x},t)}{K}\right]}_{\text{contact inhibited cell proliferation}}, \quad \mathbf{x}\in\partial\Omega,\end{aligned} \tag{1.1}$$

where $u(\mathbf{x},t)$ denotes the cell density at time t and position $\mathbf{x}$, D the cell diffusivity, λ the cell proliferation rate, K the carrying capacity density, Ω is the pore interior bounded by the red boundaries in Figure 1.2, and $\partial\Omega$ is the boundary of Ω shown in red. It is through the parameters D and λ in

(1.1), together with the likelihood-based framework in Chapter 2, that we are able to make the connection between experimental results and insights about the individual cellular mechanisms that describe the experiments in this thesis; without the model (1.1), we cannot directly estimate effects related to these cellular mechanisms.

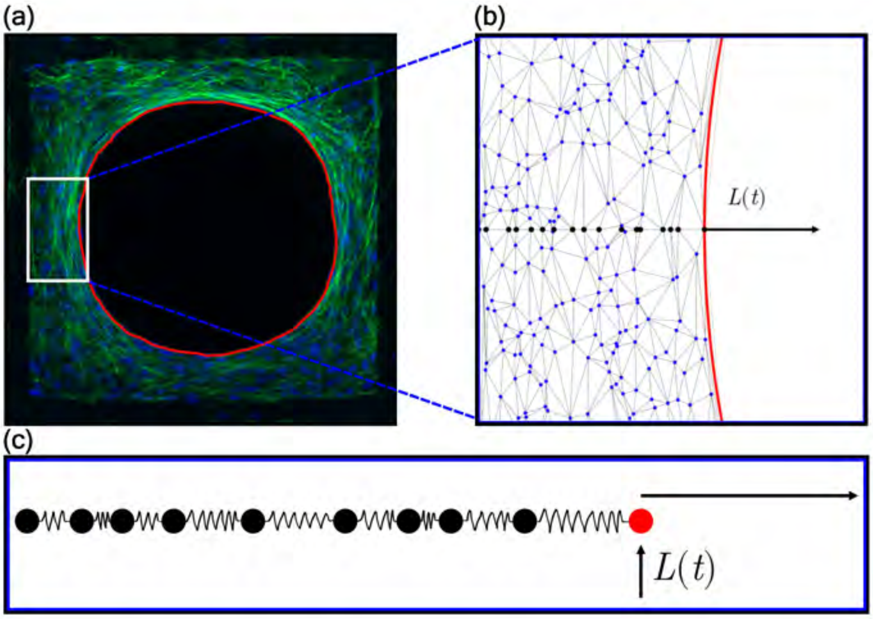

Figure 1.3: Simplifying a discrete mechanical model of cells. (a) A snapshot from a tissue growth experiment, with the red boundary showing the void boundary. (b) The zoomed-in region from the white rectangle in (a). The blue dots represent individual cells, and the edges indicate connections between cells that allow the cells to interact. Black dots show a slice through the two-dimensional cell configuration, with the leading edge $L(t)$ shown by the base of a black arrow indicating the direction of $L(t)$. (c) Zoomed-in view of the black dots from (b), with springs showing how cell boundaries are connected and the red dot showing the position of the leading edge.

When considering more complicated experiments, it may be the case that models such as (1.1) are not known, which limits the generalisability of the framework developed in Chapter 2 since we do not have parameters such as D and λ in Equation (1.1) that we can easily calibrate and interpret according to our hypotheses. Models for such experiments are traditionally derived through empirical reasoning [29, 33], using physical and conservation arguments to determine what terms should be included in the model. For example, Equation (1.1) can be derived by assuming that the migration of a cell depends on the cells surrounding the cell, leading to contact stimulated cell migration, and assuming that the growth of cells grows logistically both inside the pore and on the boundary so that we have contacted inhibited cell proliferation [24, 32]. An alternative approach to modelling these experiments is to use individual-based discrete mechanical models, where we instead model individual cells rather than cell densities [34]. Such an approach is advantageous in this situation as individual-based models are typically easier to derive, using arguments based on Newton's laws [34], and since modelling cells individually allows properties such as proliferation rates to be assigned to cells directly. In the case of experiments like those shown in Figure 1.3(a), one approach could be to connect individual cells using a network model and then allow cells to exert forces onto other cells if they share an edge in the network [34–37], as shown in Figure 1.3(b). The edge connections could be modelled so that, from Hooke's law, [34]

$$\eta \frac{\mathrm{d}\mathbf{x}_i}{\mathrm{d}t} = \sum_{j \in \mathcal{N}_i(t)} k\left(s - \|\mathbf{x}_{ij}(t)\|\right)\hat{\mathbf{x}}_{ij}(t), \tag{1.2}$$

where $\mathbf{x}_i$ is the position of the ith cell, $\mathcal{N}_i(t)$ are the set of cells sharing an edge with $\mathbf{x}_i$ at the time t, η is the drag coefficient that representing the viscosity of the surrounding medium, k is the spring constant that controls how the cells relax, $\mathbf{x}_{ij} = \mathbf{x}_j - \mathbf{x}_i$ is the edge connecting cells $\mathbf{x}_i$ and $\mathbf{x}_j$ and $\hat{\mathbf{x}}_{ij} = \mathbf{x}_{ij}/\|\mathbf{x}_{ij}\|$ is the associated unit vector, and s is the resting spring length that controls the long-time positions of the cells; extra effects are also included in the model so that cell proliferation is included [34]. The sum in (1.2) gives the total force on the ith cell. Since the void boundaries, also called the *leading edge*, form circular fronts as in Figure 1.2, it is reasonable to simplify the model (1.2) into a one-dimensional problem as shown in Figure 1.3(c), fixing the left-most cell at $x = 0$ and allowing the right-most cell at $x = L(t)$ to be free as in Baker et al. [38]. In this work, we consider this simplified model for modelling epithelial tissue dynamics, although future work can consider the more complicated two-dimensional problem as discussed in Chapter 4.

While discrete models of tissue growth are useful for studying cellular behaviour [34, 39], interpreting collective behaviour from simulations of the models can be difficult or even computationally infeasible for large populations [40]. Thus, a related continuum model is often desired so that macroscopic details can be analysed. Continuum models describing averaged data from a discrete model are typically obtained by coarse-graining using a Taylor series expansion, resulting in a partial differential equation (PDE) model that governs the cell densities [38, 41–44]. The coarse-grained model for the one-dimensional analogue of Equation (1.2), in the context of epithelial tissues, is given by a reaction-diffusion model with nonlinear diffusion of the form [38, 41]

$$\frac{\partial q}{\partial t} = \frac{\partial}{\partial x}\left(D(q)\frac{\partial q}{\partial x}\right) + R(q) \quad 0 < x < L(t),\, t > 0, \tag{1.3}$$

where $q(x,t)$ is the macroscopic density at position x and time t, $D(q) = k/(\eta q^2)$ describes the mechanical relaxation of the springs, $R(q) = \beta q(1 - q/K)$ is the source term describing the cell proliferation with β the intrinsic proliferation rate, and there are extra terms not shown that describe the evolution of the free boundary at $x = L(t)$. This model (1.3) is only accurate when the time scale of mechanical relaxation is sufficiently fast relative to the time scale of proliferation [45]. In cases where this condition on the time scales of mechanical relaxation and proliferation do not hold, continuum models describing the cell dynamics are not known, and so an alternative approach to coarse-graining is needed to derive accurate continuum descriptions. Since the discrete models will still obey a conservation principle, we expect that, even where the solution to the model (1.3) is not accurate compared to the averaged discrete data, the macroscopic densities will be governed by some macroscopic conservation description, such as a generalised form of (1.3) [33, 46]. Thus, one approach to learning an accurate continuum model would to treat the functions in (1.3) more generally, using the recently developed field of *equation learning* for learning functions that better describe the macroscopic densities [47–50]. In particular, whereas in Chapter 2 our interest is in estimating *parameters*, we need to now estimate *functions* describing the data which is the topic of Chapter 3.

Equation learning is a means for model discovery. In the context of PDEs, equation learning typically considers models of the form $\partial q/\partial t = \mathcal{N}(q, \mathcal{D}, \boldsymbol{\theta})$, where $\mathcal{N}$ is some nonlinear function parametrised by $\boldsymbol{\theta}$, $\mathcal{D}$ is a collection of differential operators for q with respect to x, and $\boldsymbol{\theta}$ are parameters to be *learned* [51], as first considered by Brunton et al. [47] and Rudy et al. [51]. These parameters $\boldsymbol{\theta}$ control the functional form $\mathcal{N}$, and are different than traditional model parameters as in the tissue growth model (1.1). For example, in the context of ordinary differential equations (ODEs), Brunton et al. [47] write

$$\frac{\mathrm{d}q}{\mathrm{d}t} = \sum_{i=1}^{p} \theta_i \varphi_i(q) \tag{1.4}$$

for some pre-specified library of functions $\varphi_1, \ldots, \varphi_p$ with coefficients θ_i to be estimated, $i = 1, \ldots, p$. Brunton et al. [47] then compare both sides of (1.4) using provided time series data to estimate $(\theta_1, \ldots, \theta_p)$ using sparse regression, giving an ODE with few terms that describes the data; to make this point clearer, and to highlight some differences between this approach and what we develop in Chapter 3, we give in Appendix A an example of the approach of Brunton et al. [47]. Equation learning has been applied to many biological problems, such as by Lagergren et al. [49] and VandenHeuvel et al. [50] who, representing $\mathcal{N}$ as a conservation law rather than as a general nonlinear function, learn models describing simple *in vitro* experiments. In the context of discrete models, as considered in this thesis, recent work has also been used to learn continuum models from averaged discrete data by Nardini et al. [48] and Simpson et al. [52], but thus far no work has been considered for problems with a moving boundary as in (1.3).

In this thesis, we apply methods from equation learning to the problem of learning continuum models, like those in (1.3), describing macroscopic behaviours of a discrete mechanical model of epithelial tissues, especially for cases where known continuum models are no longer accurate. The framework in Chapter 3 we develop is simple to implement, interpretable, modular so that models describing complex experiments to be considered, and is generally applicable to any discrete model describing population densities. The generalisability of our approach makes it ideal for acting as a basis for future work that develops two-dimensional analogues of the discrete models considered and could, for example, be used for learning models that describe the tissue growth experiments from Chapter 2 and for more complicated experiments in the field of tissue engineering.

1.2 Structure of this thesis

This thesis is derived from two papers that respectively comprise Chapter 2 and Chapter 3. Chapter 2 is published in the *Chemical Engineering Journal* [1]. Chapter 3 is published in the *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* [2]. The text in these chapters is the same as the version's submitted to their respective journals, with minor typesetting and formatting changes. In this section, we summarise the layout of this thesis and the contents of these individual chapters. In addition to what we describe below, we highlight that the introductions in Chapter 2 and Chapter 3 each contain components of a literature review for their respective topics, thereby omitting the need for a literature review in this chapter.

In Chapter 2, we present a study of the affect of geometry on tissue growth in 3D-printed scaffolds, analysing time series data of tissue growth experiments from two geometries. We use a likelihood-based framework to calibrate a mathematical model of tissue growth to the experimental data for each geometry. Using the calibrated models, we show that we can extrapolate predictions of tissue growth and bridging times, with uncertainty quantification, on other geometries, and that these predictions match the experimental results. Thus, we demonstrate that the cellular mechanisms driving tissue growth are independent of pore geometry.

Chapter 3 is related to the problem of discovering models describing tissue growth experiments in cases where known models are no longer accurate. Using a one-dimensional discrete mechanical model of epithelial tissues as an example, we show how we can use equation learning techniques to learn continuum models describing macroscopic features of a discrete model, especially in cases where previously known models are no longer accurate.

We summarise our findings in Chapter 4, and discuss possibilities for future research extending this work.

1.3 Statement of joint authorship

This section outlines the contribution of the Master's student and the coauthors of each paper.

Chapter 2: New computational tools and experiments reveal how geometry affects tissue growth in 3D printed scaffolds

This chapter is a slightly modified form of a paper titled "New computational tools and experiments reveal how geometry affects tissue growth in 3D printed scaffolds", published in the *Chemical Engineering Journal* [1]. The contribution of each author is listed below:

- VandenHeuvel, DJ: Performed all the analyses, assisted with the methodology, developed all the code and figures, and drafted the manuscript.
- Devlin, BL: Performed all the experiments, assisted with the image processing, and critically reviewed the manuscript.
- Buenzli, PR: Initiated the research concept and methodology, supervised the analyses, and critically reviewed the manuscript.
- Woodruff, MA: Initiated the research concept and supervised the experiments.
- Simpson, MJ: Initiated the research concept and methodology, supervised the analyses, and critically reviewed the manuscript.

Chapter 3: Pushing coarse-grained models beyond the continuum limit using equation learning

This chapter is a slightly modified form of a paper titled "Pushing coarse-grained models beyond the continuum limit using equation learning", published in the *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* [2]. The contribution of each author is listed below:

- VandenHeuvel, DJ: Performed all the analyses and numerical simulations, developed, the methodology, developed all the code and figures, and drafted the manuscript.

- Buenzli, PR: Supervised the analyses and numerical simulations and critically reviewed the manuscript.
- Simpson, MJ: Initiated the research concept, supervised the analyses and numerical simulations, and critically reviewed the manuscript.

Statement of Contribution of Co-Authors

The authors listed below have certified that:

1. they meet the criteria for authorship and that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
3. there are no other authors of the publication according to these criteria;
4. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit, and
5. they agree to the use of the publication in the student's thesis and its publication on the QUT's ePrints site consistent with any limitations set by publisher requirements.

Publication title: New computational tools and experiments reveal how geometry affects tissue growth in 3D printed scaffolds

Publication status: Published in Chemical Engineering Journal

Contributor	**Statement of contribution***
Daniel J. VandenHeuvel	Performed all the analyses, assisted with the methodology, developed all the code and figures, and drafted the manuscript.
Brenna L. Devlin	Performed all the experiments, assisted with the image processing, and critically reviewed the manuscript.
Pascal R. Buenzli	Initiated the research concept and methodology, supervised the analyses, and critically reviewed the manuscript.
Maria A. Woodruff	Initiated the research concept and supervised the experiments.
Matthew J. Simpson	Initiated the research concept and methodology, supervised the analyses, and critically reviewed the manuscript.

Chapter 2

New computational tools and experiments reveal how geometry affects tissue growth in 3D printed scaffolds

A paper published in the Chemical Engineering Journal as

New computational tools and experiments reveal how geometry affects tissue growth in 3D printed scaffolds

Daniel J. VandenHeuvel, Brenna L. Devlin, Pascal R. Buenzli, Maria A. Woodruff, Matthew J. Simpson

Abstract

Understanding how tissue growth in porous scaffolds is influenced by geometry is a fundamental challenge in the field of tissue engineering. We investigate the influence of pore geometry on tissue growth using osteoblastic cells in 3D printed melt electrowritten scaffolds with square-shaped pores and non-square pores with wave-shaped boundaries. Using a reaction-diffusion model together with a likelihood-based uncertainty quantification framework, we quantify how the cellular mechanisms of cell migration and cell proliferation drive tissue growth for each pore geometry. Our results show that the rates of cell migration and cell proliferation appear to be largely independent of the pore geometries considered, suggesting that observed curvature effects on local rates of tissue growth are due to space availability rather than directly affecting cell behaviour. This result allows for simple squared-shaped pores to be used for estimating parameters and making predictions about tissue growth in more realistic pores with more realistic, complicated shapes. Our findings have important implications for the development of predictive tools for tissue engineering and experimental design, highlighting new avenues for future research.

2.1 Introduction

Tissue engineering aims to regenerate damaged or diseased tissues [4]. A key challenge in tissue engineering is to understand how various clinically-motivated experimental conditions influence tissue growth [5]. Recent advancements in three-dimensional (3D) printing allow us to investigate tissue growth in 3D-printed scaffolds of various shapes and sizes, enabling realistic migration and proliferation behaviours to be studied in well-controlled experimental conditions [10, 16]. While the effect of pore geometry and tissue curvature on tissue growth is well-known [21–23], understanding how these effects relate to cellular-level mechanisms remains poorly understood. Understanding these cellular mechanisms would enable the prediction and analysis of tissue growth in complex geometries from the calibration of mathematical models in simpler geometries, providing a valuable computational tool for screening experimental designs [53] of scaffold geometries and providing plausible results on these new geometries. This kind of computational tool could enable more personalised approaches to tissue engineering, in which a specific scaffold size and shape could be tailored to an individual patient, making use of these predictions to screen for possible patient outcomes.

3D printing technology for biofabrication has evolved rapidly [11, 12]. Melt electrowriting, a technique for high quality 3D printing, allows for micro- and nano-scale fibres to created, enabling great control over the fibres and the pore geometry, making it possible to produce realistic scaffold geometries with a regular array of pores for growing tissue [17–20]. These scaffolds are designed so that cells and tissues experience a similar mechanical support as they would, for example, in skin and bone tissues, enabling realistic cell migration and cell proliferation behaviours to be observed and measured [10]. Previous work has focused primarily on squared-shaped pores [13, 24], although more complicated scaffolds can also be produced [12, 54]. One important factor in understanding tissue growth is curvature [55, 56]. In the context of bone tissue, Bidan et al. [57] suggest that cell tension can influence tissue curvature that, in turn, can stimulate tissue growth. Callens et al. [58] discuss how cells respond to their surrounding geometry, even across large spatial scales, and how this affects bone tissue growth. Mathematical modelling studies performed by Alias and Buenzli [59–61] and Hegarty-Cremer et al. [62] also investigate the role of geometry and curvature on tissue growth and cell crowding in bone tissue growth experiments.

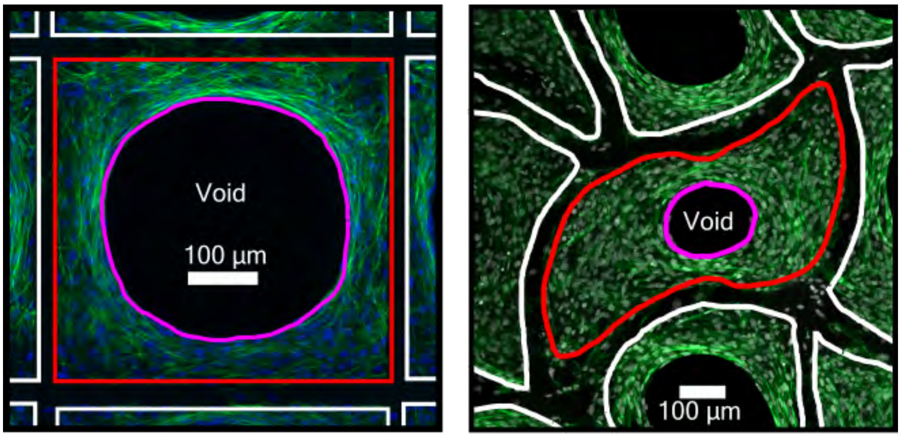

Figure 2.1: Scaffold pore boundary (red boundary), void (black interior), void boundary (magenta boundary), tissue (blue/green region), and fibres (exterior black boundary). The white outlines show the scaffold boundaries. The blue channel in the microscope images shows the cell nuclei (DAPI), and the green channel shows the tissue and cytoskeleton (phalloidin). The DAPI on the right image is shown in grey.

Tissue growth experiments in porous scaffolds are of great importance in tissue engineering [13,14]. In these experiments, cells are seeded onto the perimeter of a scaffold, leading to cell migration and cell proliferation that produces an inward-growing tissue. The shape of the region that is devoid of cell and tissue material, referred to as the *void* in Figure 2.1, matches the shape of the scaffold boundary for early times, rounding off over time, eventually forming a circular front until the void closes, which we call *pore bridging*. The feature of interest in these experiments is the time that the tissue bridges, called the *bridging time* [13, 14]. A circular front arises with time in many pore scaffold geometries such as square, triangular, and hexagonally shaped pores [13, 24–26], though the precise mechanisms driving the cells into these circular fronts in general geometries remains unclear. The most common way to report a set of pore bridging experiments is to record snapshot images of the tissue growth process [13, 14, 24]. This approach allows us to estimate the bridging time within an interval instead of identifying the precise time of bridging [13, 14], thereby introducing some uncertainty into the experimental estimate of the bridging time. To interpret such measurements meaningfully, we are interested in developing mathematical modelling tools that mechanistically capture cell migration and cell proliferation within a framework that explicitly incorporates uncertainty in the experimental measurements, uncertainty in the parameter estimates in the mathematical model, and that is capable of making predictions of new experiments that incorporate these uncertainties. This will allow us, for example, to study how variability in bridging times can be explicitly integrated into the mathematical model, as well as interpreting the predictions of the mathematical model.

In this work, we study mechanistic cell behaviour in pore bridging experiments performed within pores on 3D-printed scaffolds made from polycaprolactone [12,14,16]. We consider tissue growth in two different shaped pores, a square-shaped pore (Figure 2.1(a)) and a wave-shaped pore (Figure 2.1(b)), with the aim to understand whether the details of cell migration and cell proliferation are affected by differences in the pore geometry. In particular, we are interested in determining whether the cellular mechanisms driving tissue growth in the more realistic wave-shaped pore are indistinguishable from the cellular mechanisms driving tissue growth in the simpler square geometry. Moreover, we are interested in obtaining estimates of bridging time with uncertainty, for both pore geometries, through probability distributions that allow a user to predict probabilities of bridging time occurring within a specified time window for any pore geometry. All experiments reported in this work involve tissue growth using murine calvarial osteoblastic cells (MC3T3-E1) [27]. Our group has previously used these experiments on the square pores with different sizes [13, 14, 24], investigating the relationship between the cell migration rate measured in terms of the cell diffusivity D, cell proliferation rate λ, and scaffold size. This previous work showed that the product $D\lambda$, which controls the long-time rate of tissue production, appeared to be unaffected by the pore size, but did not consider the role of pore shape. These previous experiments displayed variability in the time to bridge, even in well controlled experiments with the same pore size. This variability motivates the need for developing mathematical modelling tools that incorporate variability and uncertainty quantification into predictions. Hence, our analysis uses a combination of numerical simulations from a mathematical model and with statistical analysis that takes numerical simulations and quantifies results together with uncertainty. These considerations will be used to answer the following broad questions:

1. Are the cellular mechanisms driving tissue growth independent of pore shape?

2. Can we use results on one pore shape to make predictions, with uncertainty, on another geometry?

3. What data, and how much data, is sufficient for accurately comparing results for different shaped pores?

We address these questions using a model-based approach, using the Porous-Fisher partial differential equation (PDE) as a model for tissue growth driven by combined cell migration and cell proliferation [32]. Solving this mathematical model requires parameter estimates for the cell diffusivity, D, the cell proliferation rate, λ, and the initial cell density on the scaffold boundary, u_0. Estimates for D and λ cannot be obtained directly from experimental images, hence methods that use information about the images are needed. For each pore shape, we calibrate this mathematical model with an experimental dataset containing information about the position of the tissue front over time. We apply a likelihood-based analysis to this dataset [63] with the aim to estimate the combined effect of proliferation and migration rates, by estimating the product $D\lambda$. Profile likelihoods are used to quantify the uncertainty in $D\lambda$, providing confidence intervals for $D\lambda$ [64], and allowing us to determine what parameters or parameter combinations can be estimated [65,66], providing insights into the second and third research questions listed above. The confidence intervals obtained on each geometry can be used to compare the tissue growth mechanisms for each pore geometry to answer the first research question. This analysis also enables us to make predictions with uncertainty about the pore bridging time. By using this likelihood-based approach to make predictions, we can take results from the square-shaped pores and estimate, with uncertainty, the pore bridging times on the wave-shaped pores. Similarly, we can use the wave-shaped pores to make predictions on the square-shaped pores. Comparing both situations allows us to answer the second research question. The simplest interpretation of our results is that the cell migration and cell proliferation rates are independent of the pore scaffold geometry.

2.2 Materials and methods

In this section, we describe the methods used for the experiments and the data that we collect from these experiments. Following this description, we introduce the mathematical models we use and how we apply likelihood analysis for performing statistical inference from these experiments.

2.2.1 Tissue growth experiments

All reagents are sourced from Thermo Fisher unless otherwise stated. Using a melt electrowriting printer described previously [14], polycaprolactone (45 kDa, Sigma Aldrich) fibres of diameter 50 µm are fabricated into a three-layer scaffold which was then biopsy punched to form a 6 mm disc. The resultant scaffold has a thickness of approximately 150 µm. The code used to produce the outline of each pore is adjusted to produce square-shaped and wave-shaped pores of comparable size, both being derived from a unit of cell of 500 µm. Prior to cell seeding, scaffolds are sterilised under UV light overnight. The cells used are murine calvarial osteoblastic cells (MC3T3-E1) [27] that are cultured in α-MEM, 10 % fetal bovine serum, and 1 % penicillin-streptomycin, and are approximately 20–30 µm in diameter [14]. Cells were expanded in a T75 culture flask and at 80 % confluency were detached with TrypLE. Cells were seeded at a density of 10,000 cells per scaffold in 48-well plates. After allowing 4 h for the cells to attach to each scaffold after seeding, an additional 500 µL was added. Cell-seeded scaffolds were cultured in a humidified environment at 37 °C in 5 % CO2 for 28 days. The media is changed every 2–3 days from day 5 to day 14, every 1–2 days from day 15 to 28. The viability of the cells are assessed at day 10, 14, and 28 using calcein-AM and ethidium homodimer-1 to stain live and dead cells, respectively. At specific timepoints, cell-seeded scaffolds were fixed with 4 % paraformaldehyde and stained with 4’,6-diamidino-2-phenylindole (DAPI) and Alexa Fluor™ 488 Phalloidin, which stain cell nuclei and actin filaments, respectively. High resolution images of the centre of each scaffold are obtained using fluorescent microscopy (Zeiss, AxioObserver 7). For each pore shape and timepoint, fixation, staining, and microscopy are repeated across two or three identically prepared replicates. Each experimental replicate provides information of several pores, giving information about tissue growth data from day 5 to day 28. This procedure gives 41 data points for the square pore, and 3 data points for the wave pore; while this number of data points is relatively small, it is sufficient for our analysis as we use a likelihood-based uncertainty quantification framework that naturally incorporates variability associated with finite sample sizes, as introduced in Section 2.2.4. More information on the procedure for capturing these images is provided in [14]. These experiments lead to a tissue composed of a monolayer of cells within the pore. The growth of monolayers is achieved through the design of the pore geometry as the width of the pores is large compared to the pore depth. The vertical pore depth is approximately equal to the average cell diameter, 20–30 µm [14], which means that we do not observe cells growing on top of each other in the vertical direction.

2.2.2 Data and image processing

The experiments provide us with several images at four time points (days 7, 14, 25, and 28), and each image contains information about several pores. To summarise the tissue growth processes, for each pore we calculate two quantities:

$$y_{\mathrm{c}}^{i,j} = \frac{\text{area of void}}{\text{area of pore}}, \quad y_{\mathrm{p}}^{i,j} = \frac{\text{perimeter of void}}{\text{perimeter of pore}}, \tag{2.1}$$

where $y_{\mathrm{c}}^{i,j}$ and $y_{\mathrm{p}}^{i,j}$ denote the *void coverage* and *normalised void perimeter*, respectively, for the jth pore at the ith time, t_i. We do not collect data from any pores that have bridged, or pores that have bridged in a way so that the void splits into multiple disconnected regions. All images at $t = 25$ day and $t = 28$ day show that all pores have bridged before 25 days, so we focus on measuring (2.1) at $t_1 = 7$ day and $t_2 = 14$ day. A precise description of how we compute the quantities in (2.1) from the images is given in Appendix 2.A.

To complete the processing, we select the computational representations for the square and wave geometries for use in the mathematical model described Section 2.2.3 (Figure 2.2) for comparison with each data point. In the square case, we construct this independent of the images, and simply define a boundary for a square with its lower-left corner at the origin and side length $L = 475$ µm. For the wave geometry, we take a single image from the experiments and choose its boundary as a representative boundary for each experiment, which is reasonable as all scaffolds are uniformly printed.

2.2.3 Mathematical model

We use the Porous-Fisher PDE to model tissue growth, as this model explicitly describes how cell migration and cell proliferation leads to tissue growth with sharp fronts that we observe in the experiments (Figures 2.3 and 2.4) [24, 32]. Since the pore bridging process starts after 5 days, the PDE is solved for $t > 5$ day. Letting $\tilde{u}(\mathbf{x}, t)$ [cells/µm^2] denote the density of cells at a point $\mathbf{x} = (x, y)^\mathsf{T}$ and time t [day], and $\tilde{K}$ denoting the maximum carrying capacity density [cells/µm^2], we define a normalised density $u(\mathbf{x}, t) \in [0, 1]$ by $u(\mathbf{x}, t) = \tilde{u}(\mathbf{x}, t)/\tilde{K}$. Thus, the model for $u(\mathbf{x}, t)$ is given by

$$\frac{\partial u(\mathbf{x}, t)}{\partial t} = \overbrace{D\boldsymbol{\nabla}\cdot\left[u(\mathbf{x}, t)\boldsymbol{\nabla}u(\mathbf{x}, t)\right]}^{\text{contact stimulated cell migration}} + \overbrace{\lambda u(\mathbf{x}, t)\left[1 - u(\mathbf{x}, t)\right]}^{\text{contact inhibited cell proliferation}}, \quad \mathbf{x} \in \Omega, \tag{2.2}$$

$$\frac{\mathrm{d}u(\mathbf{x}, t)}{\mathrm{d}t} = \underbrace{\lambda u(\mathbf{x}, t)\left[1 - u(\mathbf{x}, t)\right]}_{\text{contact inhibited cell proliferation}}, \quad \mathbf{x} \in \partial\Omega, \tag{2.3}$$

$$u(\mathbf{x}, 5) = \begin{cases} u_0 & \mathbf{x} \in \partial\Omega, \\ 0 & \mathbf{x} \in \Omega. \end{cases} \tag{2.4}$$

where (2.2) is applied on the interior scaffold pore space Ω, the space inside the red curves of Figure 2.2; (2.3) is applied on the boundary $\partial\Omega$, the red curve in Figure 2.2. We note that while our scaffolds are three-dimensional, their thickness is small compared to their width, and so this two-dimensional depth-averaged model (2.2)–(2.4) is reasonable [67]. This model is characterised by three parameters (D, λ, u_0), where D [µm^2/day] is the cell diffusivity that controls the rate of cell migration, λ [day^{-1}] is the cell proliferation rate, and u_0 is the normalised density of cells on the scaffold boundary $\partial\Omega$ at $t = 5$ day. We solve Equations (2.2)–(2.4) numerically using the finite volume method with an unstructured triangular mesh (Figure 2.2), as described in Appendix 2.B.

Solutions of Equations (2.2)–(2.4) are used to compute model predictions corresponding to the data $y_\mathrm{c}^{i,j}$ and $y_\mathrm{p}^{i,j}$ from (2.1). Using the numerical solution for $u(\mathbf{x}, t)$, we identify the contour $u(\mathbf{x}, t) = 1/2$ to indicate the location of the tissue front [13, 24]. We define predictions of the $y_\mathrm{c}^{i,j}$ and

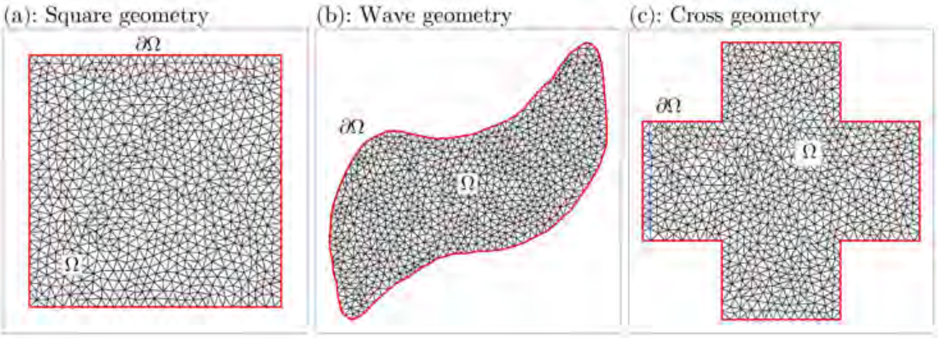

Figure 2.2: Schematics for triangular meshes used for numerical solutions of Equations (2.2)–(2.4) on the (a) square-shaped pore, (b) the wave-like pore, and (c) the cross-shaped pore considered in Section 2.3.3. Denser meshes are used for the actual solutions. The triangles represent the mesh we use for studying these geometries computationally, as discussed in Section 2.2.3. The red curves represent the boundary $\partial\Omega$, and the region bounded by these curves is the interior domain Ω.

$y_{\mathrm{p}}^{i,j}$ by

$$\mu_{\mathrm{c}}(t) = \frac{1}{A(\Omega)}\sum_{k=1}^{n} A_k(t), \quad \mu_{\mathrm{p}}(t) = \frac{1}{\ell(\partial\Omega)}\sum_{k=1}^{n} \ell_k(t), \tag{2.5}$$

respectively, where n is the number of triangular elements in the mesh, $A_k(t)$ is the area of the portion of the kth element at the time t that is inside the contour $u(\mathbf{x}, t) = 1/2$, $\ell_k(t)$ is the length of the line through the kth element at the time t that is on the contour $u(\mathbf{x}, t) = 1/2$ or zero if the contour does not go through the element, and $A(\Omega)$ and $\ell(\partial\Omega)$ are the area and perimeter of the domain Ω, respectively. More details on how we compute the coverage μ_{c} and normalised perimeter μ_{p} are given in Appendix 2.C.

We remark that our definition of the PDE (2.2)–(2.4) involves working with a nondimensional dependent variable u and nondimensional parameter u_0, while retaining the dimensional parameters D and λ and dimensional variables x, y, and t, so that spatial and temporal features can be compared with experimental images, similar to [68]. Our interpretation of the dependent variable, $u(\mathbf{x}, t)$, is different, though. While it is possible to work with the dimensional density $\tilde{u}(\mathbf{x}, t)$ [24, 30], this would require manually counting cells to estimate cell densities in space and time [68–70]. Thus, to be consistent with the fact that we only treat the leading edge in the experimental images for computing $y_{\mathrm{c}}^{i,j}$ and $y_{\mathrm{p}}^{i,j}$, it makes sense to consider the dimensionless ratio $u(\mathbf{x}, t) = \tilde{u}(\mathbf{x}, t)/\tilde{K}$ rather than $\tilde{u}(\mathbf{x}, t)$ itself.

2.2.4 Parameter estimation

We use a likelihood-based approach to estimate model parameters [28,65]. This approach takes predictions $\mu_{\mathrm{c}}(t)$ and $\mu_{\mathrm{p}}(t)$ from solutions of Equations (2.2)–(2.4) and compares them with the noisy experimental observations, $y_{\mathrm{c}}^{i,j}$ and $y_{\mathrm{p}}^{i,j}$. We assume that these noisy experimental observations are all independent realisations of random variables Y_{c}^{i} and Y_{p}^{i}, respectively, that are defined by [24]

$$Y_{\mathrm{c}}^{i} \sim \mathcal{N}\left(\mu_{\mathrm{c}}(t_i; \boldsymbol{\theta}), \sigma_{\mathrm{c}}^2\right), \quad \text{and} \quad Y_{\mathrm{p}}^{i} \sim \mathcal{N}\left(\mu_{\mathrm{p}}(t_i; \boldsymbol{\theta}), \sigma_{\mathrm{p}}^2\right), \tag{2.6}$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. These random variables in (2.6) each have a mean that depends on $\boldsymbol{\theta} = (D, \lambda, u_0)$, and the variances σ_{c}^2 and σ_{p}^2 need to be estimated. The values of σ_{c} and σ_{p} are treated as constants that are pre-estimated using the sample standard deviation of the experimental data aggregated for each t_i and each j, as described in Appendix 2.C. We note that, for measurements taken from the same experiment, there may be some dependence between $y_{\mathrm{c}}^{i,j}$ and $y_{p}^{i,j}$, due to area and perimeter being slightly related to each other, contradicting our assumption that Y_{c}^{i} and Y_{p}^{i} are independent. Typically, falsely assuming independence might cause issues with statistical inference, however for this type of data this will not be the case [71].

Given sufficient experimental data we could, in theory, estimate all model parameters $\boldsymbol{\theta}$ and σ_{c}^2 and σ_{p}^2 directly, however, as we will show, our data is insufficient for this purpose. We find that our numerical simulations of Equations (2.2)–(2.4) on the time scale of our experiments are relatively independent of u_0, and so we show results for a range of pre-specified values of u_0 rather than focusing on any single value, demonstrating this independence. While it would be ideal to estimate both D and λ, we find that it is difficult to treat them separately, as we show in Appendix 2.D. [13,30]. For our purposes, though, we are mainly interested in the combined effect $D\lambda$, as this is the variable that affects the velocity of the tissue and thus the bridging time [13]. Thus, rather than using $\boldsymbol{\theta} = (D, \lambda, u_0)$, we instead re-parametrise the vector of model parameters as $\boldsymbol{\theta} = (D\lambda, \lambda)$, omitting u_0 and following [72]. With this definition, $\mu_{\mathrm{c}}(t_i; \boldsymbol{\theta})$ and $\mu_{\mathrm{p}}(t_i; \boldsymbol{\theta})$

still refer to predictions from the solution of the PDE with parameters D and λ, and the choice of the fixed value of u_0 is left implicit.

Log-likelihood function

The log-likelihood function $\ell(\boldsymbol{\theta} \mid \mathbf{y})$ is a function of the model parameters that describes the likelihood that the model has parameter values $\boldsymbol{\theta}$ given that the data observed is $\mathbf{y}$ [63, 73]. In this work we have

$$\ell(\boldsymbol{\theta} \mid \mathbf{y}) = \sum_{i=1}^{2} \sum_{j=1}^{J(i)} \left[\log \phi\left(y_{\mathrm{c}}^{i,j}; \mu_{\mathrm{c}}\left(t_i; \boldsymbol{\theta}\right), \sigma_{\mathrm{c}}^2\right) + \log \phi\left(y_{\mathrm{p}}^{i,j}; \mu_{\mathrm{p}}\left(t_i; \boldsymbol{\theta}\right), \sigma_{\mathrm{p}}^2\right)\right], \tag{2.7}$$

where $\mathbf{y}$ is the vector of observations,

$$\phi(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

is the normal probability density function, and $J(i)$ is the number of pores included at the time t_i. For the wave pore, the log-likelihood (2.7) only includes the sum at $i = 2$ as there is only data at $t_2 = 14$ day in this case. For notational convenience we write $\ell(\boldsymbol{\theta} \mid \mathbf{y})$ as $\ell(\boldsymbol{\theta})$. The log-likelihood depends on the parameters $\boldsymbol{\theta} = (D\lambda, \lambda)$ and the fixed values for u_0, σ_{c}, and σ_{p}, with $\{D, \lambda, u_0\}$ governing the solution to the PDE and σ_{p} and σ_{c} governing the measurement model.

Maximum likelihood estimation

We obtain a best-fit estimate for the parameters $\boldsymbol{\theta}$ by maximising $\ell(\boldsymbol{\theta})$. This procedure is called maximum likelihood estimation [63], and it results in a maximum likelihood estimate (MLE) for $\boldsymbol{\theta}$, denoted $\hat{\boldsymbol{\theta}}$. For this maximisation, we constrain the values of D and λ so that $0\,\mu\mathrm{m}^2/\mathrm{day}^2 < D\lambda \leq 10\,000\,\mu\mathrm{m}^2/\mathrm{day}^2$. For λ we use $0\,\mathrm{day}^{-1} < \lambda \leq 5\,\mathrm{day}^{-1}$ in the square and $0\,\mathrm{day}^{-1} < \lambda \leq 10\,\mathrm{day}^{-1}$ in the wave. These bounds do not affect the results significantly. More detail on how we perform this maximum likelihood estimation is given in Appendix 2.E.

Uncertainty quantification

One limitation of maximum likelihood estimation is that we obtain a single point estimate for the MLE $\hat{\boldsymbol{\theta}}$, and the asymptotic uncertainty in this point estimate depends upon the curvature of the log-likelihood function [74]. To quantify the uncertainty in this estimate, we combine two approaches. For the first approach, we evaluate the log-likelihood function $\ell(\boldsymbol{\theta})$ over a large grid of $D\lambda$ and λ values. We then use this grid to find all points where $\ell(\boldsymbol{\theta}) - \ell^* \geq -\chi^2_{2,1-\alpha}/2$, where $\chi^2_{d,q}$ is the qth quantile of the χ^2 distribution with d degrees of freedom and $\ell^* = \ell(\hat{\boldsymbol{\theta}})$ is the maximum likelihood, as the resulting set of values defines a $100(1-\alpha)\%$ confidence region (CR) for $\boldsymbol{\theta}$ [63, 75]. We use $\alpha = 0.05$ in this work, giving $-\chi^2_{2,0.95}/2 \approx -3$.

The approach above gives us a two-dimensional region representing the uncertainty in $\boldsymbol{\theta}$. It will also be useful to reduce these regions to confidence intervals for each parameter, and most importantly for the parameter combination $D\lambda$, which allows us to make predictions about the variability in tissue growth later in Section 2.2.4. We take a profile likelihood approach to consider each parameter individually, specifying a range of values for an interest parameter and using numerical optimisation, reducing the log-likelihood function to a series of interpretable univariate functions. These univariate results then provide insight into the curvature of the log-likelihood function, and hence the uncertainty in the MLE point estimates. The resulting univariate function results in what is known as the *profile log-likelihood function* [63]. Following [24, 28, 65], we define the profile log-likelihood in terms of the interest parameter $D\lambda$. For a given value of $D\lambda$, we define the normalised profile log-likelihood:

$$\ell_p(D\lambda) = \max_{\lambda \in \Lambda} \left[\ell(D\lambda, \lambda)\right] - \ell^*, \tag{2.8}$$

where $\ell(D\lambda, \lambda) = \ell(\boldsymbol{\theta})$, $\ell^* = \ell(\hat{\boldsymbol{\theta}})$, and $\Lambda = \{\lambda : 0\,\mathrm{day}^{-1} < \lambda \leq 5\,\mathrm{day}^{-1}\}$. This definition gives a simple univariate function of $D\lambda$ that reaches a maximum of zero at the MLE, and the curvature of this function is related to inferential precision — a profile log-likelihood function with a well-defined peak at zero indicates a parameter that has been well estimated and identified, while a flat profile means that the data was insufficient for estimating or obtaining any inference for that parameter [28, 76]. A useful feature of (2.8) is that it can be used for constructing approximate confidence intervals for $D\lambda$, with an approximate $100(1-\alpha)\%$ confidence interval given by the set of all $D\lambda$ such that $\ell_p(D\lambda) \geq -\chi^2_{1,1-\alpha}/2$ [63]. In this work, we use $\alpha = 0.05$ so that we are constructing 95% confidence intervals, giving $c^* = -\chi^2_{1,0.95}/2 \approx -1.92$. The procedure we use for computing profile likelihoods is implemented in the Julia package `ProfileLikelihood.jl` [77], and a summary of the procedure is outlined in Appendix 2.E.

Parameter-wise prediction intervals

The two-dimensional likelihood function allows us to propagate the uncertainty in $D\lambda$ through to give us a prediction interval in terms of the outcome of the mathematical model for a variable of interest, such as the bridging time or cell density. Using the approach developed by [28, 65] which builds on basic properties of likelihood function [63], we are able to quantify the uncertainty in the cell densities $u(\mathbf{x}, t)$ and the bridging time t_b directly from our likelihood function. In particular, by taking pairs of parameter values inside of the 95% confidence region from the log-likelihood function and computing the variable of interest at each pair, we obtain a sample of values that gives the uncertainty in our variable of interest, as described in Appendix 2.E. We use this method to obtain prediction intervals for $y_{\mathrm{c}}^{i,j}$ and $y_{\mathrm{p}}^{i,j}$ over time. Moreover, we can obtain prediction intervals for the bridging time, t_b, at which $\mu_{\mathrm{c}}(t)$ first becomes zero. We emphasise that these prediction intervals are computed by treating the average prediction as a *predictive* quantity [28], so that the uncertainty is being computed relative to the mean prediction rather than relative to the experimental data points themselves, although a portion of this uncertainty does come from the data since σ_{c}^2 and σ_{p}^2 are pre-estimated using the data. Prediction intervals using the latter approach would be wider than with our current approach, but our approach is more useful in this work for assessing the calibration of the model and for predicting expected quantities on other geometries. In addition to prediction intervals, this procedure returns a sample of bridging times, which we use to obtain probability distributions for the bridging time via `KernelDensity.jl` [78]. We represent this probability distribution using a probability density function (PDF) $p(t_b)$ for the bridging time which can be understood as [73]

$$p(t_b)\Delta t \approx \mathbb{P}(t < t_b < t + \Delta t), \tag{2.9}$$

where $\mathbb{P}(t < t_b < t + \Delta t)$ is the probability that the bridging time t_b is between t and $t + \Delta t$, given the uncertainty in the parameters $\boldsymbol{\theta}$, and Δt is some small sufficiently interval of time. This PDF $p(t_b)$ allows us to compute probabilities that the bridging time occurs in *any* given interval.

The complete procedure for how we obtain these results is implemented in `ProfileLikelihood.jl` [77], and the method that we implement is outlined in Appendix 2.E.

Predicting variability in tissue growth

We further extend our results to provide a more qualitative approach to assess the uncertainty in the tissue growth on these pores. Taking values for the parameters inside their confidence intervals from the profile likelihoods, we can produce time series model predictions of the solution to Equations (2.2)–(2.4), indicating the variability that we might expect in the tissue growth. For these predictions, we take three parameter values for $\boldsymbol{\theta}$: (1) $\hat{\boldsymbol{\theta}}$, the MLE; (2) $\hat{\boldsymbol{\theta}}_L$, where we take $D\lambda$ to be the lower endpoint of its confidence interval from $\ell_p(D\lambda)$ and λ to be the lower endpoint of its confidence interval from $\ell_p(\lambda)$; (3) $\hat{\boldsymbol{\theta}}_U$, where we take $D\lambda$ to be the upper endpoint of its confidence interval from $\ell_p(D\lambda)$ and λ to be the upper endpoint of its confidence interval from $\ell_p(\lambda)$. Solving Equations (2.2)–(2.4) with these three combinations of $\boldsymbol{\theta}$ provides a simple way of giving a visual interpretation of the uncertainty in cell density as an approximation to the true uncertainty bounds.

2.3 Results

We now give the results from our experiments and from our likelihood analysis. Following these results, we conclude with a description of how we can use these results to predict future pore bridging experiments.

2.3.1 Experimental images

A subset of the results for the pore bridging experiments on the square geometry are shown in Figure 2.3 for days 7, 14, 25, and 28, where we see most of the pores take longer than 14 days to bridge, although there is some significant variability in this bridging time as we can even see some pores have completely bridged by day 14. In each pore, the growing tissue always forms a circle before bridging. In total, we have $n = 41$ imaged pores included in the dataset for the square, with 26 at day 7 and 15 at day 14.

The results we use for the pore bridging experiments on the wave geometry come only from day 14, and they are shown in Figure 2.4. Just as we saw in Figure 2.3, there is significant variability in the bridging time – while most pores appear to have bridged by day 14, some are still open, with a few being far from closed. In these wave pores, the void appears to initially close in as an oval before the void boundary eventually forms a circle. Our interest is in comparing the cell migration and cell proliferation rates between the pores of Figure 2.3 and Figure 2.4, but from these images it is not immediately clear whether these are similar or not. Section 2.3.2 shows results making this comparison using a mathematical model. In total, we only have $n = 3$ pore images for the wave geometry, all at day 14, since all other pores are closed and thus no other data is available for $y_{\text{c}}^{i,j}$ and $y_{\text{p}}^{i,j}$, or parts of the scaffolds were not imaged as in the leftmost pore in the first image of Figure 2.4. The pores used for the data are given in the first, fourth, and sixth images in Figure 2.4.

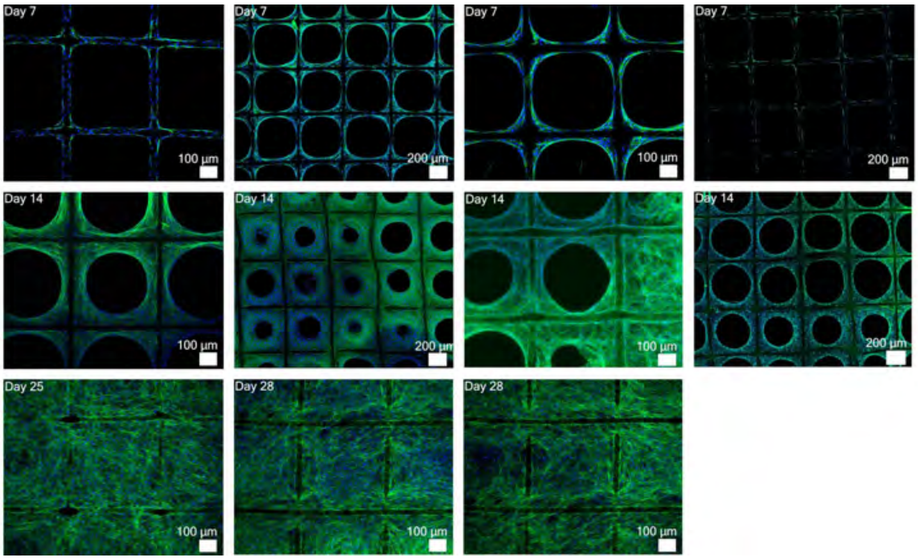

Figure 2.3: Experimental images for the square geometry. The images are composite fluoresence microscopy images of pore bridging experiments, with the blue channel showing the cell nuclei (stained with DAPI); the green channel showing the tissue and cytoskeleton (stained with phalloidin). Note that each image is from an independent experiment.

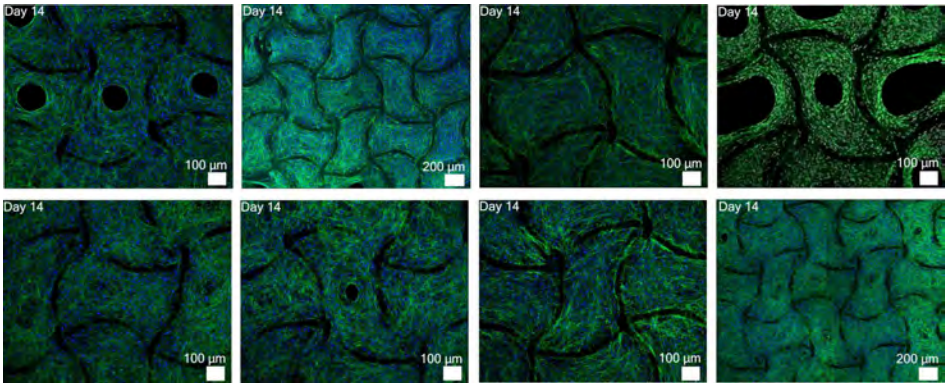

Figure 2.4: Experimental images for the wave geometry on day 14 of each experiment. The images are composite fluoresence microscopy images of pore bridging experiments, with the blue channel showing the cell nuclei (stained with DAPI); the green channel showing the tissue and cytoskeleton (stained with phalloidin). Note that each image is from an independent experiment.

2.3.2 Parameter estimation and parameter identifiability

We now consider the likelihoods, profile likelihoods, prediction intervals, and tissue growth predictions for the square and wave pores, demonstrating how well we can calibrate our model to the experimental data and make predictions between the two geometries.

Square pore

Figure 2.5(a) shows that the confidence regions for $(D\lambda, \lambda)$ for each u_0 have a similar shape, and the MLEs for $D\lambda$ are all around the same value. The boundary of each confidence region is well-defined in $D\lambda$ but not for λ, indicating that we are only able to obtain reliable estimates for $D\lambda$ but not for λ as we might anticipate [13, 30]. The confidence intervals we obtain from the profile log-likelihoods shown in Figure 2.5(b) for $D\lambda$ are approximately the same for each u_0, given approximately by $90\,\mu\mathrm{m}^2/\mathrm{day}^2 < D\lambda < 300\,\mu\mathrm{m}^2/\mathrm{day}^2$ for each u_0. The width of this interval is relatively small, noting that previously reported estimates of D in the literature vary across several orders of magnitude [24, 79]. The predictions for $\mu_c(t)$ on each geometry are shown in Figure 2.5(c)–(d). We see in (c) that we can recover the data on the square, with the prediction intervals capturing the average experimental data points, and the prediction intervals are indistinguishable for each u_0. The dashed lines show the predictions from

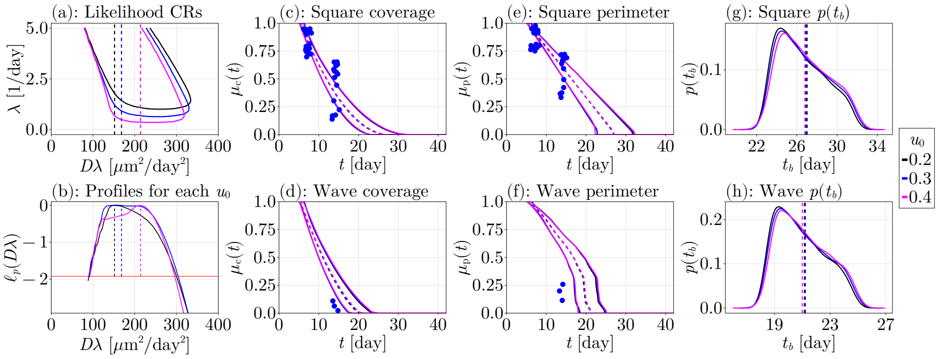

Figure 2.5: Likelihood analysis results for the square pore. In (a), the lines give the boundaries of the 95% confidence region for $\boldsymbol{\theta}$ for each u_0, and the vertical dashed lines show the MLE for $D\lambda$ (see Table 2.1). The profile log-likelihoods for $D\lambda$ for each u_0 are shown in (b), with the threshold $c^* \approx -1.92$ shown with a horizontal red line and the vertical dashed lines show the MLEs for $D\lambda$. In (c)–(f), predictions for $\mu_c(t)$ and $\mu_p(t)$ on each pore geometry are shown, with the blue dots showing the experimental data, the surrounding solid lines giving 95% prediction intervals for each u_0, and the dashed lines showing the corresponding estimates at the MLE $\hat{\boldsymbol{\theta}}$. The blue data points have been slightly jittered horizontally to help distinguish them. The estimates for the PDF $p(t_b)$ of the bridging time on each pore geometry are shown in (g)–(h). The results in (d), (f), and (h) are predictions on the wave geometry using parameters inferred from the square pore data.

the MLE $\hat{\boldsymbol{\theta}}$, indicating the most likely outcome of the experiments, and these curves too pass through the average of the experimental data points and are indistinguishable for each u_0. In contrast, we see for $\mu_c(t)$ on the wave geometry that we do not capture the precise values for $y_c^{i,j}$, although if we had more data points then we would likely capture more values due to their variability, noting that the far simpler shape, the square, has high variance. These results are also independent for u_0. The corresponding figures for $\mu_p(t)$ are shown in Figure 2.5(e)–(f), where we again capture the data on the square pore but not the data from the wave pore, and again the curves are all independent of u_0. Lastly, we show the probability distributions for the bridging time (2.9) on each pore geometry in Figure 2.5(g)–(h). These distributions have a similar shape for each u_0. The mode for t_b on the square pore appears to be around $t_b \approx 24$ days, and the distribution shows that we expect more pores to bridge between 23 to 30 days, consistent with our experiments in Figure 2.3. Similarly, the mode for t_b is around 20 days on the wave geometry, with most pores bridging within 19 to 24 days. The observation that the computed results in Figure 2.5, and in Figures 2.6–2.10, are all indistinguishable for each u_0 is an important result – it suggests that both the mean and variability in our estimates are relatively insensitive to u_0, indicating that precise measurements of u_0 are not critical.

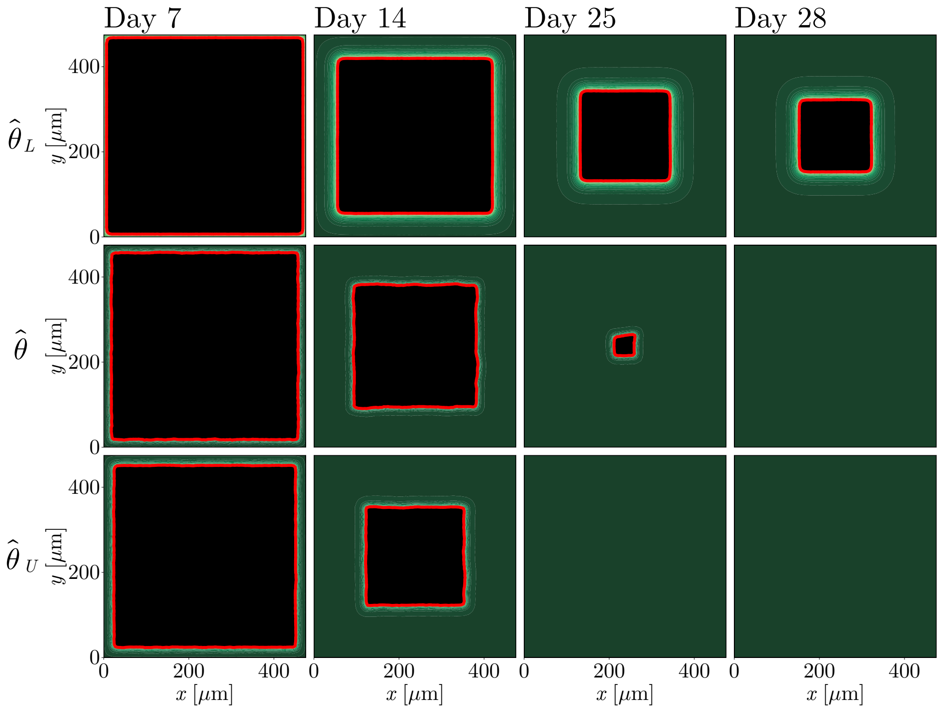

Figure 2.6: Model predictions for the variability in the tissue growth behaviour for the square geometry for $u_0 = 0.2$ at days 7, 14, 25, and 28. The parameters used are $\hat{\boldsymbol{\theta}}_L = (\hat{D}_L, \hat{\lambda}_L) = (73\,\mu\mathrm{m}^2/\mathrm{day}, 1.2\,\mathrm{day}^{-1})$, $\hat{\boldsymbol{\theta}} = (\hat{D}, \hat{\lambda}) = (32\,\mu\mathrm{m}^2/\mathrm{day}, 4.9\,\mathrm{day}^{-1})$, and $\hat{\boldsymbol{\theta}}_L = (\hat{D}_U, \hat{\lambda}_U) = (60\,\mu\mathrm{m}^2/\mathrm{day}, 5\,\mathrm{day}^{-1})$. The red boundary marks the position of the void boundary where $u(\mathbf{x}, t) = 1/2$. The top row of plots shows the predictions for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_L$, the middle row of plots shows the prediction for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, and the bottom row of plots show the predictions for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_U$; each individual column thus shows a rough confidence interval for the prediction at the respective time.

Figure 2.6 shows a summary of model predictions where we explore the variability in the behaviour of the experiments over time. In particular, we show numerical solutions of Equations (2.2)–(2.4) at $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\theta}}_L$, and $\hat{\boldsymbol{\theta}}_U$, each for $u_0 = 0.2$. In this figure, the columns show predictions for a given time, with each row corresponding to a different value for $\boldsymbol{\theta}$. The middle row corresponds to $\hat{\boldsymbol{\theta}}$, meaning the prediction that we expect to be most likely. The first two columns show model predictions for the two days that we use for calibrating the model, while the last two columns are genuine predictions since our mathematical model is not calibrated to data from these predictions. We see that there is a lot of variability in the position of the void boundary, especially at day 14, depending on the choice of $\hat{\boldsymbol{\theta}}_L$, $\hat{\boldsymbol{\theta}}$, or $\hat{\boldsymbol{\theta}}_U$, which approximately matches the observed variability in the experimental images in Figure 2.3. The numerical results on day 14 show that, at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_L$, the void boundary is still close to the pore boundary, but the bottom row shows that, at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_U$, the pore is half-way to being filled. The tissue boundaries do not round off as clearly as in Figure 2.3, although there is some rounding in the corners of these boundaries. This lack of rounding is a limitation of how we define the void boundary as the contour $u(\mathbf{x}, t) = 1/2$, with values higher than $1/2$ giving rounder boundaries.

The model predictions of the tissue growth in Figure 2.7 show predictions of how the wave pores will evolve over time for $u_0 = 0.2$, using parameter estimates obtained by calibrating Equations (2.2)–(2.4) to data from the square pores. Similar to what we noted in Figure 2.6, we do not see the same circular voids in Figure 2.7 as we do in the experimental images in Figure 2.4, though this is expected as the predictions average over many curves. Similarly, we see high variability in the results, with pores at day 14 ranging from being slightly closed to more than half-way closed. This variability is a positive result, matching the variability in the experimental images (Figure 2.4).

Together, these model predictions indicate that our model has been well-calibrated to the experimental data on the square pore, as we have captured the experimental data with our predictions and computed sensible probability distributions for the bridging time. The model predictions applied to the wave pore are reasonable, giving evidence of the similarities between the cell migration and cell proliferation mechanisms between the two geometries, although this is difficult to assess with few data points. Section 2.3.2 discusses the analogous results where we instead consider the model predictions from the experimental data on the wave.

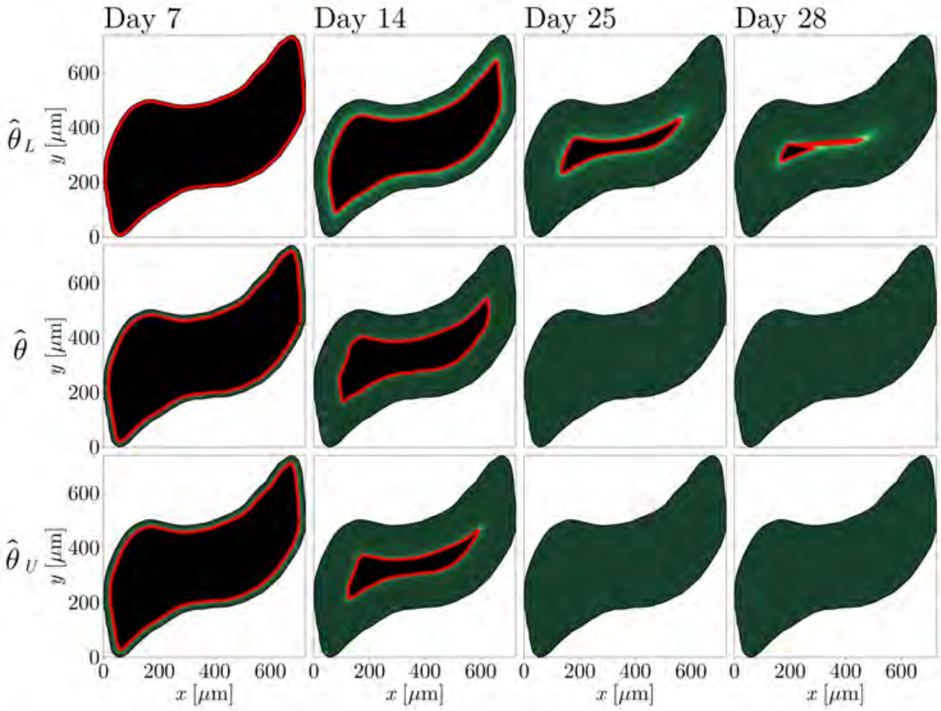

Figure 2.7: Model predictions for the variability in the tissue growth behaviour for the wave geometry using results from the square geometry for $u_0 = 0.2$ at days 7, 14, 25, and 28. The parameters used are $\hat{\boldsymbol{\theta}}_L = (\hat{D}_L, \hat{\lambda}_L) = (73\,\mu\text{m}^2/\text{day}, 1.2\,\text{day}^{-1})$, $\hat{\boldsymbol{\theta}} = (\hat{D}, \hat{\lambda}) = (32\,\mu\text{m}^2/\text{day}, 4.9\,\text{day}^{-1})$, and $\hat{\boldsymbol{\theta}}_L = (\hat{D}_U, \hat{\lambda}_U) = (60\,\mu\text{m}^2/\text{day}, 5\,\text{day}^{-1})$. The red boundary marks the position of the void boundary where $u(\mathbf{x}, t) = 1/2$. The top row of plots shows the predictions for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_L$, the middle row of plots shows the prediction for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, and the bottom row of plots show the predictions for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_U$; each individual column thus shows a rough confidence interval for the prediction at the respective time.

Wave-like pore

We now show the results obtained when we instead estimate $D\lambda$ from the data on the wave pore. The results have greater uncertainty here than in the square case since we have far fewer data points and only one day is covered by the data. The confidence regions in Figure 2.8(a) are much wider and flatter at the bottom than they were in the square case (Figure 2.5(a)), meaning the estimates for $D\lambda$ are less precise. The confidence region is not well-defined for larger values of $D\lambda$, indicating that we are unable to give any estimate for an upper bound on $D\lambda$. The corresponding profile log-likelihoods in Figure 2.8(b) for each u_0 do not intersect the

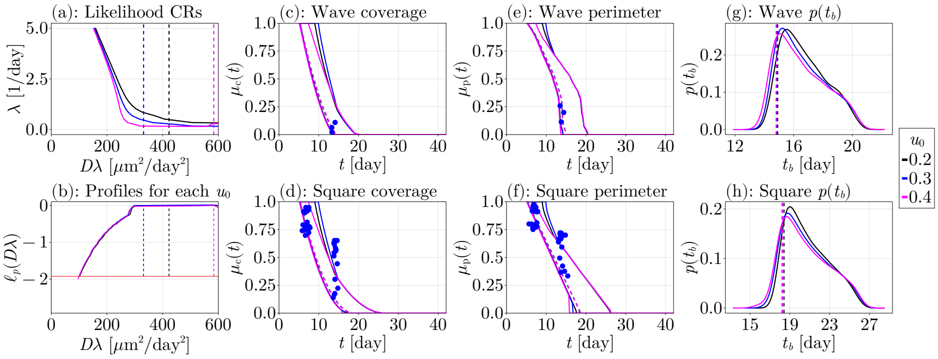

Figure 2.8: Likelihood analysis results for the wave pore. In (a), the lines give the boundaries of the 95% confidence region for $\boldsymbol{\theta}$ for each u_0, and the vertical dashed lines show the MLE for $D\lambda$ (see Table 2.1). The profile log-likelihoods for $D\lambda$ for each u_0 are shown in (b), with the threshold $c^* \approx -1.92$ shown with a horizontal red line and the vertical dashed lines show the MLEs for $D\lambda$. In (c)–(f), predictions for $\mu_\mathrm{c}(t)$ and $\mu_\mathrm{p}(t)$ on each pore geometry are shown, with the blue dots showing the experimental data, the surrounding solid lines giving 95% prediction intervals for each u_0, and the dashed lines showing the corresponding estimates at the MLE $\hat{\boldsymbol{\theta}}$. The blue data points have been slightly jittered horizontally to help distinguish them. The estimates for the PDF $p(t_b)$ of the bridging time on each pore geometry are shown in (g)–(h). The results in (d), (f), and (h) are predictions on the square geometry using parameters inferred from the wave pore data.

threshold $c^* \approx -1.92$, independently of u_0, and so we are unable to give any estimate for the upper limit of the confidence intervals for $D\lambda$, as was already suggested from Figure 2.8(a). Despite these difficulties, Figures 2.8(c)–(f) suggest that we are able to recover values for the experimental data $y_\mathrm{c}^{i,j}$ and $y_\mathrm{p}^{i,j}$ on each pore geometry, with the predictions from the MLEs going through the experimental data points on each geometry. The probability distribution for the bridging time for the wave pore is shown in Figure 2.8(g), where we see a mode for t_b around 14 days, with most pores predicted to close between 10 and 18 days, which is consistent with Figure 2.4. The corresponding probability distribution for the square geometry is shown in Figure 2.8(h), where we see that most pores are expected to close between 15 and 25 days, which is a shift from Figure 2.5(g) but is still consistent with the experimental images in Figure 2.3. It is important to emphasise that the recovery of these results on the square pore with so few data points is remarkable, as it (1) demonstrates the similarity between the migration and proliferation mechanisms on the two geometries, providing further evidence for the first research question, and (2) shows that we do not need such detailed, or even plentiful, data to recover these cellular mechanisms from another geometry.

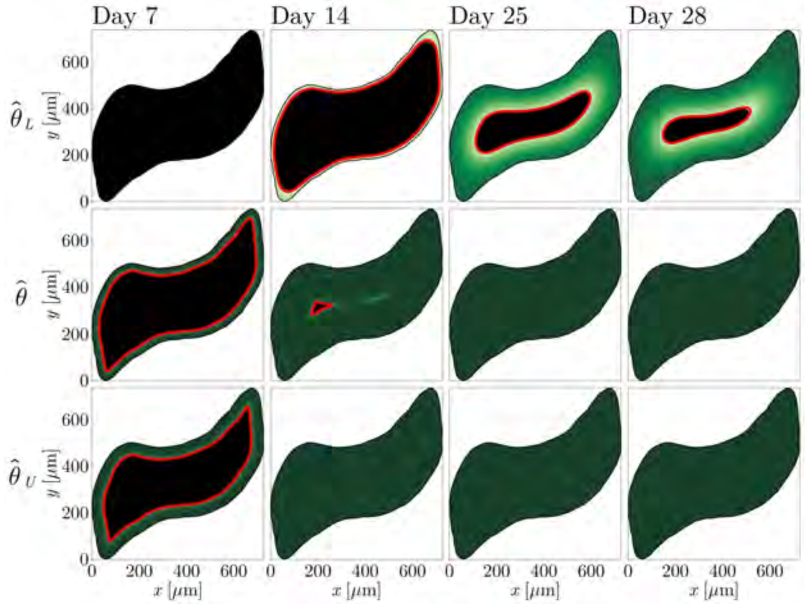

Figure 2.9: Model predictions for the variability in the tissue growth behaviour for the wave geometry for $u_0 = 0.2$ at days 5, 7, 14, 25, and 28. The parameters used are $\hat{\boldsymbol{\theta}}_L = (\hat{D}_L, \hat{\lambda}_L) = (581\,\mu\text{m}^2/\text{day}, 0.18\,\text{day}^{-1})$, $\hat{\boldsymbol{\theta}} = (\hat{D}, \hat{\lambda}) = (58\,\mu\text{m}^2/\text{day}, 7.29\,\text{day}^{-1})$, and $\hat{\boldsymbol{\theta}}_L = (\hat{D}_U, \hat{\lambda}_U) = (117\,\mu\text{m}^2/\text{day}, 10\,\text{day}^{-1})$. The red boundary marks the position of the void boundary where $u(\mathbf{x}, t) = 1/2$. The top row of plots shows the predictions for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_L$, the middle row of plots shows the prediction for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, and the bottom row of plots show the predictions for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_U$; each individual column thus shows a rough confidence interval for the prediction at the respective time.

Predictions of tissue growth for the wave geometry, based on experimental data on the wave geometry, are shown in Figure 2.9 for $u_0 = 0.2$. These model simulations are consistent with our experimental observations in Figure 2.4. The middle row, displaying the most likely outcome, shows that essentially all pores will be closed by 14, which matches Figure 2.4. The model simulations at day 14 show that while some pores may be completely open at this time, some may be closed or almost closed, and we expect this significant variability as we have so few data points.

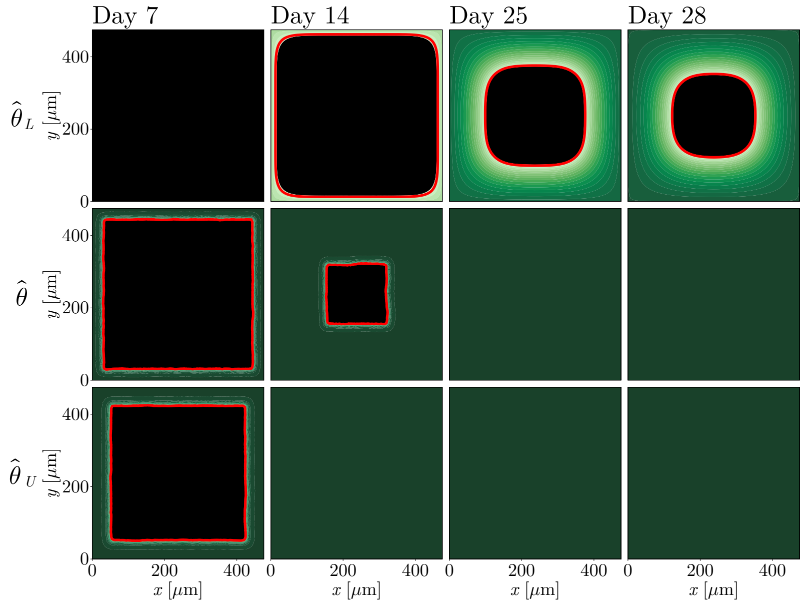

Figure 2.10: Model predictions for the variability in the tissue growth behaviour for the square geometry using results from the wave geometry for $u_0 = 0.2$ at days 7, 14, 25, and 28. The parameters used are $\hat{\boldsymbol{\theta}}_L = (\hat{D}_L, \hat{\lambda}_L) = (581\,\mu\text{m}^2/\text{day}, 0.18\,\text{day}^{-1})$, $\hat{\boldsymbol{\theta}} = (\hat{D}, \hat{\lambda}) = (58\,\mu\text{m}^2/\text{day}, 7.29\,\text{day}^{-1})$, and $\hat{\boldsymbol{\theta}}_L = (\hat{D}_U, \hat{\lambda}_U) = (117\,\mu\text{m}^2/\text{day}, 10\,\text{day}^{-1})$. The red boundary marks the position of the void boundary where $u(\mathbf{x}, t) = 1/2$. The top row of plots shows the predictions for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_L$, the middle row of plots shows the prediction for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, and the bottom row of plots show the predictions for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_U$; each individual column thus shows a rough confidence interval for the prediction at the respective time.

The predictions we make for the tissue growth on the square geometry using model results on the wave pore are given in Figure 2.10. We observe significant variations in the closing time. In particular, while there could be some pores that are only halfway bridged by day 25 or day 28, most are close to closing by day 14 and completely closed by day 25, consistent with Figure 2.3. The middle row shows that the most likely outcome, according to our model, is that the majority of pores will be halfway bridged at day 14 and closed by day 25, which again matches Figure 2.3.

Overall, these model predictions indicate that, despite only having three data points, we have been able to calibrate our mathematical model sufficiently well so that we capture the original experimental data on the wave from our model predictions and, most importantly, we can recover the experimental data from the square pore from our experimental data on the wave pores. Moreover, the estimated probability distributions for the bridging times on each geometry are a good match to the experimental images in Figures 2.3–2.4, as are the predictions of the tissue growth from the model simulations. Thus, not only have we demonstrated the practicality and utility of our method for obtaining these probability distributions, we have provided much stronger evidence than in Section 2.3.2 that the cellular mechanisms driving tissue growth are similar between the two geometries.

Tabulated comparison between the two geometries

Table 2.1 compares values of $D\lambda$ and t_b, for $u_0 = 0.2$, for the two geometries; the choice of $u_0 = 0.2$ is not significant as the model results are relatively insensitive to this choice. We see that, while we cannot estimate the upper limit of the confidence interval for $D\lambda$ using the wave geometry, the lower limits are similar between the two geometries, as are the MLEs. Note that while the MLEs differ by a factor of three, this is an insignificant amount when we note that estimates for D could vary by many orders of magnitude [24]. These values for $D\lambda$ provide strong evidence that the cellular mechanisms driving tissue growth on the two geometries are the same. The estimates of the bridging times for each geometry are not too dissimilar when using either the same geometry or predicting from the other geometry.

Table 2.1: Estimates for $D\lambda$ and t_b on each geometry for $u_0 = 0.2$. The 95% *CI* column gives the 95% confidence interval for the respective quantities, and the *MLE* column shows the corresponding MLEs. The second row gives predictions of the bridging time on the square geometry, while the third row is for the bridging time on the wave geometry.

	Square		Wave	
	MLE	95% CI	MLE	95% CI
$D\lambda$ [µm^2/day^2]	152	$(90, 299)$	423	$(107, -)$
t_b (square) [day]	27	$(24, 31)$	19	$(15, 22)$
t_b (wave) [day]	21	$(19, 24)$	15	$(13, 18)$

Overall, these model results support the hypothesis that the cellular mechanisms driving the tissue growth on each geometry are similar. Moreover, the ability to calculate reasonable estimates and probability distributions for the bridging time provides evidence that the results obtained from one geometry can be used to make predictions about tissue growth on another geometry using the available data.

2.3.3 Prediction of tissue growth on a hypothetical geometry

We now take the results on the square geometry and use them to make predictions on a new geometry that is yet to be experimentally tested. The purpose of this exercise is to demonstrate how our mathematical modelling tools could be used for making predictions on a new geometry from experimental results on a simpler geometry, such as about bridging times, without having to conduct any (potentially expensive and time-consuming) experiments. The geometry we consider is a cross-shaped pore. Using the same values for $D\lambda$ as were used in making the predictions in Figures 2.6 and 2.7, we produce the model predictions for the variability in the tissue growth in this new geometry in Figure 2.11. We see a similar variance in the previous predictions, namely most pores are closed by day 25 but at $\hat{\boldsymbol{\theta}}_L$ there are still some pores that remain half closed. The void maintains the symmetry of the geometry, forming a diamond shape during the early part of the growth process. These results are also largely independent of u_0, as we find when plotting these predictions for other u_0 values (not shown). We similarly show predictions for the $y_{\mathrm{c}}^{i,j}$ and $y_{\mathrm{p}}^{i,j}$ in Figure 2.12(a)–(b) and the hypothetical bridging time distribution $p(t_b)$ in Figure 2.12(c), all for $u_0 = 0.2$

An important feature of working with these predictions is that, once the likelihood results have been obtained, producing these predictions in Figures 2.11–2.12, or for any new geometry, is not a significantly time consuming task. The snapshots in Figure 2.11 take only around a minute to compute and visualise, and the model predictions in Figure 2.12 may take around 10 minutes to an hour, depending on the number of samples requested for the prediction intervals. Thus, this type of exploratory analysis of a new geometry can be efficiently performed in a reasonable time.

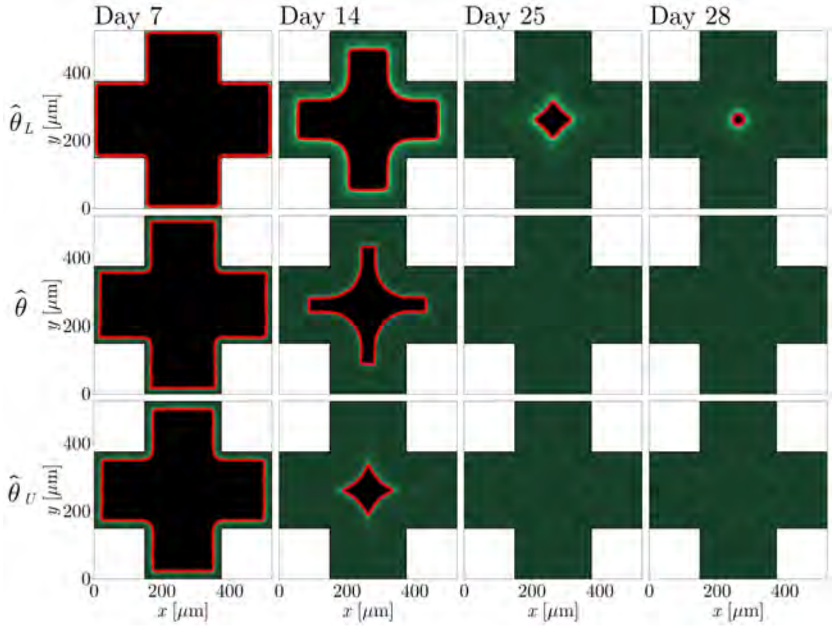

Figure 2.11: Model predictions for the variability in the tissue growth behaviour for the hypothetical geometry using results from the square geometry for $u_0 = 0.2$ at days 7, 14, 25, and 28. The parameters used are $\hat{\boldsymbol{\theta}}_L = (\hat{D}_L, \hat{\lambda}_L) = (73\,\mu\text{m}^2/\text{day}, 1.2\,\text{day}^{-1})$, $\hat{\boldsymbol{\theta}} = (\hat{D}, \hat{\lambda}) = (32\,\mu\text{m}^2/\text{day}, 4.9\,\text{day}^{-1})$, and $\hat{\boldsymbol{\theta}}_L = (\hat{D}_U, \hat{\lambda}_U) = (60\,\mu\text{m}^2/\text{day}, 5\,\text{day}^{-1})$. The red boundary marks the position of the void boundary where $u(\mathbf{x}, t) = 1/2$. The top row of plots shows the predictions for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_L$, the middle row of plots shows the prediction for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, and the bottom row of plots show the predictions for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_U$; each individual column thus shows a rough confidence interval for the prediction at the respective time.

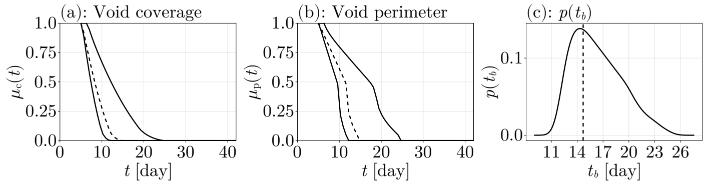

Figure 2.12: Predictions for the summary statistics and bridging distributions on the hypothetical geometry using results from the square geometry in Figure 2.5 at $u_0 = 0.2$. In (a)–(b), the dashed lines show the estimated curve corresponding to the maximum likelihood estimate for $(D\lambda, \lambda)$ from the square pore, and the curves surrounding it define a 95% uncertainty band for the curve.

2.4 Discussion

New experimental images and modelling predictions in Section 2.3 provide answers to the three research questions posed in Section 2.1. The first question, namely whether the cellular mechanisms driving tissue growth are independent of pore shape, appears to be true for the geometries considered. The second question asks whether we can make predictions of bridging times, with uncertainty, on a geometry from a separate geometry. We have found that we can produce reliable predictions with uncertainty between separate geometries, both in the form of probability distributions and prediction intervals. The third question concerns the type and quantity of data required for making predictions between geometries. We found that the data, and the amount of data, we use for summarising the images is sufficient for making predictions of tissue growth with uncertainty, namely information about the tissue void — even with only three data points on the wave geometry. Interestingly, the answers to these questions require only very simple measurements of the experiments, rather than performing cell counting [80]. This observation agrees with previous work that has compared methods using leading edge detection and cell counting, demonstrating that tracking the leading edge is sufficient for estimating the cell migration and cell proliferation rates [80]. We show in Appendix 2.F that if we considered only void area for the analysis of the square geometry then we would obtain the same conclusions, but both area and perimeter are necessary for the wave geometry to answer the research questions. The fact that both area and perimeter are required in general can be expected since area and perimeter together give a detailed description of a shape, but not separately.

These answers have important implications. Firstly, we have demonstrated the ability to extrapolate from experimental results on one geometry to another geometry, making predictions with uncertainty. This facilitates fast and inexpensive pilot studies to be performed for new pore geometries without conducting the experiments or even fabricating the scaffolds, as with our exploratory analysis in Section 2.3.3. Numerical simulations could be performed in a few minutes of computation on a standard desktop computer, while conducting the necessary experiments will require more than one month for tissue growth and a considerable amount of effort and expense to fabricate the scaffolds with melt electrowriting. We do not mean to imply that these predictions can replace experimental verification, instead we view this suite of predictive tools as complementary screening tools that can be used to plan and interpret experiments efficiently. Secondly, the novel method we present for obtaining probability distributions for the bridging time provides a useful tool for meeting certain needs by helping us to understand the amount of time required for a tissue to form and bridge the pore, and also for understanding how long tissue growth needs to be incubated for in tissue engineering constructs. Together with the type of exploratory analysis demonstrated in Section 2.3.3, these probability distributions can help facilitate the construction of a geometry that is likely to bridge within some time window for certain clinical needs, and for determining how long an experiment should be run for.

The mathematical model we use in this study is relative simple as it involves just three parameters: D, λ, and u_0. A model that better incorporates other effects such as cell adhesion or the different phases of tissue growth, in particular the initial phase where cells move off the scaffold or the later phase where the pore is closing and cells overlap [26], could be of interest to provide more biological insight. A key limitation of working with a more detailed mathematical model, however, would be the need to collect significantly more data so that the necessary additional parameters can be properly estimated [28, 76].

2.5 Conclusion and future work

In this study, we use a reaction-diffusion model together with a likelihood-based uncertainty quantification framework to study how pore geometry affects tissue growth, particularly in how we can make inference about tissue growth on complicated pore geometries using data from tissue growth on simpler square geometries, providing new tools for studying tissue growth with uncertainty and providing probability distributions for bridging times. We use data from pore bridging experiments to perform this analysis, considering a square geometry and a wave-like geometry.

Our combined experimental and mathematical modelling results suggest that the cell migration and cell proliferation mechanisms driving tissue growth appear to be independent of the pore geometry, giving evidence that observed curvature effects are due to space availability rather than cellular mechanisms. We can make predictions of the bridging time on a new geometry in the form of a probability density function, a powerful tool for understanding both quantitatively and qualitatively what may happen in a pore bridging experiment on a new geometry, including the estimation of probabilities of bridging times over a given time interval.

There are several avenues for future work based on our findings in this study. First, our computational tools can be applied to new pore bridging experiments involving different geometries or different cell lines since our methods are independent of these two features. Secondly, it would be of interest to collect more data across more time points to explore the extent to which additional parameters, such as D and λ separately, can be estimated [24]. If future works consider more than one variable, we note that it would not be feasible to work with the plots of the log-likelihood function as we have done, and instead the profile log-likelihood would be required to obtain uncertainty quantification. Thirdly, the ability to make predictions on new geometries can facilitate a systematic study of how bridging times depend on curvature, such as by defining a geometry that depends directly on a specified curvature and comparing the probability distributions over many curvatures. This would provide a plausible set of outcomes to be analysed prior to running full-scale experiments exploring these features. Lastly, it would be worthwhile to consider more complex quantities for summarising the experimental images for inclusion in the likelihood function (2.7), in particular quantities that capture the complicated nature of the void boundaries better than coverage and perimeter. These quantities could help with the calibration of our models, leading to void boundaries that better match those in the experimental images.

The predictions made on geometries from data on a separate geometry can be useful for facilitating a pilot study for pore bridging experiments on the geometry, such as the geometry demonstrated in Figure 2.11 and Figure 2.12. It would be of interest to see how well these predictions can help with preparing and investing into future experiments, for example in estimating what time scales an experiment may need to be run for by assessing the uncertainty in the bridging times. All code and data to reproduce this work are available on GitHub at https://github.com/DanielVandH/PoreBridging.jl in the Julia language [81].

Chapter 2: Supplementary material

2.A Summarising images from the experiments

Here we present the formulae for computing the area and perimeter summary statistics. As discussed in Section 2.2, for an image $I(t)$ at time t we identify a boundary $\mathcal{P}^I(t)$ for the pore and a boundary $\mathcal{V}^I(t)$ for the void. Precisely, we identify the sets $\mathcal{P}^I(t) = \{\mathbf{p}_1^I(t), \ldots, \mathbf{p}_n^I(t), \mathbf{p}_{n+1}^I(t)\}$ and $\mathcal{V}^I(t) = \{\mathbf{v}_1^I(t), \ldots, \mathbf{v}_m^I(t), \mathbf{v}_{m+1}^I(t)\}$, where $\mathbf{p}_{n+1}^I(t) = \mathbf{p}_1^I(t)$ and $\mathbf{v}_{m+1}^I(t) = \mathbf{v}_1^I(t)$ and the boundary points are arranged in counter-clockwise order. An example of these sets is shown in Figure 2.A.1. Using these sets, the area of the pore and the area of the void for this image $I(t)$, denoted $A[\mathcal{P}^I(t)]$ and $A[\mathcal{V}^I(t)]$, respectively, can be computed [82]

$$\begin{aligned} A[\mathcal{P}^I(t)] &= \frac{1}{2}\sum_{i=1}^{n} \det\bigl(\mathbf{p}_i^I(t), \mathbf{p}_{i+1}^I(t)\bigr), \\ A[\mathcal{V}^I(t)] &= \frac{1}{2}\sum_{i=1}^{m} \det\bigl(\mathbf{v}_i^I(t), \mathbf{v}_{i+1}^I(t)\bigr). \end{aligned} \tag{2.10}$$

Similarly, the perimeters $\ell[\mathcal{P}^I(t)]$ and $\ell[\mathcal{V}^I(t)]$ for the pore and void boundaries, respectively, are simply

$$\begin{aligned} \ell[\mathcal{P}^I(t)] &= \sum_{i=1}^{n} \left\|\mathbf{p}_{i+1}^I(t) - \mathbf{p}_i^I(t)\right\|, \\ \ell[\mathcal{V}^I(t)] &= \sum_{i=1}^{n} \left\|\mathbf{v}_{i+1}^I(t) - \mathbf{v}_i^I(t)\right\|, \end{aligned} \tag{2.11}$$

summing up each length like the one annotated in Figure 2.A.1.

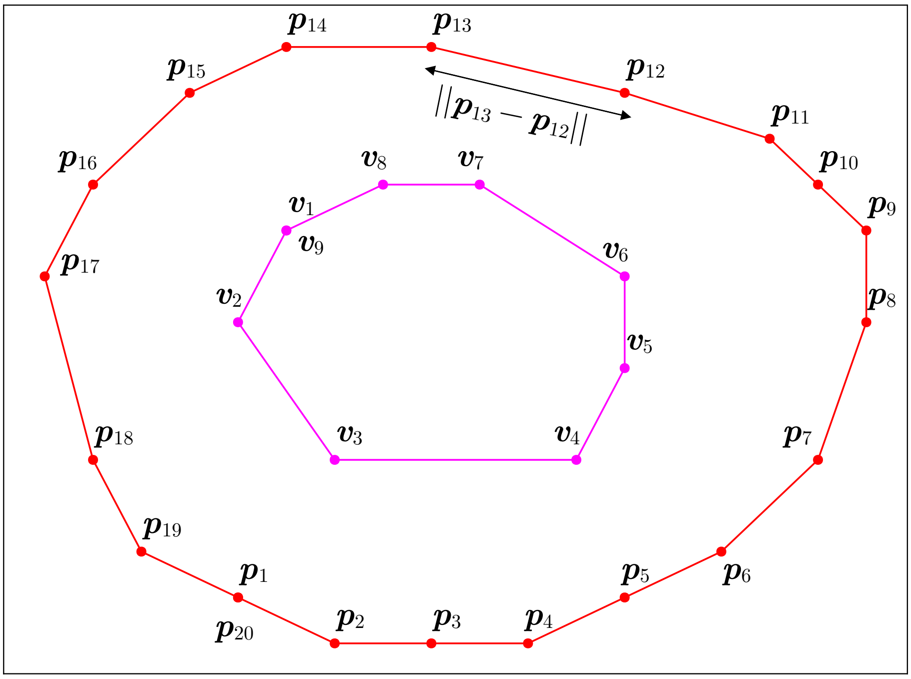

Figure 2.A.1: Example of the configuration of the point sets $\mathcal{P}$ and $\mathcal{V}$, omitting the superscript I. These are $n = 19$ points on the pore boundary and $m = 8$ on the void boundary, with an extra point at the end of each set to close the boundary. Note also the order of the points.

2.B Finite volume method

In this section, we give the details for how we solve the PDE

$$\frac{\partial u(\mathbf{x},t)}{\partial t} = D\boldsymbol{\nabla}\cdot\left[u(\mathbf{x},t)\boldsymbol{\nabla}u(\mathbf{x},t)\right] + \lambda u(\mathbf{x},t)\left[1-u(\mathbf{x},t)\right], \quad \mathbf{x}\in\Omega, \tag{2.12}$$

$$\frac{\mathrm{d}u(\mathbf{x},t)}{\mathrm{d}t} = \lambda u(\mathbf{x},t)\left[1-u(\mathbf{x},t)\right], \quad \mathbf{x}\in\partial\Omega, \tag{2.13}$$

$$u(\mathbf{x},t_a) = \begin{cases} u_0 & \mathbf{x}\in\partial\Omega, \\ 0 & \mathbf{x}\in\Omega, \end{cases} \tag{2.14}$$

using the finite volume method [83]. The first step is to compute a triangulation of the domain Ω, denoted $\mathcal{T}(\Omega)$, which we accomplish using `DelaunayTriangulation.jl` [84]. For some interior point $\mathbf{x}_i = (x_i, y_i)^{\mathsf{T}} \in \mathcal{T}(\Omega)$, we take the centroids of the triangles neighbouring $\mathbf{x}_i$ and connect these centroids to the midpoints of the associated triangle, giving a closed polygon that we denote by $\partial\Omega_i$ and show in Figure 2.B.1. The interior of this polygon is denoted Ω_i, which we call a control volume, and has some volume V_i. This polygon is defined by a set of edges $\mathcal{E}_i$, and for each $\mathbf{x}_\sigma \in \mathcal{E}_i$ there is an associated length L_σ, midpoint $\mathbf{x}_\sigma$, and unit normal $\hat{\mathbf{n}}_{i,\sigma}$ which is normal to σ and directed outwards to Ω_i with unit length. It is with these control volumes that we can now discretise (2.12).

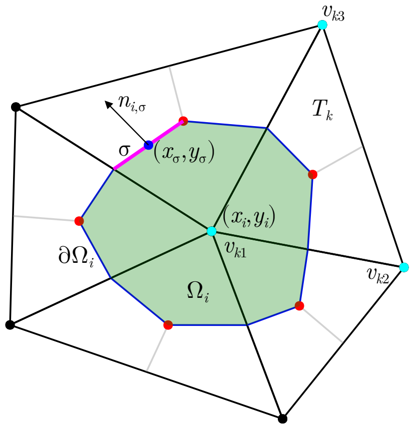

Figure 2.B.1: Example of a control volume around a point $\mathbf{x}_i = (x_i, y_i)^{\mathsf{T}}$. The control volume is the region in green, and its boundary $\partial\Omega_i$ is shown in blue. The edge $\sigma \in \mathcal{E}_i$ is shown in magenta. Lastly, the cyan points show an example counter-clockwise ordering (v_{k1}, v_{k2}, v_{k3}) of a triangle $T_k \in \mathcal{T}(\Omega)$.

We integrate (2.12) over Ω_i,

$$\frac{\mathrm{d}}{\mathrm{d}t}\iint_{\Omega_i} u(\mathbf{x},t)\,\mathrm{d}A = D\iint_{\Omega_i} \boldsymbol{\nabla}\cdot[u(\mathbf{x},t)\boldsymbol{\nabla}u(\mathbf{x},t)]\,\mathrm{d}A + \lambda\iint_{\Omega_i} u(\mathbf{x},t)\,(1-u(\mathbf{x},t))\,\mathrm{d}A\,. \tag{2.15}$$

The first integral on the right of (2.15) can be re-written as a line integral using the divergence theorem, and then re-written as a sum by integrating across each edge of $\partial\Omega_i$:

$$\begin{aligned} D\iint_{\Omega_i} \boldsymbol{\nabla}\cdot[u(\mathbf{x},t)\boldsymbol{\nabla}u(\mathbf{x},t)]\,\mathrm{d}A &= D\oint_{\partial\Omega_i}[u(\mathbf{x},t)\boldsymbol{\nabla}u(\mathbf{x},t)]\cdot\hat{\mathbf{n}}_{i,\sigma}(\mathbf{x},t)\,\mathrm{d}s \\ &= D\sum_{\sigma\in\mathcal{E}_i}\int_\sigma [u(\mathbf{x},t)\boldsymbol{\nabla}u(\mathbf{x},t)]\cdot\hat{\mathbf{n}}_{i,\sigma}\,\mathrm{d}s\,, \end{aligned} \tag{2.16}$$

with $\hat{\mathbf{n}}_i(\mathbf{x},t)$ the unit normal vector field on $\partial\Omega_i$. Next, defining the control volume averages

$$\bar{u}_i = \frac{1}{V_i}\iint_{\Omega_i} u(\mathbf{x},t)\,\mathrm{d}A \quad \text{and} \quad \bar{R}_i = \frac{\lambda}{V_i}\iint_{\Omega_i} u(\mathbf{x},t)\,(1-u(\mathbf{x},t))\,\mathrm{d}A\,,$$

our integral formulation (2.15) becomes

$$\frac{\mathrm{d}\bar{u}_i}{\mathrm{d}t} = \frac{D}{V_i}\sum_{\sigma\in\mathcal{E}_i}\int_\sigma [u(\mathbf{x},t)\boldsymbol{\nabla}u(\mathbf{x},t)]\cdot\hat{\mathbf{n}}_{i,\sigma}\,\mathrm{d}s + \bar{R}_i. \tag{2.17}$$

To now approximate the integral in (2.17), we take $\bar{u}_i \approx u(\mathbf{x}_i,t)$, $\bar{R}_i \approx \lambda u(\mathbf{x}_i,t)[1-u(\mathbf{x}_i,t)]$, and use the midpoint rule:

$$\int_\sigma [u(\mathbf{x},t)\boldsymbol{\nabla}u(\mathbf{x},t)]\cdot\hat{\mathbf{n}}_{i,\sigma}\,\mathrm{d}s \approx \left\{[u(\mathbf{x}_\sigma,t)\boldsymbol{\nabla}u(\mathbf{x}_\sigma,t)]\cdot\hat{\mathbf{n}}_{i,\sigma}\right\}L_\sigma.$$

To approximate $\boldsymbol{\nabla}u(\mathbf{x}_\sigma,t)$, we let $\mathcal{T}_i$ be the set of triangles in $\mathcal{T}(\Omega)$ that have $\mathbf{x}_i$ as a node, and take a triangle $T_k\in\mathcal{T}_i$. Linearly interpolating u over the element T_k,

$$u(\mathbf{x},t) = \alpha_k(t)x + \beta_k(t)y + \gamma_k(t), \quad (x,y)\in T_k, \tag{2.18}$$

where the coefficients come from the values of u at each vertex of T_k, gives $\boldsymbol{\nabla}u(\mathbf{x},t) = (\alpha_k(t),\beta_k(t))^{\mathsf{T}}$ inside T_k. Thus, our approximation becomes, for each time step,

$$\frac{\mathrm{d}u_i}{\mathrm{d}t} = \frac{D}{V_i} \sum_{\sigma \in \mathcal{E}_i} \left\{ \left[\left(\alpha_{k(\sigma)}(t) x_\sigma + \beta_{k(\sigma)}(t) y_\sigma + \gamma_{k(\sigma)}(t) \right) \right. \right.$$
$$\left. \left. \left(\alpha_{k(\sigma)}(t), \beta_{k(\sigma)}(t) \right)^{\mathsf{T}} \right] \cdot \hat{\mathbf{n}}_{i,\sigma} \right\} L_\sigma + \lambda u_i \left(1 - u_i\right), \tag{2.19}$$

where $u_i = u(\mathbf{x}_i, t)$ and the $k(\sigma)$ notation is used to refer to the edge σ inside the triangle $T_{k(\sigma)}$.

To complete the approximation, the boundary condition (2.13) is given by $\mathrm{d}u_i/\mathrm{d}t = \lambda u_i(1 - u_i)$. Thus, our discretisation is given by (2.19) in the interior, i.e. the regions bounded by the red curves in Figure 2, while on the red curve we have $\mathrm{d}u_i/\mathrm{d}t = \lambda u_i(1 - u_i)$. The initial condition for this system of ODEs comes from (2.14), letting $u_i = u_0$ on the boundary and $u_i = 0$ in the interior at the initial time. We solve the system of ordinary differential equations using `DifferentialEquations.jl` [85] with the `TRBDF2` algorithm and the `KLUFactorization` linear solver [86, 87] together with the package `FiniteVolumeMethod.jl` [88] that computes the equations.

To assess the accuracy of our implementation of the finite volume, we applied several test cases, including setting up a domain to compare with one-dimensional travelling waves and comparisons with exact solutions. Moreover, we ensured that the size of the mesh used was sufficient by checking that increasing the number of mesh elements did not change the quality of the solution. Tests for the implementation itself are examined clearly in the documentation of the `FiniteVolumeMethod.jl` package [88].

2.C Computing summary statistics from model realisations

In this appendix, we consider the problem of computing the summary statistics $\mu_{\mathrm{c}}(t)$ and $\mu_{\mathrm{p}}(t)$ as defined in the text. We let $C_\tau(t) = \{\mathbf{x} \in \Omega : u(\mathbf{x}, t) = \tau\}$; note that $\tau = 1/2$ in the text. The objective is to compute the area and perimeter of $C_\tau(t)$, together with a polygonal representation of $C_\tau(t)$ assuming $C_\tau(t)$ is simply connected or $C_\tau(t) = \emptyset$. In what follows, we instead compute the area of the region where $u(\mathbf{x}, t) > \tau$, i.e. $A(\Omega \setminus C_\tau(t)) = A(\Omega) - A(C_\tau(t))$, where $A(\Omega)$ is the area of Ω and $A(C_\tau(t))$ the area of $C_\tau(t)$. The area of $C_\tau(t)$ is then obtained by simply computing $A(C_\tau(t)) = A(\Omega) - A(\Omega \setminus C_\tau(t))$.

Let us take our triangular mesh $\mathcal{T}(\Omega)$ of our domain, and consider some triangle $T(\Omega)$ with vertices $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k$ and associated solution values at time t given by $u_i = u(\mathbf{x}_i, t)$, $u_j = u(\mathbf{x}_j, t)$, and $u_k = u(\mathbf{x}_k, t)$. The finite volume method allows us to represent $u(\mathbf{x}, t)$ with a linear interpolant inside T, giving

$$u(\mathbf{x}, t) = \alpha x + \beta y + \gamma, \quad (x, y) \in T,$$

where the coefficients (α, β, γ) depend on t; these coefficients are defined in Appendix 2.B. This linearity then implies that, to find intersections of u with the plane $u = \tau$ inside T, we need only consider intersections with the edges. We denote the edge connecting u_i to u_j by $\overrightarrow{u_iu_j}$, and the edge connecting $\mathbf{x}_i$ to $\mathbf{x}_j$ by $\overrightarrow{\mathbf{x}_i\mathbf{x}_j}$. This edge $\overrightarrow{\mathbf{x}_i\mathbf{x}_j}$ is parametrised by $\mathbf{x}(s) = \mathbf{x}_i + (\mathbf{x}_j - \mathbf{x}_i)s$, $0 \leq s \leq 1$. With this parametrisation, we see that, if an intersection does exist on $\overrightarrow{u_iu_j}$, it occurs when $s^* = (\tau - u_i)/(u_j - u_i)$, in particular at $\mathbf{x}(s^*) = \mathbf{x}_i + (\mathbf{x}_j - \mathbf{x}_i)(\tau - u_i)/(u_j - u_i)$.

By considering the eight possible values of the u_i, u_j, u_k relative to $u = \tau$, we can easily determine whether an intersection exists. These possibilities are shown in Table 2.C.1, which show that we can check each possibility and compute the area accordingly. All the cases in Table 2.C.1, except for the first and last cases, imply that there is a line going through T where $u = \tau$, and the length of this line can be easily computed by simply taking the magnitude of the difference of the intersections on the two associated edges. Moreover, if we are interested in getting a representation of the leading edge itself for plotting, we can simply store all these intersection points which we can then sort counter-clockwise and clear duplicated

Table 2.C.1: Possible configurations of the nodal values relative to the threshold τ. In the first three columns, the symbol refers to u_i's value relative to τ. For example, a $<$ in the u_j column means $u_j < \tau$. In the intersection columns, $\overrightarrow{u_i u_j}$ is the edge from u_i to u_j, and the text refers to whether the plane $u(\mathbf{x}, t) = \alpha x + \beta y + \gamma$ can intersect with the plane defined by the plane $u = \tau$, with "N" meaning no intersection and "Y" meaning there is an intersection. The notation $A(u_i, u_j, u_k)$ means the area formed by these points projected onto the plane, and a point u_{ij} denotes the intersection point on the edge connecting u_i and u_j.

Nodal values			Intersection?			Area contribution
u_i	u_j	u_k	$\overrightarrow{u_i u_j}$	$\overrightarrow{u_j u_k}$	$\overrightarrow{u_k u_i}$	
$<$	$<$	$<$	N	N	N	0
$<$	$<$	$>$	N	Y	Y	$A(u_{ki}, u_{jk}, u_k)$
$<$	$>$	$<$	Y	Y	N	$A(u_{ij}, u_j, u_{jk})$
$<$	$>$	$>$	Y	N	Y	$A(u_i, u_j, u_k) - A(u_{ij}, u_{ki}, u_i)$
$>$	$<$	$<$	Y	N	Y	$A(u_{ki}, u_i, u_{ij})$
$>$	$<$	$>$	Y	Y	N	$A(u_i, u_j, u_k) - A(u_{jk}, u_{ij}, u_j)$
$>$	$>$	$<$	N	Y	Y	$A(u_i, u_j, u_k) - A(u_{jk}, u_k, u_{ki})$
$>$	$>$	$>$	N	N	N	$A(u_i, u_j, u_k)$

intersections, giving a vector of points that can be plotted. For the wave geometry considered in the text, sorting the leading edge is not as simple and so we instead plot the concave hull of these points, computed with the `ConcaveHull.jl` package [89, 90].

Now that we understand how to compute the area of the part of a triangle that is above the plane $u = \tau$, which we denote by A_T, the total area where $u(\mathbf{x}, t) \geq \tau$ is given by $\sum_{T \in \mathcal{T}(\Omega)} A_T$, meaning $A(C_\tau(t)) = A(\Omega) - \sum_{T \in \mathcal{T}(\Omega)} A_T$. Thus, normalising by $A(\Omega)$, we have

$$\mu_{\mathrm{c}}(t) = 1 - \frac{1}{A(\Omega)} \sum_{T \in \mathcal{T}(\Omega)} A_T. \tag{2.20}$$

Similarly, letting ℓ_T be the length of the line in T where $u = \tau$, which is zero if there is no such line, the perimeter of $C_\tau(t)$ is $\ell(C_\tau(t)) = \sum_{T \in \mathcal{T}(\Omega)} \ell_T$, giving $\mu_{\mathrm{p}}(t) = [1/\ell(\partial\Omega)] \sum_{T \in \mathcal{T}(\Omega)} \ell_T$. We note that it is possible to have $\sum_{T \in \mathcal{T}(\Omega)} \ell_T = 0$, which means that there is no part of $u(\mathbf{x}, t)$ where $u > \tau$, but this means the whole of Ω is the void, i.e. $u < \tau$ in all of Ω. Thus, the correct definition is

$$\mu_{\mathrm{p}}(t) = \begin{cases} \dfrac{1}{\ell(\partial\Omega)} \displaystyle\sum_{T\in\mathcal{T}(\Omega)} \ell_T & \displaystyle\sum_{T\in\mathcal{T}(\Omega)} \ell_T \neq 0, \\ 1 & \displaystyle\sum_{T\in\mathcal{T}(\Omega)} \ell_T = 1. \end{cases} \tag{2.21}$$

As described in the manuscript, these quantities μ_S are used to model the distribution that our data $y_S^{i,j}$ are realisations of, in particular $Y_S^i \sim \mathcal{N}(\mu_S(t_i;\boldsymbol{\theta}), \sigma_S^2)$. Here we give the formula used for σ_S^2. We simply aggregate all the data for the quantity S into a single set, giving the sample standard deviation

$$\sigma_S^2 = \frac{1}{n_1+n_2-1} \sum_{i=1}^{2} \sum_j \left(y_S^{i,j} - \overline{y}_S^{i,j} \right)^2,$$

where n_i is the total number of data points at $t = t_i$ for $i = 1, 2$, and

$$\overline{y}_S^{i,j} = \frac{1}{n_1+n_2-1} \sum_{i=1}^{2} \sum_j y_S^{i,j}$$

is the aggregated mean of the $y_S^{i,j}$; the second sum in each term denotes a sum over all pore indices j. For the square geometry, $n_1 + n_2 = 41$, and for the wave geometry we have $n_1 + n_2 = 3$.

2.D Reparametrisation of the likelihood function

Here we discuss issues with working with D and λ separately in the likelihood function. For this discussion, we will take $u_0 = 0.2$, but note that the results are the same for any other u_0. To start, let us take our log-likelihood function $\ell(\boldsymbol{\theta} \mid \mathbf{y})$ with $\boldsymbol{\theta} = (D, \lambda)$. We evaluate this log-likelihood over a grid of points, obtaining the surface shown in Figure 2.D.1(a). We see that the log-likelihood in this case is banana-shaped, indicating that D and λ are related [28], and so we would expect problems when trying to compute univariate confidence intervals from the profile log-likelihoods. One way to overcome this issue is to reparametrise in terms of $(D\lambda, \lambda)$, motivated by noting that our likelihood function uses data based on the void boundary which is known to have a speed that depends directly on the product $D\lambda$ [30]. The surface we obtain under this reparametrisation is given in Figure 2.D.1(b), where we see that we can now assign a finite interval to $D\lambda$, meaning we will be able to obtain confidence intervals from the profile likelihood for $D\lambda$, but we can still not assign any upper bound to λ — λ is not identifiable. This latter issue with λ is not important for us, though, as we only need $D\lambda$ to describe the cellular mechanisms driving tissue growth in our experiments.

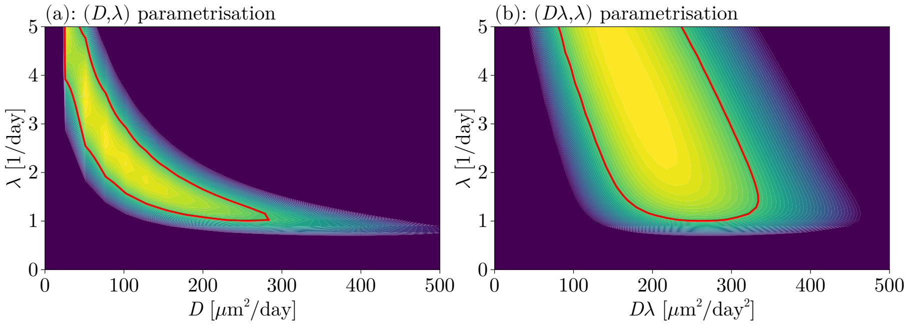

Figure 2.D.1: Log-likelihood surfaces for $\ell(\boldsymbol{\theta} \mid \mathbf{y})$ using (a) the (D, λ) parametrisation and (b) the $(D\lambda, \lambda)$ parametrisation. The red curves show the 95% confidence region for $\boldsymbol{\theta}$.

We note that, at first glance, it might appear that the surface in Figure 2.D.1(b) could eventually stop on the vertical axis for larger values of λ. We have computed this surface up to $\lambda = 25\,\text{day}^{-1}$ previously and find that this is not the case. Moreover, we note that even the maximum value $\lambda = 5\,\text{day}^{-1}$ shown in Figure 2.D.1(b) is large, as the proliferation time for these cells is typically between half a day to two days, corresponding to a value of λ between $0.5\,\mathrm{day}^{-1}$ and $2\,\mathrm{day}^{-1}$ [13]. Thus, even this value of $\lambda = 5\,\mathrm{day}^{-1}$ is a conservative upper bound, and certainly λ is not identifiable.

2.E Profile likelihood analysis

In this section, we will outline (1) how we compute the MLE, (2) how profile likelihoods are computed, and (3) how prediction intervals are computed.

2.E.1 Computing the MLE

The MLE is obtained by maximising $\ell(\boldsymbol{\theta} \mid \mathbf{y})$ over a certain rectangle defined by the bounds on $D\lambda$ and λ. We maximise $\ell(\boldsymbol{\theta} \mid \mathbf{y})$ using `NLopt.jl` with the derivative-free algorithm `LN_BOBYQA` [91–93]. To construct an initial estimate for the optimiser, we evaluate the log-likelihood on a 40×40 grid, taking 40 values for $D\lambda$ in $0 < D\lambda < 500$ and 40 values for λ in $0 < \lambda < 5$. We then take the pair $(D\lambda, \lambda)$ in this grid that gives the greatest value for $\ell(\boldsymbol{\theta} \mid \mathbf{y})$, and this then gives the initial estimate we use for the optimiser.

2.E.2 Computing profile likelihoods

We describe here how we compute $\ell_p(D\lambda)$ as defined in Equation (8). We use a simple iterative approach, although other approaches that exploit the PDE for improving the computation could be used [94]. The basic idea is to step to the left and right of the MLE $\widehat{D\lambda}$ until we find where $\ell_p(D\lambda) \leq c^*$ in each direction, or until we reach the bounds of $D\lambda$; recall that $c^* = -\chi^2_{1,1-\alpha}/2 \approx -1.92$ in this work, taking $\alpha = 0.05$. At each step, we solve the optimisation problem (8) to get a new value for $\ell_p(D\lambda)$ at the given $D\lambda$. This optimisation problem starts with an initial estimate given by the MLE if we have only taken one step, or via linear interpolation of the optimised values for $\lambda^*(D\lambda)$ from the previous two steps, with $\lambda^*(D\lambda)$ denoting the optimised value of λ that together gives the value for $\ell_p(D\lambda) = \max_{\lambda \in \Lambda}[\ell(D\lambda, \lambda)] - \ell^*$, meaning $\ell_p(D\lambda) = \ell(D\lambda, \lambda^*(D\lambda)) - \ell^*$. If we find points on each side of the MLE where $\ell_p(D\lambda) \leq c^*$, we stop iterating and fit a spline to the data $(D\lambda_i, \ell_p(D\lambda_i))$, using a bisection algorithm on each side of the MLE to find the two points where $\ell_p(D\lambda) = c^*$. These two points define the endpoints of the confidence interval. This procedure is implemented in the Julia package `ProfileLikelihood.jl` [77].

2.E.3 Computing prediction intervals

Let us now describe how prediction intervals are computed, following the approach developed by [28]. We note that while we describe the procedure below for propagating uncertainty from the full log-likelihood $\ell(\boldsymbol{\theta} \mid \mathbf{y})$, we could just as easily propagate uncertainty from the profile likelihoods, again following [28]. The results turn out to be essentially the same, and so we only describe the former approach here.

We start with the same 40×40 grid that we use for finding an estimate estimate for computing the maximum likelihood, as described in Appendix 2.E.1. We then find all pairs $\boldsymbol{\theta} = (D\lambda, \lambda)$ in this grid such that $\ell(\boldsymbol{\theta}) - \ell^* \geq -\chi^2_{2,1-\alpha}/2$, where $\chi^2_{2,q}$ is the qth quantile of the χ^2 distribution with two degrees of freedom and $\ell^* = \ell(\hat{\boldsymbol{\theta}})$ is the maximum log-likelihood. With $\alpha = 0.05$, $-\chi^2_{2,0.95}/2 \approx -3$. We enumerate the points satisfying this condition as $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r\}$. For each point $\boldsymbol{\theta}_i$ we compute $\mathbf{q}_i = \mathbf{q}(\boldsymbol{\theta}_i)$ for a prediction function $\mathbf{q}$, giving a sample $(\mathbf{q}_1, \ldots, \mathbf{q}_r)$. Now, letting q_{ij} denote the jth element of $\mathbf{q}_i$, define $\mathbf{q}_L = (\min_{i=1}^r q_{i1}, \ldots, \min_{i=1}^r q_{i|q|})$ and $\mathbf{q}_U = (\max_{i=1}^r q_{i1}, \ldots, \max_{i=1}^r q_{i|q|})$, where $|q|$ is the length of outputs of $\mathbf{q}$. A parameter-wise prediction interval for $\mathbf{q}$ is then given by $\mathbf{q}_L \leq \mathbf{q} \leq \mathbf{q}_U$, where the vector inequality $\mathbf{a} \leq \mathbf{b} \leq \mathbf{c}$ means $a_i \leq b_i \leq c_i$ for each i.

For our application, the prediction function $\mathbf{q}$ is defined by

$$\mathbf{q}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{A}(\boldsymbol{\theta}; \mathbf{t}^*) \\ \mathbf{P}(\boldsymbol{\theta}; \mathbf{t}^*) \\ t_b(\boldsymbol{\theta}) \end{bmatrix}. \tag{2.22}$$

For these functions, we let $\mathbf{t}^*$ be a vector of $m = 361$ equally spaced points between $t = 5$ day and $t = 70$ day. Then, $\mathbf{A}(\boldsymbol{\theta}; \mathbf{t}^*)$ is the vector of coverages $(\mu_{\mathrm{c}}(t_1^*), \ldots, \mu_{\mathrm{c}}(t_m^*))$ for the given $\boldsymbol{\theta}$; $\mathbf{P}(\boldsymbol{\theta}; \mathbf{t}^*)$ is the corresponding vector of normalised perimeters $(\mu_{\mathrm{p}}(t_1^*), \ldots, \mu_{\mathrm{p}}(t_m^*))$ for the given $\boldsymbol{\theta}$; $t_b(\boldsymbol{\theta})$ is the time at which the area of the void first becomes zero, in particular this is the bridging time for the given $\boldsymbol{\theta}$, computing using the continuous callback interface from `DifferentialEquations.jl` [85] to find when $\mu_{\mathrm{c}}(t) \approx 0$ by applying rootfinding to the function $g(t) = \mu_{\mathrm{c}}(t) - 10^{-9}$.

2.F Likelihood results using area only

In this appendix, we show some of the results when we include only area in the likelihood function rather than both area and perimeter. For the tissue growth predictions, we only show the results at $u_0 = 0.2$, noting that the results for other u_0 are mostly indistinguishable.

2.F.1 Square pore

Analogous figures to those in Figures 5–7 are shown in Figures 2.F.1–2.F.3. The results are very similar, with the main difference being that the uncertainty is much wider than when we also include perimeter information, as should be expected. The differences in the model predictions are also not too distinguishable compared to their counterparts when including perimeter information. Overall, we see that for this data on the square, perimeter does not contribute significantly to our understanding of these effects.

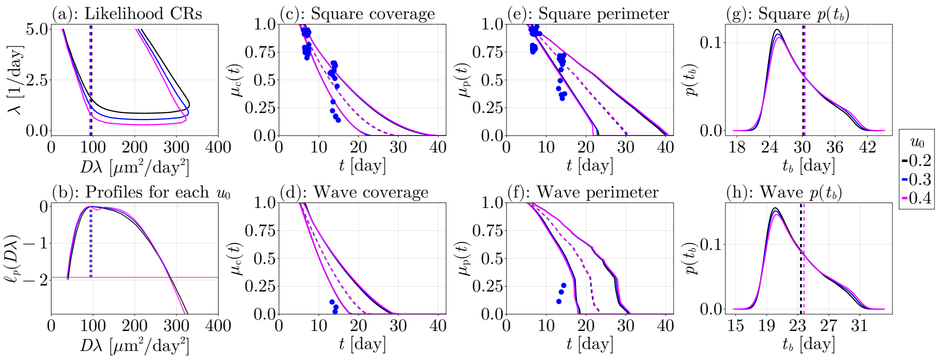

Figure 2.F.1: Likelihood analysis results for the square pore without perimeter information. In (a), the lines give the boundaries of the 95% confidence region for $\boldsymbol{\theta}$ for each u_0, and the vertical dashed lines show the MLE for $D\lambda$ (see Table 1). The profile log-likelihoods for $D\lambda$ for each u_0 are shown in (b), with the threshold $c^* \approx -1.92$ shown with a horizontal red line and the vertical dashed lines show the MLEs for $D\lambda$. In (c)–(f), predictions for $\mu_c(t)$ and $\mu_p(t)$ on each pore geometry are shown, with the blue dots showing the experimental data, the surrounding solid lines giving 95% prediction intervals for each u_0, and the dashed lines showing the corresponding estimates at the MLE $\hat{\boldsymbol{\theta}}$. The blue data points have been slightly jittered horizontally to help distinguish them. The estimates for the PDF $p(t_b)$ of the bridging time on each pore geometry are shown in (g)–(h). The results in (d), (f), and (h) are predictions on the wave geometry using parameters inferred from the square pore data.

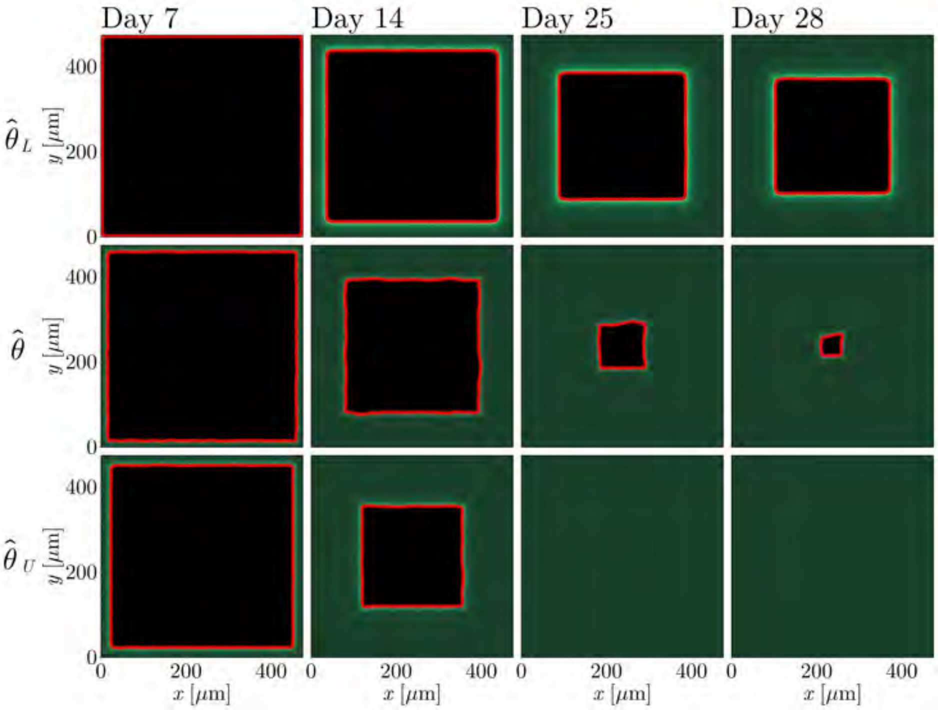

Figure 2.F.2: Model predictions for the variability in the tissue growth behaviour for the square geometry for $u_0 = 0.2$ at days 7, 14, 25, and 28 without perimeter information. The parameters used are $\hat{\boldsymbol{\theta}}_L = (\hat{D}_L, \hat{\lambda}_L) = (41\,\mu\text{m}^2/\text{day}, 1\,\text{day}^{-1})$, $\hat{\boldsymbol{\theta}} = (\hat{D}, \hat{\lambda}) = (19\,\mu\text{m}^2/\text{day}, 5\,\text{day}^{-1})$, and $\hat{\boldsymbol{\theta}}_L = (\hat{D}_U, \hat{\lambda}_U) = (57\,\mu\text{m}^2/\text{day}, 5\,\text{day}^{-1})$. The red boundary marks the position of the void boundary where $u(\mathbf{x}, t) = 1/2$. The top row of plots shows the predictions for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_L$, the middle row of plots shows the prediction for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, and the bottom row of plots show the predictions for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_U$; each individual column thus shows a rough confidence interval for the prediction at the respective time.

2.F.2 Wave pore

Analogous figures to those in Figures 8–10 are shown in Figures 2.F.4–2.F.6. The results are much more problematic in this case than we include area, with the uncertainty significantly wider than before, and $D\lambda$ is no longer identifiable. We do capture the data in our uncertainty intervals, although this is difficult to judge as the uncertainty being so large implies that we might have captured this data regardless. It is impressive, though, that we recover all of the experimental data points on the square. The bridging time distributions in this case cover a much wider range, again due to the large uncertainty in the parameters.

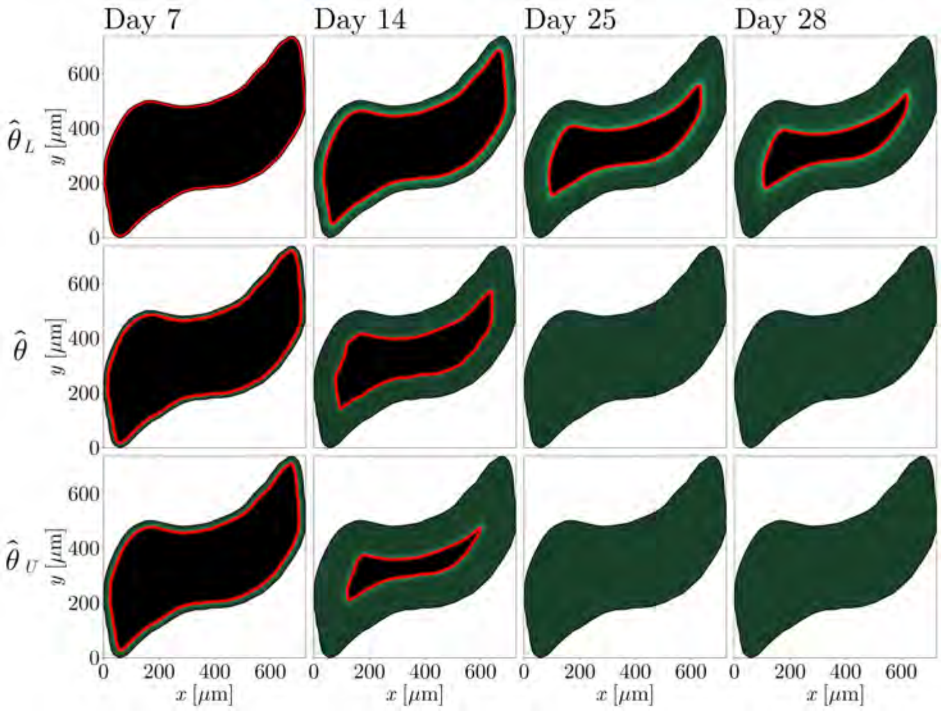

Figure 2.F.3: Model predictions for the variability in the tissue growth behaviour for the wave geometry using results from the square geometry for $u_0 = 0.2$ at days 7, 14, 25, and 28 without perimeter information. The parameters used are $\hat{\boldsymbol{\theta}}_L = (\hat{D}_L, \hat{\lambda}_L) = (41\,\mathrm{\mu m^2/day}, 1\,\mathrm{day^{-1}})$, $\hat{\boldsymbol{\theta}} = (\hat{D}, \hat{\lambda}) = (19\,\mathrm{\mu m^2/day}, 5\,\mathrm{day^{-1}})$, and $\hat{\boldsymbol{\theta}}_L = (\hat{D}_U, \hat{\lambda}_U) = (57\,\mathrm{\mu m^2/day}, 5\,\mathrm{day^{-1}})$. The red boundary marks the position of the void boundary where $u(\mathbf{x}, t) = 1/2$. The top row of plots shows the predictions for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_L$, the middle row of plots shows the prediction for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, and the bottom row of plots show the predictions for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_U$; each individual column thus shows a rough confidence interval for the prediction at the respective time.

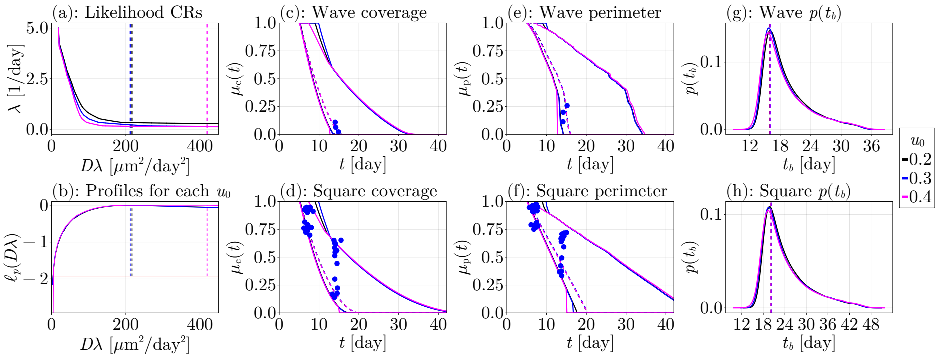

Figure 2.F.4: Likelihood analysis results for the wave pore without perimeter information. In (a), the lines give the boundaries of the 95% confidence region for $\boldsymbol{\theta}$ for each u_0, and the vertical dashed lines show the MLE for $D\lambda$ (see Table 1). The profile log-likelihoods for $D\lambda$ for each u_0 are shown in (b), with the threshold $c^* \approx -1.92$ shown with a horizontal red line and the vertical dashed lines show the MLEs for $D\lambda$. In (c)–(f), predictions for $\mu_c(t)$ and $\mu_p(t)$ on each pore geometry are shown, with the blue dots showing the experimental data, the surrounding solid lines giving 95% prediction intervals for each u_0, and the dashed lines showing the corresponding estimates at the MLE $\hat{\boldsymbol{\theta}}$. The blue data points have been slightly jittered horizontally to help distinguish them. The estimates for the PDF $p(t_b)$ of the bridging time on each pore geometry are shown in (g)–(h). The results in (d), (f), and (h) are predictions on the square geometry using parameters inferred from the wave pore data. The blue data points have been slightly jittered horizontally to help distinguish them.

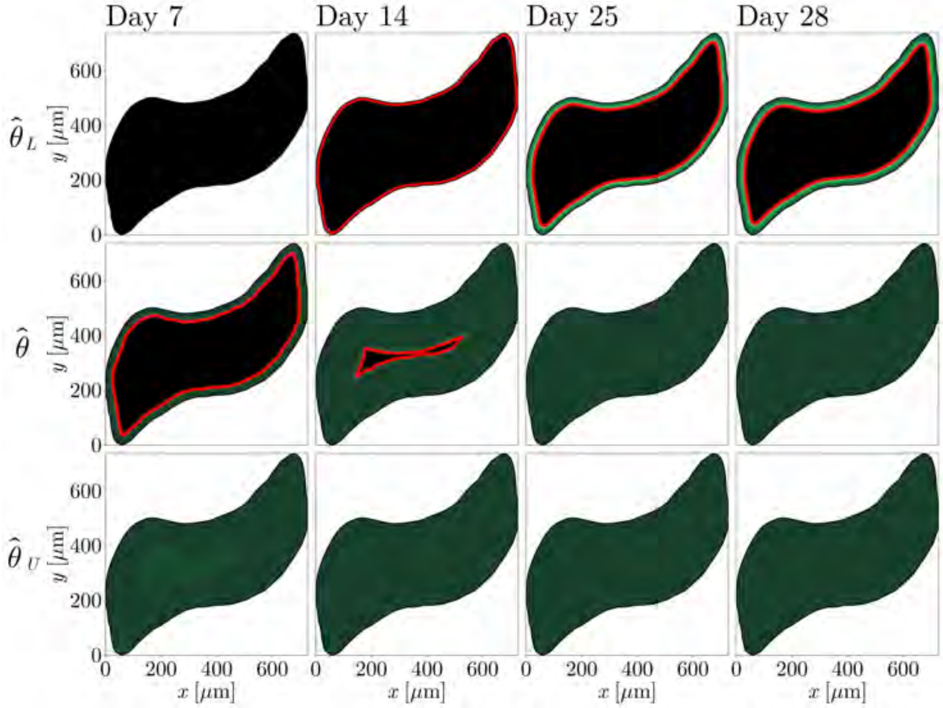

Figure 2.F.5: Model predictions for the variability in the tissue growth behaviour for the wave geometry for $u_0 = 0.2$ at days 7, 14, 25, and 28 without perimeter information. The parameters used are $\hat{\boldsymbol{\theta}}_L = (\hat{D}_L, \hat{\lambda}_L) = (25\,\mu\text{m}^2/\text{day}, 0.2\,\text{day}^{-1})$, $\hat{\boldsymbol{\theta}} = (\hat{D}, \hat{\lambda}) = (22\,\mu\text{m}^2/\text{day}, 10\,\text{day}^{-1})$, and $\hat{\boldsymbol{\theta}}_L = (\hat{D}_U, \hat{\lambda}_U) = (2000\,\mu\text{m}^2/\text{day}, 10\,\text{day}^{-1})$. The red boundary marks the position of the void boundary where $u(\mathbf{x}, t) = 1/2$. The top row of plots shows the predictions for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_L$, the middle row of plots shows the prediction for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, and the bottom row of plots show the predictions for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_U$; each individual column thus shows a rough confidence interval for the prediction at the respective time.

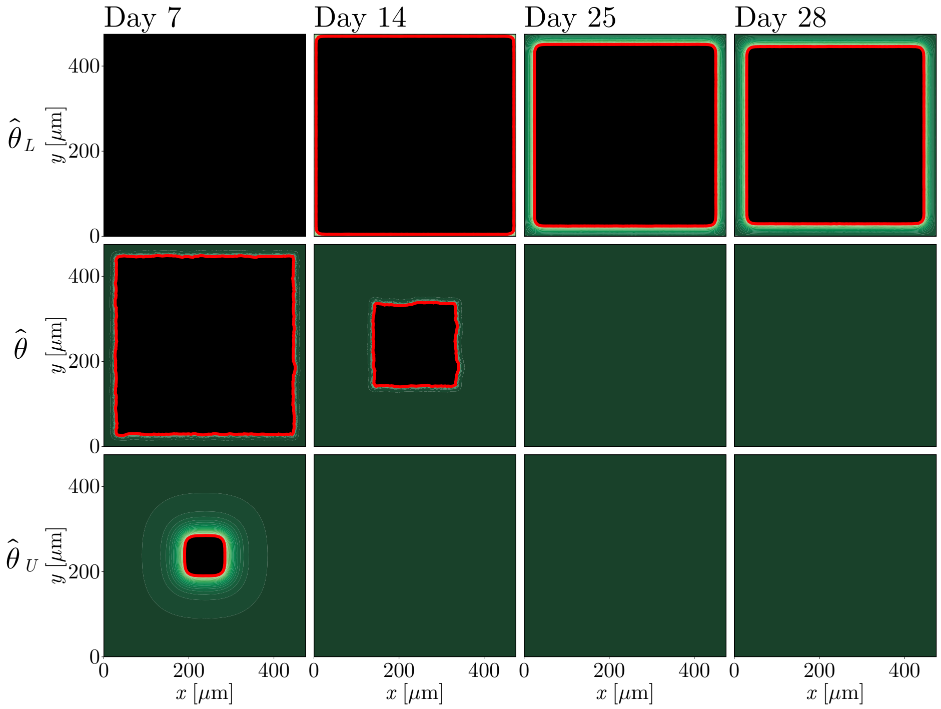

Figure 2.F.6: Model predictions for the variability in the tissue growth behaviour for the square geometry using results from the wave geometry for $u_0 = 0.2$ at days 7, 14, 25, and 28 without perimeter information. The parameters used are $\hat{\boldsymbol{\theta}}_L = (\hat{D}_L, \hat{\lambda}_L) = (25\,\mu\text{m}^2/\text{day}, 0.2\,\text{day}^{-1})$, $\hat{\boldsymbol{\theta}} = (\hat{D}, \hat{\lambda}) = (22\,\mu\text{m}^2/\text{day}, 10\,\text{day}^{-1})$, and $\hat{\boldsymbol{\theta}}_L = (\hat{D}_U, \hat{\lambda}_U) = (2000\,\mu\text{m}^2/\text{day}, 10\,\text{day}^{-1})$. The red boundary marks the position of the void boundary where $u(\mathbf{x}, t) = 1/2$. The top row of plots shows the predictions for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_L$, the middle row of plots shows the prediction for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, and the bottom row of plots show the predictions for each time at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_U$; each individual column thus shows a rough confidence interval for the prediction at the respective time.

Statement of Contribution of Co-Authors

The authors listed below have certified that:

1. they meet the criteria for authorship and that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
3. there are no other authors of the publication according to these criteria;
4. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit, and
5. they agree to the use of the publication in the student's thesis and its publication on the QUT's ePrints site consistent with any limitations set by publisher requirements.

Publication title: Pushing coarse-grained models beyond the continuum limit using equation learning

Publication status: Published in Proceedings of the Royal Society A: Mathematical, Physical, and Engineering Sciences

Contributor	Statement of contribution*
Daniel J. VandenHeuvel	Performed all the analyses and numerical simulations, developed the methodology, developed all the code and figures, and drafted the manuscript.
Pascal R. Buenzli	Supervised the analyses and numerical simulations and critically reviewed the manuscript.
Matthew J. Simpson	Initiated the research concept, supervised the analyses and numerical simulations, and critically reviewed the manuscript.

Chapter 3

Pushing coarse-grained models beyond the continuum limit using equation learning

A paper published in the Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences as

Pushing coarse-grained models beyond the continuum limit using equation learning

Daniel J. VandenHeuvel, Pascal R. Buenzli, Matthew J. Simpson

Abstract

Mathematical modelling of biological population dynamics often involves proposing high fidelity discrete agent-based models that capture stochasticity and individual-level processes. These models are often considered in conjunction with an approximate coarse-grained differential equation that captures population-level features only. These coarse-grained models are only accurate in certain asymptotic parameter regimes, such as enforcing that the time scale of individual motility far exceeds the time scale of birth/death processes. When these coarse-grained models are accurate, the discrete model still abides by conservation laws at the microscopic level, which implies that there is some macroscopic conservation law that can describe the macroscopic dynamics. In this work, we introduce an equation learning framework to find accurate coarse-grained models when standard continuum limit approaches are inaccurate. We demonstrate our approach using a discrete mechanical model of epithelial tissues, considering a series of four case studies that consider problems with and without free boundaries, and with and without proliferation, illustrating how we can learn macroscopic equations describing mechanical relaxation, cell proliferation, and the equation governing the dynamics of the free boundary of the tissue. While our presentation focuses on this biological application, our approach is more broadly applicable across a range of scenarios where discrete models are approximated by approximate continuum-limit descriptions. All code and data to reproduce this work are available at https://github.com/DanielVandH/StepwiseEQL.jl.

3.1 Introduction

Mathematical models of population dynamics are often constructed by considering both discrete and continuous descriptions, allowing for both microscopic and macroscopic details to be considered [95]. This approach has been applied to several kinds of discrete models, including cellular Potts models [96–99], exclusion processes [100–103], mechanical models of epithelial tissues [38, 40–42, 45, 104–106], hydrodynamics [107, 108], and a variety of other types of individual-based models [95, 109–116]. Continuum models are useful for describing collective behaviour, especially because the computational requirement of discrete models increases with the size of the population, and this can become computationally prohibitive for large populations, which is particularly problematic for parameter inference [117]. In contrast, the computational requirement to solve a continuous model is independent of the population size, and generally requires less computational overhead than working with a discrete approach only [40]. Continuum models are typically obtained by coarse-graining the discrete model, using Taylor series expansions to obtain continuous partial differential equation (PDE) models that govern the population densities on a continuum or macroscopic scale [41–44].

One challenge with using coarse-grained continuum limit models is that while the solution of these models can match averaged data from the corresponding discrete model for certain choices of parameters [41, 106, 118], the solution of the continuous model can be a very poor approximation for other parameter choices [40, 48, 105, 119]. More generally, coarse-grained models are typically only valid in certain asymptotic parameter regimes [118–120]. For example, suppose we have a discrete space, discrete time, agent-based model that incorporates random motion and random proliferation. Random motion involves stepping a distance Δ with probability $P_\mathrm{m} \in [0, 1]$ per time step of duration τ. The stochastic proliferation process involves undergoing proliferation with probability $P_\mathrm{p} \in [0, 1]$ per unit time step. The continuum limit description of this kind of discrete process can be written as [119]

$$\frac{\partial q}{\partial t} = \frac{\partial}{\partial x}\left(D(q)\frac{\partial q}{\partial x}\right) + R(q), \tag{3.1}$$

where q is the macroscopic density of individuals, $D(q)$ is the nonlinear diffusivity that describes the effects of individual migration, and $R(q)$ is a source term that describes the effects of the birth process in the discrete model [119]. Standard approaches to derive (3.1) require $D(q) = \mathcal{O}(P_\mathrm{m}\Delta^2/\tau)$ and $R(q) = \mathcal{O}(P_\mathrm{p}/\tau)$ in the limit that $\Delta \to 0$ and $\tau \to 0$. To obtain a well-defined continuum limit such that the diffusion and source terms are both present in the macroscopic model, some restrictions on the parameters in the discrete model are required [119, 120]. Typically, this is achieved by taking the limit as $\Delta \to 0$ and $\tau \to 0$ jointly such that the ratio Δ^2/τ remains finite, implying that $P_\mathrm{p} = \mathcal{O}(\tau)$ so that both the diffusion and source terms in (3.1) are $\mathcal{O}(1)$. In practice, this means that the time scale of individual migration events has to be much faster than the time scale of individual proliferation events, otherwise the continuum limit description is not well defined [119, 120]. If this restriction is not enforced, then the solution of the continuum limit model does not always predict the averaged behaviour of the discrete model [119], as the terms on the right-hand side of (3.1) are no longer $\mathcal{O}(1)$ so that the continuum limit is not well defined [120].

Regardless of whether choices of parameters in a discrete model obey the asymptotic restrictions imposed by coarse-graining, the discrete model still obeys a conservation principle, which implies that there is some alternative macroscopic conservation description that will describe population-level features of interest [33, 46]. Equation learning is a means of determining appropriate continuum models outside of the usual continuum limit asymptotic regimes. Equation learning has been used in several applications for model discovery. In the context of PDEs, a typical approach is to write $\partial q/\partial t = \mathcal{N}(q, \mathcal{D}, \boldsymbol{\theta})$, where q is the population density, $\mathcal{N}$ is some nonlinear function parametrised by $\boldsymbol{\theta}$, $\mathcal{D}$ is a collection of differential operators, and $\boldsymbol{\theta}$ are parameters to be estimated [51]. This formulation was first introduced by Rudy et al. [51], who extended previous work in learning ordinary differential equations (ODEs) proposed by Brunton et al. [47]. Equation learning methods developed for the purpose of learning biological models has also been a key interest [121, 122]. Lagergren et al. [49] introduce a biologically-informed neural network framework that uses equation learning that is guided by biological constraints, imposing a specific conservation PDE rather than a general nonlinear function $\mathcal{N}$. Lagergren et al. [49] use this framework to discover a model describing data from simple *in vitro* experiments that describe the invasion of populations of motile and proliferative prostate cancer cells. VandenHeuvel et al. [50] extend the work of Lagergren et al. [49], incorporating uncertainty quantification into the equation learning procedure through a bootstrapping approach. Nardini et al. [48] use discrete data from agent-based models to learn associated continuum ODE models, combining a user-provided library of functions together with sparse regression methods to give simple ODE models describing population densities. Regression methods have also been used as an alternative to equation learning for this purpose [52].

These previous approaches to equation learning consider various methods to estimate the parameters $\boldsymbol{\theta}$, such as sparse regression or nonlinear optimisation [47–51,121], representing $\mathcal{N}$ as a library of functions [47,51,121], neural networks [49], or in the form of a conservation law with individual components to be learned [49, 50]. In this work, we introduce a *stepwise equation learning* framework, inspired from stepwise regression [123], for estimating $\boldsymbol{\theta}$ from averaged discrete data with a given $\mathcal{N}$ representing a proposed form for the continuum model description. We incorporate or remove terms one at a time until a parsimonious continuum model is obtained whose solution matches the data well and no further improvements can be made to this match. Our approach is advantageous for several reasons. Firstly, it is computationally efficient and parallelisable, allowing for rapid exploration of results with different discrete parameters and different forms of $\mathcal{N}$ for a given data set. Secondly, the approach is modular, with different mechanistic features easily incorporated. This approach enables extensive computational experimentation by exploring the impact of including or excluding putative terms in the continuum model without any great increase in computational overhead. Lastly, it is easy to examine the results from our procedure, allowing for ease of diagnosing and correcting reasons for obtaining poor fitting models, and explaining what components of the continuum model are the most influential. We emphasise that a key difference between our approach and other work, such as the methods developed by Brunton et al. [47] and Rudy et al. [51], is that we constrain our problem so that we can only learn conservation laws rather than allow a general form through a library of functions, and that we iteratively eliminate variables from $\boldsymbol{\theta}$ rather than use sparse regression. These important features are what support the modularity and interpretability of our approach.

To illustrate our procedure, we consider a discrete, individual-based one-dimensional toy model inspired from epithelial tissues [41,106]. Epithelial tissues are biological tissue composed of cells, organised in a monolayer, and are present in many parts of the body and interact with other cells [124], lining surfaces such as the skin and the intestine [125]. They are important in a variety of contexts, such as wound healing [126,127] and cancer [128,129]. Many models have been developed for studying their dynamics, considering both discrete and continuum modelling [38,40–42,45,104,105], with most models given in the form of a nonlinear reaction-diffusion equation with a moving boundary, using a nonlinear diffusivity term to incorporate mechanical relaxation and a source term to model cell proliferation [38, 104, 105]. These continuum limit models too are only accurate in certain parameter regimes, becoming inaccurate if the rate of mechanical relaxation is slow relative to the rate of proliferation [40, 105, 119]. To apply our stepwise equation learning procedure, we let the nonlinear function $\mathcal{N}$ be given in the form of a conservation law together with equations describing the free boundary. We demonstrate this approach using a series of four biologically-motivated case studies, considering problems with and without a free boundary, and with and without proliferation,with each case study building on those before it. The first two case studies are used to demonstrate how our approach can learn known continuum limits, while the latter two case studies show how we can learn improved continuum limit models in parameter regimes where these known continuum limits are no longer accurate. We implement our approach in the Julia language [130], and all code is available on GitHub at https://github.com/DanielVandH/StepwiseEQL.jl.

3.2 Mathematical model

Following Murray et al. and Baker et al. [38, 41], we suppose that we have a set of nodes $x_1, \ldots, x_n(t)$ describing n cell boundaries at a time t. The interval $(x_i(t), x_{i+1}(t))$ represents the ith cell for $i = 1, \ldots, n-1$, where we fix $x_1 = 0$ and $x_1 < x_2(t) < \cdots < x_n(t)$. The number of nodes, n, may increase over time due to cell proliferation. We model the mechanical interaction between cells by treating them as springs, as indicated in Figure 3.1, so that each node i experiences forces $F_{i,i\pm1}$ from

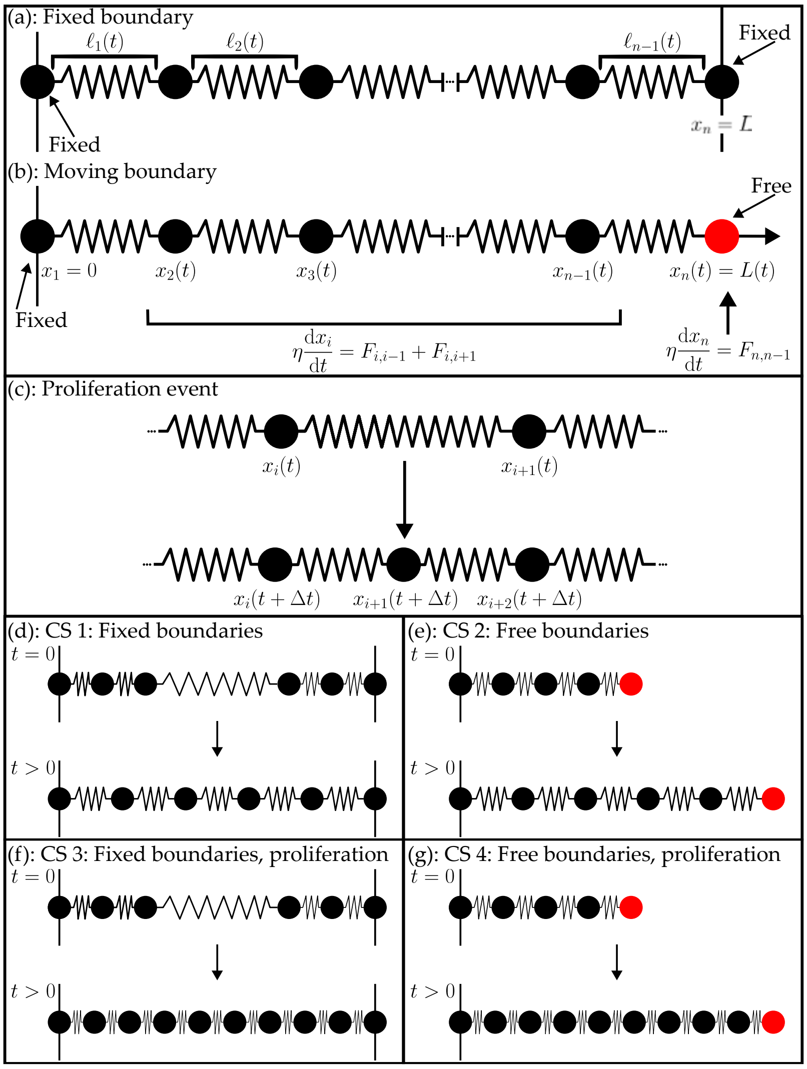

Figure 3.1: Discrete model and schematics for each case study (CS). (a) A fixed boundary problem with $x_1 = 0$ and $x_n = L$ fixed. (b) A free boundary problem with $x_1 = 0$ and $x_n(t) = L(t)$, show in red, free. (c) Proliferation schematic, showing a cell $(x_i(t), x_{i+1}(t))$ dividing into $(x_i(t+\Delta t), x_{i+1}(t+\Delta t))$ and $(x_{i+1}(t+\Delta t), x_{i+2}(t+\Delta t))$ following a proliferation event, where $x_{i+1}(t+\Delta t) = (x_i(t) + x_{i+1}(t))/2$. (d)–(g) show schematics for the four case studies considered in this work, where the first row in each panel is a representation of the initial configuration of cells at $t = 0$ and the second row a representation at a later time $t > 0$.

nodes $i \pm 1$, respectively, except at the boundaries where there is only one neighbouring force. We further assume that each of these springs has the same mechanical properties, and that the viscous force from the surrounding medium is given by $\eta \mathrm{d}x_i(t)/\mathrm{d}t$ with drag coefficient η. Lastly, assuming we are in a viscous medium so that the motion is overdamped, we can model the dynamics of each individual node $x_i(t)$, fixing $x_1 = 0$, by [38]

$$\eta \frac{\mathrm{d}x_i(t)}{\mathrm{d}t} = F_{i,i-1} + F_{i,i+1}, \quad i = 1, \ldots, n-1, \tag{3.2}$$

$$\eta \frac{\mathrm{d}x_n(t)}{\mathrm{d}t} = F_{n,n-1}, \tag{3.3}$$

where

$$F_{i,i\pm1} = F\left(|x_i(t) - x_{i\pm1}(t)|\right) \frac{x_i(t) - x_{i\pm1}(t)}{|x_i(t) - x_{i\pm1}(t)|} \tag{3.4}$$

is the interaction force that the ith node experiences from nodes $i \pm 1$ (Figure 3.1). In Case Studies 1 and 3 (see Section 3.3, below), we hold $x_n(t) = L$ constant and discard (3.3). Throughout this work, we use linear Hookean springs so that $F(\ell_i) = k(s - \ell_i)$, $\ell_i > 0$, where $\ell_i(t) = x_{i+1}(t) - x_i(t)$ is the length of the ith cell, $k > 0$ is the spring constant, and $s \geq 0$ is the resting spring length [41]; we discuss other force laws in 3.E.

The dynamics governed by (3.2)–(3.3) describe a system in which cells mechanically relax. Following previous work [38, 41, 45], we introduce a stochastic mechanism that allows the cells to undergo proliferation, assuming only one cell can divide at a time over a given interval $[t, t + \Delta t)$ for some small duration Δt. We let the probability that the ith cell proliferates be given by $G_i \Delta t$, where $G_i = G(\ell_i)$ for some length-dependent proliferation law $G(\ell_i) > 0$. As represented in Figure 3.1(c), when the ith cell proliferates, the cell divides into two equally-sized daughter cells, and the boundary between the new daughter cells is placed at the midpoint of the original cell. Throughout this work, we use a logistic proliferation law $G(\ell_i) = \beta[1 - 1/(K\ell_i)]$ with $\ell_i > 1/K$, where β is the intrinsic proliferation rate and K is the carrying capacity density; we consider other proliferation laws in 3.E. The implementation of the solution to these equations (3.2)–(3.3) and the proliferation mechanism is given in the Julia package `EpithelialDynamics1D.jl`; in this implementation, if $G(\ell_i) < 0$ we set $G(\ell_i) = 0$ to be consistent with the fact that we interpret $G(\ell_i)$ as a probability. We emphasise that, without proliferation, we need only solve (3.2)–(3.3) once for a given initial condition in order to obtain the expected behaviour of the discrete model, because the discrete model is deterministic in the absence of proliferation. In contrast, incorporating proliferation means that we need to consider several identically-prepared realisations of the same stochastic discrete model to estimate the expected behaviour of the discrete model for a given initial condition.

In practice, macroscopic models of populations of cells are described in terms of cell densities rather than keeping track of the position of individual cell boundaries. The density of the ith cell $(x_i(t), x_{i+1}(t))$ is $1/\ell_i(t)$. For an interior node $x_i(t)$, we obtain a density $q_i(t)$ by taking the inverse of the average of the cells left and right of $x_i(t)$, giving

$$q_i(t) = \frac{2}{x_{i+1}(t) - x_{i-1}(t)}, \quad i = 2, \ldots, n-1, \tag{3.5}$$

as in Baker et al. [38]. At boundary nodes, we use

$$q_1(t) = \frac{2}{x_2(t)} - \frac{2}{x_3(t)}, \quad q_n(t) = \frac{2}{x_n(t) - x_{n-1}(t)} - \frac{2}{x_n(t) - x_{n-2}(t)}, \tag{3.6}$$

derived by linear extrapolation of (3.5) to the boundary. The densities in (3.6) ensure that the slope of the density curves at the boundaries, $\partial q/\partial x$, match those in the continuum limit. We discuss the derivation of (3.6) in 3.B. In the continuum limit where the number of cells is large and mechanical relaxation is fast, the densities evolve according to the moving boundary problem [38, 41]

$$\begin{aligned}
\frac{\partial q}{\partial t} &= \frac{\partial}{\partial x}\left(D(q)\frac{\partial q}{\partial x}\right) + R(q) && 0 < x < L(t),\, t > 0, \\
\frac{\partial q}{\partial x} &= 0 && x = 0,\, t > 0, \\
\frac{\partial q}{\partial x} &= H(q) && x = L(t),\, t > 0, \\
q\frac{\mathrm{d}L}{\mathrm{d}t} &= -D(q)\frac{\partial q}{\partial x} && x = L(t),\, t > 0,
\end{aligned} \tag{3.7}$$

where $q(x, t)$ is the density at position x and time t, $D(q) = -1/(\eta q^2)F'(1/q)$, $R(q) = qG(1/q)$, $H(q) = -2qF(1/q)/[\eta D(q)]$, and $L(t) = x_n(t)$ is the leading edge position with $L(0) = x_n(0)$. The quantity $1/q$ in these equations can be interpreted as a continuous function related to the length of the individual cells. The initial condition $q(x, 0) = q_0(x)$ is a linear interpolant of the discrete densities $q_i(t)$ of the cells at $t = 0$. Similar to the discussion of (3.1), for this continuum limit to be valid so that both $D(q)$ and $R(q)$ play a role in the continuum model, constraints must be imposed on the discrete parameters. As discussed by Murphy et al. [40], we require that the time scale of mechanical relaxation is sufficiently fast relative to the time scale of proliferation. In practice this means that for a given choice of β we must have k/η sufficiently large for the solution of the continuum model to match averaged data from the discrete model. We note that, with our choices of F and G, the functions in (3.7) are given by

$$D(q) = \frac{k}{\eta q^2}, \quad R(q) = \beta q\left(1 - \frac{q}{K}\right), \quad H(q) = 2q^2(1 - qs). \tag{3.8}$$

For fixed boundary problems we take $H(q) = 0$ and $\mathrm{d}L/\mathrm{d}t = 0$. In 3.C, we describe how to solve (3.7) numerically, as well as how to solve the corresponding problem with fixed boundaries numerically.

3.3 Continuum-discrete comparison

We now consider four biologically-motivated case studies to illustrate the performance of the continuum limit description (3.7). These case studies are represented schematically in Figure 3.1(d)—(g). Case Studies 1 and 3, shown in Figure 3.1(d) and Figure 3.1(f), are fixed boundary problems, where we see cells relax mechanically towards a steady state where each cell has equal length. Case Studies 2 and 4 are free boundary problems, where the right-most cell boundary moves in the positive x-direction while all cells relax towards a steady state where the length of each cell is given by resting spring length s. Case Studies 1 and 2 have $\beta = 0$ so that there is no cell proliferation and the number of cells remains fixed during the simulations, whereas Case Studies 3 and 4 have $\beta > 0$ so that the number of cells increases during the discrete simulations. To explore these problems, we first consider cases where the continuum limit model is accurate, using the data shown in Figure 3.2, where we show space-time diagrams and a set of averaged density profiles for each problem in the left and right columns of Figure 3.2, respectively. Case Studies 1 and 3 initially place 30 nodes in $0 \leq x \leq 5$ and 30 nodes in $25 \leq x \leq 30$, or equivalently $n = 60$ with 28 cells in $0 \leq x \leq 5$ and 28 cells in $25 \leq x \leq 30$, spacing the nodes uniformly

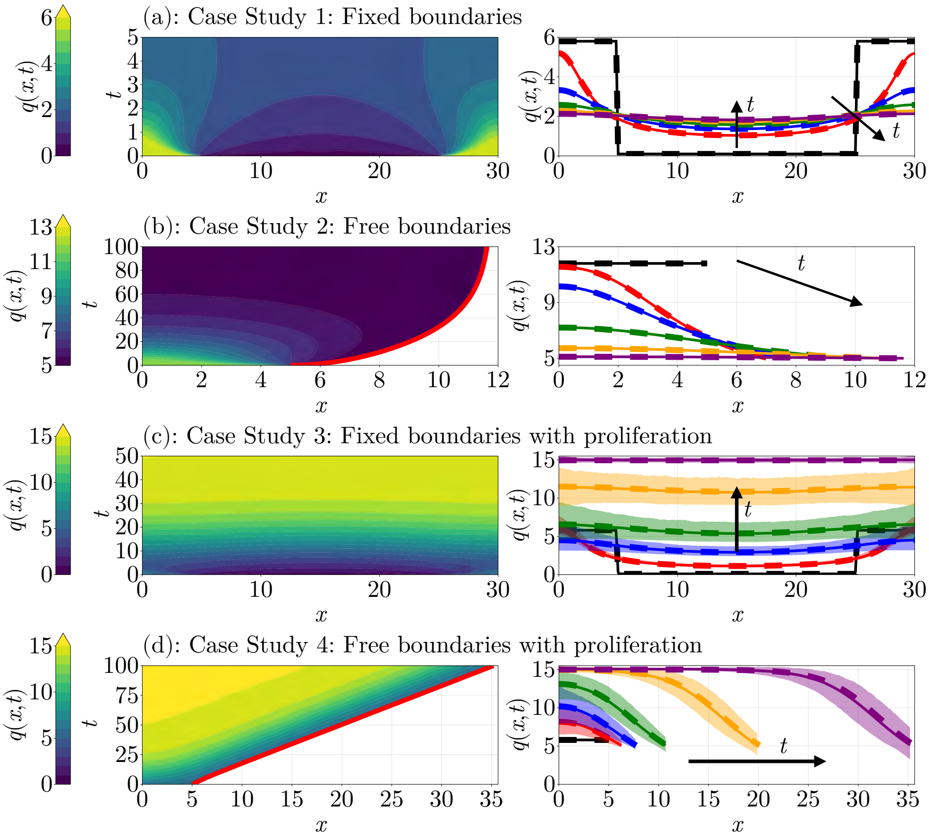

Figure 3.2: Space-time diagrams (left column) and densities (right column) for the four case studies from Figure 3.1 considered throughout this work. The left column shows the evolution of the discrete densities in space and time, with (c) and (d) showing averaged results over 2500 identically-prepared realisations of the discrete model. In (b) and (d), the red line shows the position of the free boundary. In the figures in the right column, the solid curves are the discrete densities (3.5) and the dashed curves are solutions to the continuum limit problem (3.7), and the curves are given by black, red, blue, green, orange, and purple in the order of increasing time as indicated by the black arrows. The times shown are (a) $t = 0, 1, 2, 3, 4, 5$; (b) $t = 0, 5, 10, 25, 50, 100$; (c) $t = 0, 1, 5, 10, 20, 50$; and (d) $t = 0, 5, 10, 20, 50, 100$. In (c) and (d), the shaded regions show 95% confidence bands from the mean discrete curves at each time; the curves in (a) and (b) show no shaded regions as these models have no stochasticity.

within each subinterval. Case Studies 2 and 4 initially place 60 equally spaced nodes in $0 \leq x \leq 5$.

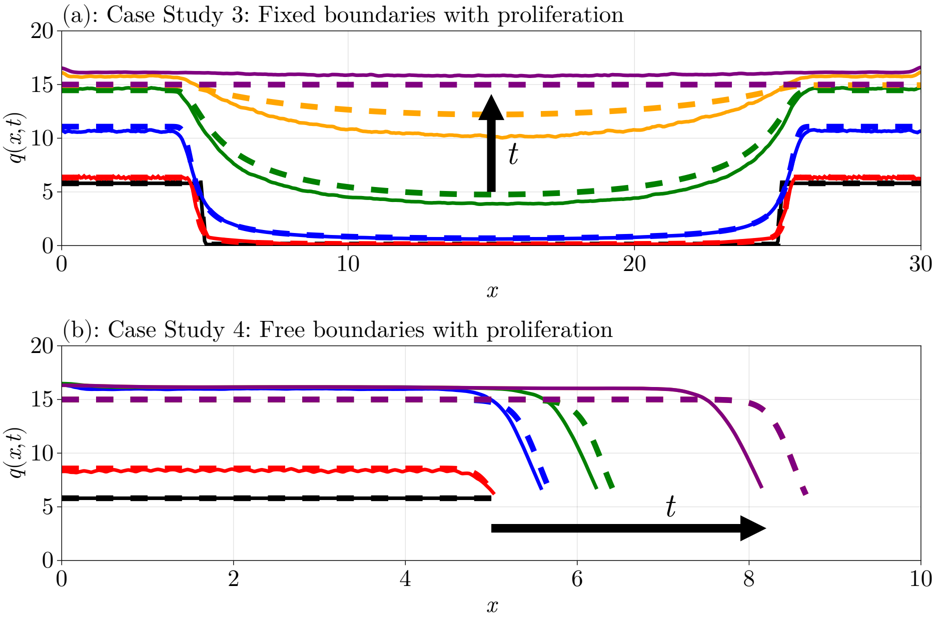

Figure 3.3: Examples of inaccurate continuum limits for (a) Case Study 3 and (b) Case Study 4, where both case studies use the same parameters as in Figure 3.2 except with $k = 1/5$ rather than $k = 50$. The solid curves are the discrete densities (3.5) and the dashed curves are solutions to the continuum limit problem (3.7). The arrows show the direction of increasing time. The density profiles are plotted in black, red, blue, green, orange, and purple for the respective times (a) $t = 0, 1, 10, 25, 40, 75$ and (b) $t = 0, 5, 25, 50, 100, 250$.

The problems shown in Figure 3.2 use parameter values such that the solution of the continuum limit (3.7) is a good match to the averaged discrete density profiles. In particular, all problems use $k = 50$, $\eta = 1$, $s = 1/5$ and, for Case Studies 3 and 4, $\Delta t = 10^{-2}$, $K = 15$, and $\beta = 0.15$. The accuracy of the continuum limit is clearly evident in the right column of Figure 3.2 where, in each case, the solution of the continuum limit model is visually indistinguishable from averaged data from the discrete model. With proliferation, however, the continuum limit can be accurate when k/η is not too much larger than β, and we use Case Studies 3 and 4 to explore this.

Figure 3.3 shows further continuum-discrete comparisons for Case Studies 3 and 4 where we have slowed the mechanical relaxation by taking $k = 1/5$. This choice of k means that $D(q)$ and $R(q)$ are no longer on the same scale and thus the continuum limit is no longer well defined, as explained in the discussion of (3.1), meaning the continuum limit solutions are no longer accurate. In both cases, the solution of the continuum limit model lags behind the averaged data from the discrete model. In 3.A, we show the 95% confidence regions for each curve in Figure 3.3, where we find that the solutions have much greater variance compared to the corresponding curves in Figure 3.2 where $k = 50$.

We are interested in developing an equation learning method for learning an improved continuum model for problems like those in Figure 3.3, allowing us to extend beyond the parameter regime where the continuum limit (3.7) is accurate. We demonstrate this in Case Studies 1–4 in Section 3.4 where we develop such a method.

3.4 Learning accurate continuum limit models

In this section we introduce our method for equation learning and demonstrate the method using the four case studies from Figures 3.1–3.3. Since the equation learning procedure is modular, adding these components into an existing problem is straightforward. All Julia code to reproduce these results is available at https://github.com/DanielVandH/StepwiseEQL.jl. A summary of all the parameters used for each case study is given in Table 3.1.

3.4.1 Case Study 1: Fixed boundaries

Case Study 1 involves mechanical relaxation only so that there is no cell proliferation and the boundaries are fixed, implying $R(q) = 0$ and $H(q) = 0$ in (3.7), respectively, and the only function to learn is $D(q)$.

Our equation learning approach starts by assuming that $D(q)$ is a linear combination of d *basis coefficients* $\{\theta_1, \ldots, \theta_d\}$ and d *basis functions* $\{\varphi_1, \ldots, \varphi_d\}$, meaning $D(q)$ can be represented as

$$D(q) = \sum_{i=1}^{d} \theta_i \varphi_i(q). \tag{3.9}$$

These basis functions could be any univariate function of q, for example the basis could be $\{\varphi_1, \varphi_2, \varphi_3\} = \{1/q, 1/q^2, 1/q^3\}$ with $d = 3$. In this work, we impose the constraint that $D(q) \geq 0$ for $q_{\min} \leq q \leq q_{\max}$, where

Parameter	**Case Study**					
	1	**2**	**3a**	**3b**	**4a**	**4b**
k	50	50	50	$1/5$	50	$1/5$
η	1	1	1	1	1	1
s	$1/5$	$1/5$	$1/5$	$1/5$	$1/5$	$1/5$
Δt	—	—	10^{-2}	10^{-2}	10^{-2}	10^{-2}
β	—	—	0.15	0.15	0.15	0.15
K	—	—	15	15	15	15
M	50	150	501	751	$(25, 50, 100, 250)$	$(20, 200, 200, 200)$
t_1	0	0	0	0	$(0, 0, 5, 10)$	$(0, 2, 10, 20)$
t_M	5	15	50	75	$(10^{-1}, 5, 10, 50)$	$(2, 10, 20, 50)$
n_s	—	—	1000	1000	1000	1000
n_k	—	—	50	200	$(25, 50, 100, 50)$	$(50, 100, 100, 100)$
τ_q	0.1	0.35	0.1	0.25	$(0.1, 0, 0, 0)$	$(0, 0, 0, 0.3)$
$\tau_{\mathrm{d}L/\mathrm{d}t}$	—	0.1	—	—	$(0, 0.2, 0, 0)$	$(0, 0.4, 0, 0)$
τ_t	0	0	0	0	0	$(0.4, 0.4, 0, 0)$

Table 3.1: Parameters used for each case study. The parameters are k, the spring constant; η, the drag coefficient; s, the resting spring length; Δt, the proliferation duration; β, the intrinsic proliferation rate; K, the carrying capacity density; M, the number of time points; t_1, the initial time; t_M, the final time; n_s, the number of identically-prepared realisations; n_k, the number of knots used for averaging over realisations; τ_q, which defines the $100\tau_q\%$ and $100(1-\tau_q)\%$ density quantiles; $\tau_{\mathrm{d}L/\mathrm{d}t}$, which defines the $100\tau_{\mathrm{d}L/\mathrm{d}t}\%$ and $100(1-\tau_{\mathrm{d}L/\mathrm{d}t})\%$ velocity quantiles; and τ_t, which defines the $100\tau_t\%$ and $100(1-\tau_t)\%$ temporal quantiles. Values indicated by a line are not relevant for the corresponding case study. For Case Study 3 and 4, the label "a" refers to the accurate continuum limit case, and "b" refers to the inaccurate continuum limit case. For Case Study 4, some parameters are given by a set of four parameters, with the ith value of this set referring to the value used when learning the ith mechanism; see Section 43.4.4 for details.

$q_{\min}$ and $q_{\max}$ are the minimum and maximum densities observed in the discrete simulations, respectively. This constraint enforces the condition that the nonlinear diffusivity function is positive over the density interval of interest. While it is possible to work with some choices of nonlinear diffusivity functions for which $D(q) < 0$ for some interval of density [131–133], we wish to avoid the possibility of having negative nonlinear diffusivity functions and our results support this approach.

The aim is to estimate $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^{\mathsf{T}}$ in (3.9). We use ideas similar to the basis function approach from VandenHeuvel et al. [50], using (3.9) to construct a matrix problem for $\boldsymbol{\theta}$. In particular, let us take the PDE (3.7), with $R(q) = 0$ and $H(q) = 0$, and expand the spatial derivative term so that we can isolate the θ_k terms,

$$\frac{\partial q_{ij}}{\partial t} = \sum_{k=1}^{d} \left\{ \frac{\mathrm{d}\varphi_k(q_{ij})}{\mathrm{d}q} \left(\frac{\partial q_{ij}}{\partial x}\right)^2 + \varphi_k(q_{ij}) \frac{\partial^2 q_{ij}}{\partial x^2} \right\} \theta_k, \tag{3.10}$$

where we let q_{ij} denote the discrete density at position $x_{ij} = x_i(t_j)$ and time t_j. We note that while q_{ij} is discrete, we assume it can be approximated by a smooth function, allowing us to define these derivatives $\partial q_{ij}/\partial t$, $\partial q_{ij}/\partial x$, and $\partial^2 q_{ij}/\partial x^2$ in (3.10); this assumption is appropriate since, as shown in Figure 3.2, these discrete densities can be well approximated by smooth functions. These derivatives are estimated using finite differences, as described in 3.D. We also emphasise that, while (3.10) appears similar to results in [47, 51], the crucial difference is that we are specifying forms for the *mechanisms* of the PDE rather than the *complete* PDE itself; one other important difference is in how we estimate $\boldsymbol{\theta}$, defined below and in (3.15). We save the solution to the discrete problems (3.2)–(3.3) at M times $0 = t_1 < t_2 < \cdots < t_M$ so that $i \in \{1, \ldots, n\}$ and $j \in \{2, \ldots, M\}$, where $n = 60$ is the number of nodes and we do not deal with data at $j = 1$ since the PDE does not apply at $t = 0$. We can therefore convert (3.10) into a rectangular matrix problem $\mathbf{A}\boldsymbol{\theta} = \mathbf{b}$, where the rth row in $\mathbf{A}$, $r = 1, \ldots, n(M-1)$, corresponding to the point (x_{ij}, t_j) is given by $\mathbf{a}_{ij} \in \mathbb{R}^{1\times d}$, where

$$\mathbf{a}_{ij}^{\mathsf{T}} = \begin{bmatrix} \dfrac{\mathrm{d}\varphi_1(q_{ij})}{\mathrm{d}q} \left(\dfrac{\partial q_{ij}}{\partial x}\right)^2 + \varphi_1(q_{ij}) \dfrac{\partial^2 q_{ij}}{\partial x^2} \\ \vdots \\ \dfrac{\mathrm{d}\varphi_d(q_{ij})}{\mathrm{d}q} \left(\dfrac{\partial q_{ij}}{\partial x}\right)^2 + \varphi_d(q_{ij}) \dfrac{\partial^2 q_{ij}}{\partial x^2} \end{bmatrix}, \tag{3.11}$$

with each element of $\mathbf{a}_{ij}$ corresponding to the contribution of the associated basis function in (3.10). Thus, we obtain the system

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_{12} \\ \mathbf{a}_{22} \\ \vdots \\ \mathbf{a}_{nM} \end{bmatrix} \in \mathbb{R}^{n(M-1)\times d} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} \partial q_{12}/\partial t \\ \partial q_{22}/\partial t \\ \vdots \\ \partial q_{nM}/\partial t \end{bmatrix} \in \mathbb{R}^{n(M-1)\times 1}. \tag{3.12}$$

The solution of $\mathbf{A}\boldsymbol{\theta} = \mathbf{b}$, given by $\boldsymbol{\theta} = (\mathbf{A}^{\mathsf{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{b}$, is obtained by minimising the residual $\|\mathbf{A}\boldsymbol{\theta} - \mathbf{b}\|_2^2$, which keeps all terms present in the learned model. We expect, however, just as in (3.8), that $\boldsymbol{\theta}$ is sparse so that $D(q)$ has very few terms, which makes the interpretation of these terms feasible [47,51]. There are several ways that we could solve $\mathbf{A}\boldsymbol{\theta} = \mathbf{b}$ to obtain a sparse vector, such as with sparse regression [51], but in this work we take a *stepwise equation learning* approach inspired by stepwise regression [123] as this helps with both the exposition and modularity of our approach. For this approach, we first let $\mathcal{I} = \{1, \ldots, d\}$ be the set of basis function indices. We let $\mathcal{A}_k$ denote the set of *active coefficients* at the kth iteration, meaning the indices of non-zero values in $\boldsymbol{\theta}$, starting with $\mathcal{A}_1 = \mathcal{I}$. The set of indices of zero values in $\boldsymbol{\theta}$, $\mathcal{I}_k = \mathcal{I} \setminus \mathcal{A}_k$, is called the set of *inactive coefficients*. To obtain the next set, $\mathcal{A}_{k+1}$, from a current set $\mathcal{A}_k$, we apply the following steps:

1. Let the vector $\boldsymbol{\theta}_{\mathcal{A}}$ denote the solution to $\mathbf{A}\boldsymbol{\theta} = \mathbf{b}$ subject to the constraint that each inactive coefficient θ_i is zero, meaning $\theta_i = 0$ for $i \in \mathcal{I} \setminus \mathcal{A}$ for a given set of active coefficients $\mathcal{A}$. We compute $\boldsymbol{\theta}_{\mathcal{A}}$ by solving the reduced problem in which the inactive columns of $\mathbf{A}$ are not included. The vector with $\mathcal{A} = \mathcal{A}_k$ at step k is denoted $\boldsymbol{\theta}_k$. With this definition, we compute the sets

$$\mathcal{M}_k^+ = \left\{\boldsymbol{\theta}_{\mathcal{A}_k \cup \{i\}} : i \notin \mathcal{A}_k\right\} \quad \text{and} \quad \mathcal{M}_k^- = \left\{\boldsymbol{\theta}_{\mathcal{A}_k \setminus \{i\}} : i \in \mathcal{A}_k\right\}. \tag{3.13}$$

$\mathcal{M}_k^+$ is the set of all coefficient vectors $\boldsymbol{\theta}$ obtained by making each active coefficient at step k inactive one at a time. $\mathcal{M}_k^-$, is similar to $\mathcal{M}_{k+1}^-$ except we make each inactive coefficient at step k active one at a time. We then define $\mathcal{M}_k = \{\boldsymbol{\theta}_k\} \cup \mathcal{M}_k^+ \cup \mathcal{M}_k^-$, so that $\mathcal{M}_k$ is the set of all coefficient vectors obtained from activating coefficients one at a time, deactivating coefficients one at a time, or retaining the current vector $\boldsymbol{\theta}_k$.

2. Choose one of the vectors in $\mathcal{M}_k$ by defining a loss function $\mathcal{L}(\boldsymbol{\theta})$:

$$\underbrace{\mathcal{L}(\boldsymbol{\theta})}_{\text{loss}} = \underbrace{\log\left[\frac{1}{n(M-1)}\sum_{j=2}^{M}\sum_{i=1}^{n}\left(\frac{q_{ij} - q(x_{ij}, t_j; \boldsymbol{\theta})}{q_{ij}}\right)^2\right]}_{\text{goodness of fit}} + \underbrace{\|\boldsymbol{\theta}\|_0}_{\text{model complexity}}, \tag{3.14}$$

where $q(x, t; \boldsymbol{\theta})$ is the solution of the PDE (3.7) with $R(q) = H(q) = 0$

and $D(q)$ uses the coefficients $\boldsymbol{\theta}$ in (3.9), $q(x_{ij}, t_j; \boldsymbol{\theta})$ is the linear interpolant of the PDE data at $t = t_j$ evaluated at $x = x_{ij}$, and $\|\boldsymbol{\theta}\|_0$ is the number of non-zero terms in $\boldsymbol{\theta}$. This loss function balances the goodness of fit with model complexity. If, for some $\boldsymbol{\theta}$, $D(q) < 0$ within $q_{\min} \leq q \leq q_{\max}$, which we check by evaluating $D(q)$ at $n_c = 100$ equally spaced points in $q_{\min} \leq q \leq q_{\max}$, we set $\mathcal{L}(\boldsymbol{\theta}) = \infty$. With this loss function, we compute the next coefficient vector

$$\boldsymbol{\theta}_{k+1} = \underset{\boldsymbol{\theta} \in \mathcal{M}_k}{\operatorname{argmin}} \, \mathcal{L}(\boldsymbol{\theta}). \tag{3.15}$$

If $\boldsymbol{\theta}_{k+1} = \mathbf{0}$, so that all the coefficients are inactive, we instead take the vector that attains the second-smallest loss so that a model with no terms cannot be selected.

3. If $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k$, then there are no more local improvements to be made and so the procedure stops. Otherwise, we recompute $\mathcal{A}_{k+1}$ and $\mathcal{I}_{k+1}$ from $\boldsymbol{\theta}_{k+1}$ and continue iterating.

The second step prevents empty models from being returned, allowing the algorithm to more easily find an optimal model when starting with no active coefficients. We note that Nardini et al. [48] consider a loss based on the regression error, $\|\mathbf{A}\boldsymbol{\theta} - \mathbf{b}\|_2^2$, that has been useful for a range of previously-considered problems [47, 48, 51]. We do not consider the regression error in this work as we find that it typically leads to poorer estimates for $\boldsymbol{\theta}$ compared to controlling the density errors as we do in (3.15).

Let us now apply the procedure to our data from Figure 3.2, where we know that the continuum limit with $D(q) = 50/q^2$ is accurate. We use the basis functions $\varphi_i = 1/q^i$ for $i = 1, 2, 3$ so that

$$D(q) = \frac{\theta_1}{q} + \frac{\theta_2}{q^2} + \frac{\theta_3}{q^3}, \tag{3.16}$$

and we expect to learn $\boldsymbol{\theta} = (0, 50, 0)^\mathsf{T}$. We save the solution to the discrete model at $M = 50$ equally spaced time points between $t_1 = 0$ and $t_M = 5$. With this setup, and starting with all coefficients initially active so that $\mathcal{A}_1 = \{1, 2, 3\}$, we obtain the results in Table 3.2. The first iterate gives us $\boldsymbol{\theta}_1$ such that $D(q) < 0$ for some range of q as we show in Figure 3.4(a), and so we assign $\mathcal{L}(\boldsymbol{\theta}_1) = \infty$. To get to the next step, we remove θ_1, θ_2, and θ_3 one a time and compute the loss for each resulting vector, and we find that removing θ_3 leads to a vector that gives the least loss out of those considered. We thus find $\mathcal{A}_2 = \{1, 2\}$ and $\boldsymbol{\theta}_2 = (-1.46, 47.11, 0)^{\mathsf{T}}$. Continuing, we find that out of the choice of removing θ_1 or θ_2, or putting θ_3 back into the model, removing θ_1 decreases the loss by the greatest amount, giving $\mathcal{A}_3 = \{2\}$. Finally, we find that there are no more improvements to be made, and so the algorithm stops at $\boldsymbol{\theta}_3 = (0, 43.52, 0)^{\mathsf{T}}$, which is close to the continuum limit. We emphasise that this final $\boldsymbol{\theta}_3$ is a least squares solution with the constraint $\theta_1 = \theta_3 = 0$, thus there is no need to refine $\boldsymbol{\theta}_3$ further by eliminating θ_1 and θ_3 directly in (3.16), as the result would be the same. Comparing the densities from the solution of the learned PDE with $\boldsymbol{\theta} = \boldsymbol{\theta}_3$ with the discrete densities in Figure 3.5(a), we see that the curves are nearly visually indistinguishable near the center, but there are some visually discernible discrepancies near the boundaries. We show the form of $D(q)$ at each iteration in Figure 3.4(a), where we observe that the first iterate captures only the higher densities, the second iterate captures the complete range of densities, and the third iterate removes a single term which gives no noticeable difference.

Table 3.2: Stepwise equation learning results for the density data for Case Study 1: Fixed boundaries using the basis expansion (3.16), saving the results at $M = 50$ equally spaced times between $t_1 = 0$ and $t_M = 5$ and starting with all coefficients active, $\mathcal{A}_1 = \{1, 2, 3\}$. Coefficients highlighted in blue show the coefficient chosen to be removed or added at the corresponding step.

Step	θ_1	θ_2	θ_3	**Loss**
1	-5.97	70.73	**-27.06**	∞
2	**-1.46**	47.11	0.00	-4.33
3	0.00	43.52	0.00	-5.18

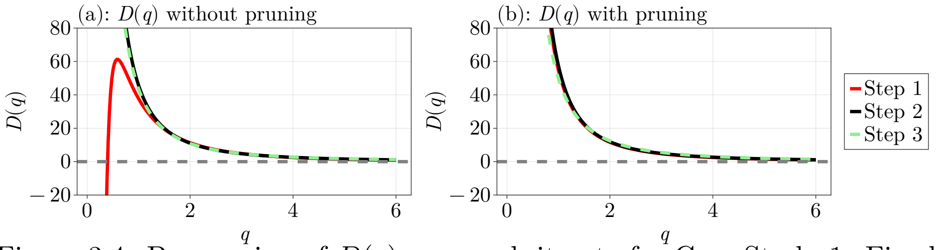

Figure 3.4: Progression of $D(q)$ over each iterate for Case Study 1: Fixed boundaries. (a) Progression from the results in Table 3.2 (dashed curves). (b) As in (a), except with the results from Table 3.3 using matrix pruning.

To improve our learned model we introduce *matrix pruning*, inspired from the data thresholding approach in VandenHeuvel et al. [50], to improve the estimates for $\boldsymbol{\theta}$. Visual inspection of the space-time diagram in Figure 3.2(a) shows that the most significant density changes occur at early time and near to locations where q changes in the initial condition, and a significant portion of the space-time diagram involves regions where q is almost constant. These regions where q has minimal change are problematic as points which lead to a higher residual are overshadowed, affecting the least squares problem and consequently degrading the estimates for $\boldsymbol{\theta}$ significantly, and so it is useful to only include important points in the construction of $\mathbf{A}$. To resolve this issue, we choose to only include points if the associated densities falls between the 10% and 90% quantiles for the complete set of densities, which we refer to by *density quantiles*; more details on this pruning procedure are given in 3.D. This choice of density quantiles is made using trial and error, starting at 0% and 100%, respectively, and shrinking the quantile range until suitable results are obtained. When we apply this pruning and reconstruct $\mathbf{A}$, we obtain the improved results in Table 3.3 and associated densities in Figure 3.5(b). Compared with Table 3.2, we see that the coefficient estimates for $\boldsymbol{\theta}$ all lead to improved losses, and our final model now has $\boldsymbol{\theta} = (0, 49.83, 0)^{\mathsf{T}}$, which is much closer to the the continuum limit, as we see in Figure 3.5(b) where the solution curves are now visually indistinguishable everywhere. Moreover, we show in Figure 3.4(b) how $D(q)$ is updated at each iteration, where we see that the learned nonlinear diffusivity functions are barely different from the expected continuum limit result. These results demonstrate the importance of only including the most important points in $\mathbf{A}$.

Table 3.3: Improved results for Case Study 1: Fixed boundaries from Table 3.2, now using matrix pruning so that densities outside of the 10% and 90% density quantiles are not included. Coefficients highlighted in blue show the coefficient chosen to be removed or added at the corresponding step.

Step	θ_1	θ_2	θ_3	**Loss**
1	**-1.45**	42.48	13.76	-4.19
2	0.00	37.79	**19.69**	-5.46
3	0.00	49.83	0.00	-7.97

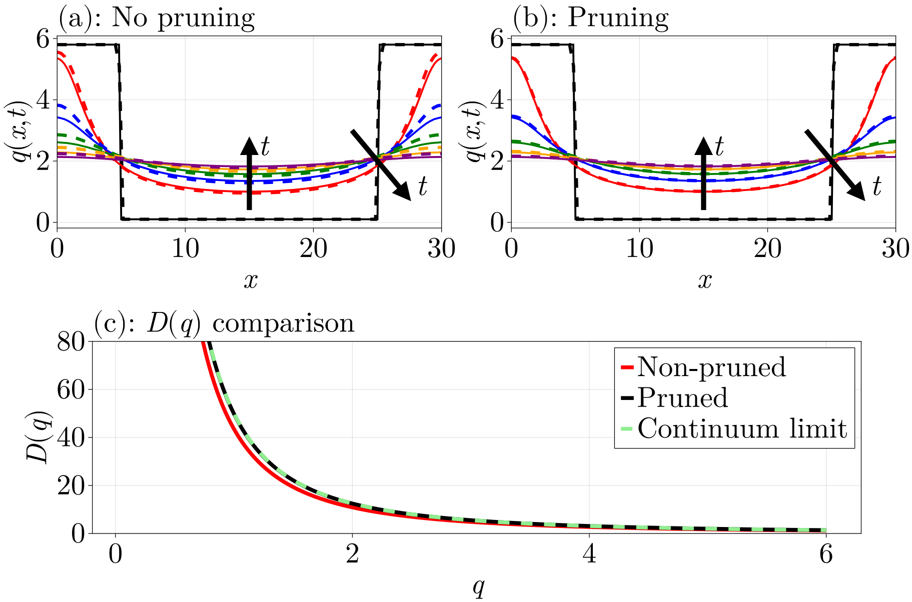

Figure 3.5: Stepwise equation learning results for Case Study 1: Fixed boundaries. (a) Comparisons of the discrete density profiles (solid curves) with those learned from PDEs obtained from the results in Table 3.2 (dashed curves). (b) As in (a), except with the results from Table 3.3 using matrix pruning so that densities outside of the 10% and 90% density quantiles are not included. (c) Comparisons of the learned $D(q)$ from Table 3.2 without pruning, Table 3.3 with pruning, and the continuum limit from (3.8). In (a)–(b), the arrows show the direction of increasing time, and the density profiles shown are at times $t = 0, 1, 2, 3, 4, 5$ in black, red, blue, green, orange, and purple, respectively.

3.4.2 Case Study 2: Free boundaries

Case Study 2 extends Case Study 1 by allowing the right-most cell boundary to move so that $H(q) \neq 0$. We do not consider proliferation, giving $R(q) = 0$ in (3.7).

The equation learning procedure for this case study is similar to Case Study 1, namely we expand $D(q)$ as in (3.9) and constrain $D(q) \geq 0$. In addition to learning $D(q)$, we need to learn $H(q)$ and the evolution equation describing the free boundary. In (3.7), this evolution equation is given by a conservation statement, $q\mathrm{d}L/\mathrm{d}t = -D(q)\partial q/\partial x$ with $q = q(L(t), t)$. Here we treat this moving boundary condition more generally by introducing a function $E(t)$ so that

$$q\frac{\mathrm{d}L}{\mathrm{d}t} = -E(q)\frac{\partial q}{\partial x} \tag{3.17}$$

at $x = L(t)$ for $t > 0$. While (3.17) could lead to local loss of conservation at the moving boundary, our approach is to for the possibility that coefficients in $D(q)$ and $E(q)$ differ and to explore the extent to which this is true, or otherwise, according to our equation learning procedure. We constrain $E(q) \geq 0$ so that (3.17) makes sense for our problem and we expand $D(q)$, $H(q)$, and $E(q)$ as follows

$$D(q) = \sum_{i=1}^{d} \theta_i^d \varphi_i^d(q), \quad H(q) = \sum_{i=1}^{h} \theta_i^h \varphi_i^h(q), \quad E(q) = \sum_{i=1}^{h} \theta_i^e \varphi_i^e(q). \tag{3.18}$$

The matrix system for $\boldsymbol{\theta}^d = (\theta_1^d, \ldots, \theta_d^d)^{\mathsf{T}}$ is the same as it was in Case Study 1 in (3.12), which we now write as $\mathbf{A}^d \boldsymbol{\theta}^d = \mathbf{b}^d$ with $\mathbf{A}^d \in \mathbb{R}^{n(M-1)\times d}$ and $\mathbf{b}^d \in \mathbb{R}^{n(M-1)\times 1}$ given by $\mathbf{A}$ and $\mathbf{b}$ in (3.12), and we can construct two other independent matrix systems for $\boldsymbol{\theta}^h = (\theta_1^h, \ldots, \theta_h^h)^{\mathsf{T}}$ and $\boldsymbol{\theta}^e = (\theta_1^e, \ldots, \theta_e^e)^{\mathsf{T}}$. To construct these matrix systems, for a given boundary point (x_{nj}, t_j) we write

$$\frac{\partial q_{nj}}{\partial x} = \sum_{k=1}^{h} \theta_k^h \varphi_k^h(q_{nj}), \quad q_{nj}\frac{\mathrm{d}L_j}{\mathrm{d}t} = -\frac{\partial q_{nj}}{\partial x} \sum_{k=1}^{e} \theta_k^e \varphi_k^e(q_{nj}), \tag{3.19}$$

where $L_j = x_{nj}$ is the position of the leading edge at $t = t_j$. In (3.19) we assume that L_j can be approximated by a smooth function so that $\mathrm{d}L_j/\mathrm{d}t$ can be defined. With (3.19) we have $\mathbf{A}^h \boldsymbol{\theta}^h = \mathbf{b}^h$ and $\mathbf{A}^e \boldsymbol{\theta}^e = \mathbf{b}^e$, where

$$\mathbf{A}^h = \begin{bmatrix} \varphi_1^h(q_{12}) & \cdots & \varphi_h^h(q_{12}) \\ \vdots & \ddots & \vdots \\ \varphi_1^h(q_{nM}) & \cdots & \varphi_h^h(q_{nM}) \end{bmatrix}, \quad \mathbf{b}^h = \begin{bmatrix} \dfrac{\partial q_{12}}{\partial x} \\ \vdots \\ \dfrac{\partial q_{nM}}{\partial x} \end{bmatrix} \tag{3.20}$$

with $\mathbf{A}^h \in \mathbb{R}^{(M-1)\times h}$ and $\mathbf{b}^h \in \mathbb{R}^{(M-1)\times 1}$, and

$$\mathbf{A}^e = \begin{bmatrix} \varphi_1^e(q_{12})\dfrac{\partial q_{n2}}{\partial x} & \cdots & \varphi_e^e(q_{12})\dfrac{\partial q_{n2}}{\partial x} \\ \vdots & \ddots & \vdots \\ \varphi_1^e(q_{nM})\dfrac{\partial q_{nM}}{\partial x} & \cdots & \varphi_e^e(q_{nM})\dfrac{\partial q_{nM}}{\partial x} \end{bmatrix}, \quad \mathbf{b}^e = -\begin{bmatrix} q_{n2}\dfrac{\mathrm{d}L_2}{\mathrm{d}t} \\ \vdots \\ q_{nM}\dfrac{\mathrm{d}L_M}{\mathrm{d}t} \end{bmatrix} \tag{3.21}$$

with $\mathbf{A}^e \in \mathbb{R}^{(M-1)\times e}$ and $\mathbf{b}^e \in \mathbb{R}^{(M-1)\times 1}$. Then, writing

$$\begin{aligned} \mathbf{A} &= \operatorname{diag}(\mathbf{A}^d, \mathbf{A}^h, \mathbf{A}^e) \in \mathbb{R}^{(n+2)(M-1)\times(d+h+e)}, \\ \mathbf{b} &= \begin{bmatrix} \mathbf{b}^d \\ \mathbf{b}^h \\ \mathbf{b}^e \end{bmatrix} \in \mathbb{R}^{(n+2)(M-1)\times 1}, \end{aligned} \tag{3.22}$$

we obtain

$$\mathbf{A}\boldsymbol{\theta} = \mathbf{b}, \quad \boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}^d \\ \boldsymbol{\theta}^h \\ \boldsymbol{\theta}^e \end{bmatrix} \in \mathbb{R}^{(d+h+e)\times 1}. \tag{3.23}$$

The solution of $\mathbf{A}\boldsymbol{\theta} = \mathbf{b}$ is the combined solution of the individual linear systems as $\mathbf{A}$ is block diagonal. Estimates for $\boldsymbol{\theta}^d$, $\boldsymbol{\theta}^h$, and $\boldsymbol{\theta}^e$ are independent, which demonstrates the modularity of our approach, where these additional features, in particular the leading edge, are just an extra independent component of our procedure in addition to the procedure for estimating $D(q)$.

In addition to the new matrix system $\mathbf{A}\boldsymbol{\theta} = \mathbf{b}$ in (3.23), we augment the loss function (3.14) to incorporate information about the location of the moving boundary. Letting $L(t; \boldsymbol{\theta})$ denote the leading edge from the solution of the PDE (3.7) with parameters $\boldsymbol{\theta}$, the loss function is

$$\begin{aligned} \underbrace{\mathcal{L}(\boldsymbol{\theta})}_{\text{loss}} = \overbrace{\log\left[\frac{1}{n(M-1)} \sum_{j=2}^{M} \sum_{i=1}^{n} \left(\frac{q_{ij} - q\left(x_{ij}, t_j; \boldsymbol{\theta}\right)}{q_{ij}}\right)^2\right]}^{\text{density goodness of fit}} \\ \underbrace{+ \log\left[\frac{1}{M-1} \sum_{j=2}^{M} \left(\frac{L_j - L\left(t_j; \boldsymbol{\theta}\right)}{L_j}\right)^2\right]}_{\text{leading edge goodness of fit}} + \underbrace{\|\boldsymbol{\theta}\|_0}_{\text{model complexity}}. \end{aligned} \tag{3.24}$$

Let us now apply our stepwise equation learning procedure with (3.23) and (3.24). We consider the data from Figure 3.2, where we know in advance that the continuum limit with $D(q) = 50/q^2$, $H(q) = 2q^2 - 0.4q^3$, and $E(q) = 50/q^2$ is accurate. The expansions we use for $D(q)$, $H(q)$, and

$E(q)$ are given by

$$
\begin{aligned}
D(q) &= \frac{\theta_1^d}{q} + \frac{\theta_2^d}{q^2} + \frac{\theta_3^d}{q^3}, \\
H(q) &= \theta_1^h q + \theta_2^h q^2 + \theta_3^h q^3 + \theta_4^h q^4 + \theta_5^h q^5, \\
E(q) &= \frac{\theta_1^e}{q} + \frac{\theta_2^e}{q^2} + \frac{\theta_3^e}{q^3}.
\end{aligned}
\tag{3.25}
$$

With these expansions, we expect to learn the vectors $\boldsymbol{\theta}^d = (0, 50, 0)^\mathsf{T}$, $\boldsymbol{\theta}^h = (0, 2, -0.4, 0, 0)^\mathsf{T}$, and $\boldsymbol{\theta}^e = (0, 50, 0)^\mathsf{T}$. We initially consider saving the solution at $M = 1000$ equally spaced times between $t_1 = 0$ and $t_M = 100$, and using matrix pruning so that only points whose densities fall within the 35% and 65% density quantiles are included. The results with this configuration are shown in Table 3.4, where we see that we are only able to learn $H(q) = E(q) = 0$ and $D(q) = 25.06/q^3$. This outcome highlights the importance of choosing an appropriate time interval, since Figure 3.2(b) indicates that mechanical relaxation takes place over a relative short interval which means that working with data in $0 < t \leq 100$ can lead to a poor outcome.

Table 3.4: Stepwise equation learning results for Case Study 2: Free boundaries, using the basis expansions (3.25), saving the results at $M = 1000$ equally spaced times between $t_1 = 0$ and $t_M = 100$, pruning so that densities outside of the 35% and 65% density quantiles are not included, and starting with all terms inactive. Coefficients highlighted in blue show the coefficient chosen to be removed or added at the corresponding step.

Step	θ_1^d	θ_2^d	θ_3^d	θ_1^h	θ_2^h	θ_3^h	θ_4^h	θ_5^h	θ_1^e	θ_2^e	θ_3^e	**Loss**
1	0.00	0.00	**0.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-1.40
2	0.00	0.00	25.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.40

We proceed by restricting our data collection to $0 \leq t \leq 15$, now saving the solution at $M = 200$ equally spaced times between $t_1 = 0$ and $t_M = 15$. Keeping the same quantiles for the matrix pruning, the new results are shown in Table 3.5 and Figure 3.6. We see that the densities and leading edges are accurate for small time, but the learned mechanisms do not extrapolate as well for $t \geq 15$, for example $L(t)$ in Figure 3.6(b) does not match the discrete data. To address this issue, we can further limit the information that we include in our matrices, looking to only include boundary points where $\mathrm{d}L/\mathrm{d}t$ is neither too large not too small. We implement this by excluding all points (x_{nj}, t_j) from the construction of

$(\mathbf{A}^e, \mathbf{b}^e)$ in (3.21) such that $\mathrm{d}L_j/\mathrm{d}t$ is outside of the 10% or 90% quantiles of the vector $(\mathrm{d}L_2/\mathrm{d}t, \ldots, \mathrm{d}L_M/\mathrm{d}t)$, called the *velocity quantiles*.

Table 3.5: Stepwise equation learning results for Case Study 2: Free boundaries, using the basis expansions (3.25), saving the results at $M = 200$ equally spaced times between $t_1 = 0$ and $t_M = 15$, pruning so that densities outside of the 35% and 65% density quantiles are not included, and starting with all terms inactive. Coefficients highlighted in blue show the coefficient chosen to be removed or added at the corresponding step.

Step	θ_1^d	θ_2^d	θ_3^d	θ_1^h	θ_2^h	θ_3^h	θ_4^h	θ_5^h	θ_1^e	θ_2^e	θ_3^e	**Loss**
1	0.00	0.00	0.00	0.00	**0.00**	0.00	0.00	0.00	0.00	0.00	0.00	-3.37
2	0.00	0.00	0.00	0.00	-0.03	0.00	0.00	0.00	**0.00**	0.00	0.00	-2.37
3	0.00	**0.00**	0.00	0.00	-0.03	0.00	0.00	0.00	8.74	0.00	0.00	-3.68
4	0.00	47.38	0.00	**0.00**	-0.03	0.00	0.00	0.00	8.74	0.00	0.00	-4.02
5	0.00	47.38	0.00	8.41	-1.69	0.00	0.00	0.00	8.74	0.00	0.00	-8.14

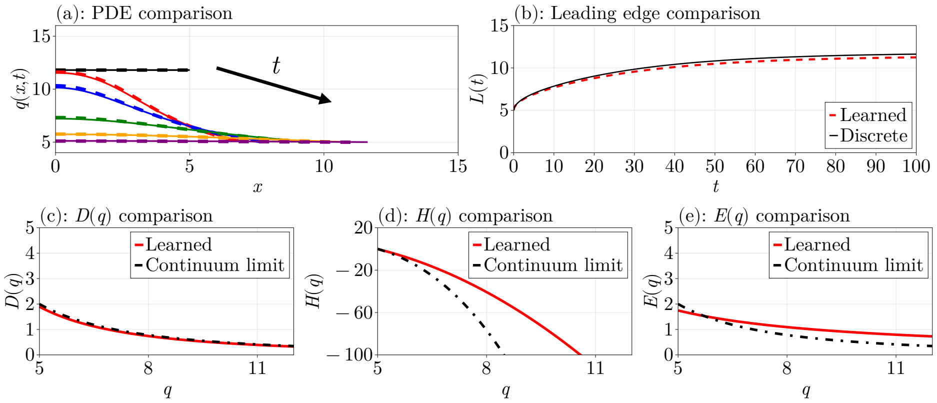

Figure 3.6: Stepwise equation learning results from Table 3.5 for Case Study 2: Free boundaries. (a) Comparisons of the discrete density profiles (solid curves) with those learned from PDEs obtained from the results in Table 3.5 (dashed curves), plotted at the times $t = 0, 5, 10, 25, 50, 100$ in black, red, blue, green, orange, and purple, respectively. The arrow shows the direction of increasing time. (b) As in (a), except comparing the leading edges. (c)–(e) are comparisons of the learned forms of $D(q)$, $H(q)$, and $E(q)$ with the forms from the continuum limit (3.8).

Implementing thresholding on $\mathrm{d}L/\mathrm{d}t$ leads to the results presented in Figure 3.7. We see that the learned densities and leading edges are both visually indistinguishable from the discrete data. Since $H(q)$ and $E(q)$ are only ever evaluated at $x = L(t)$, and $q(L(t), t) \approx 5$ for $t > 0$, we see that $H(q)$ and $E(q)$ only match the continuum limit at $q \approx 5$, which means that our learned continuum limit model conserves mass and is consistent with the traditional coarse-grained continuum limit, as expected. We discuss in

3.E how we can enforce $D(q) = E(q)$ to guarantee conservation mass from the outset, however our approach in Figure 3.7 is more general in the sense that our learned continuum limit is obtained without making any *a priori* assumptions about the form of $E(q)$.

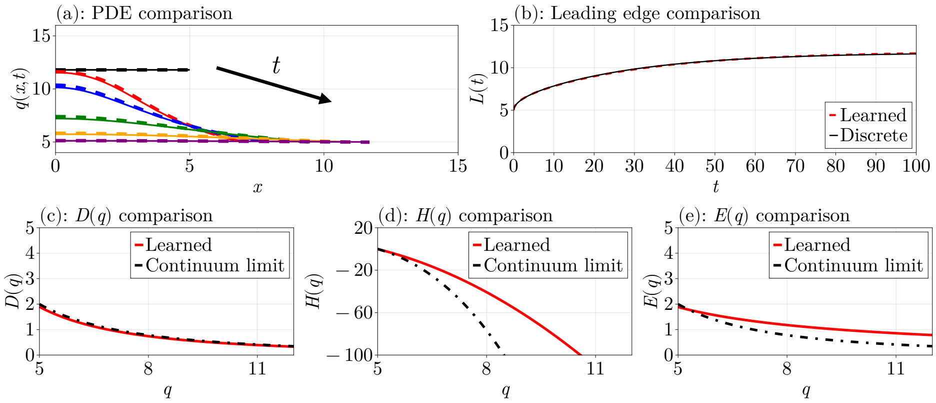

Figure 3.7: Stepwise equation learning results from Table 3.5 for Case Study 2: Free boundaries, except also using matrix pruning on $(\mathbf{A}_3, \mathbf{b}_3)$ so points where $\mathrm{d}L_j/\mathrm{d}t$ falls outside of the 10% and 90% velocity quantiles are excluded, giving $\theta_1^e = 9.42$ rather than 8.74. (a) Comparisons of the discrete density profiles (solid curves) with those from the learned PDE (dashed curves), plotted at the times $t = 0, 5, 10, 25, 50, 100$ in black, red, blue, green, orange, and purple, respectively. The arrow shows the direction of increasing time. (b) As in (a), except comparing the leading edges. (c)–(e) are comparisons of the learned forms of $D(q)$, $H(q)$, and $E(q)$ with the forms from the continuum limit (3.8).

3.4.3 Case Study 3: Fixed boundaries with proliferation

Case Study 3 is identical to Case Study 1 except that we incorporate cell proliferation, implying $R(q) \neq 0$ in (3.7). This case is more complicated than with mechanical relaxation only, as we have to consider how we combine the repeated realisations to capture the average density data as well. For this work, we average over each realisation at each time using linear interpolants as described in 3.D. This averaging procedure gives n_k points $\bar{x}_{ij}$ between $x = 0$ and $x = 30$ at each time t_j, $j = 1, \ldots, M$, with corresponding density value $\bar{q}_{ij}$. The quantities $\bar{x}_{ij}$ and $\bar{q}_{ij}$ play the same role as x_{ij} and q_{ij} in the previous case studies.

To apply equation learning we note there is no moving boundary, giving

$H(q) = 0$ in (3.8). We proceed by expanding $D(q)$ and $R(q)$ as follows

$$D(q) = \sum_{i=1}^{d} \theta_i^d \varphi_i^d(q), \quad R(q) = \sum_{i=1}^{r} \theta_i^r \varphi_i^r(q), \tag{3.26}$$

with the aim of estimating $\boldsymbol{\theta}^d = (\theta_1^d, \ldots, \theta_d^d)^{\mathsf{T}}$ and $\boldsymbol{\theta}^r = (\theta_1^r, \ldots, \theta_r^r)^{\mathsf{T}}$, again constraining $D(q) \geq 0$. We expand the PDE from (3.10), as in Section 3.4(a), and the only difference is the additional term $\sum_{m=1}^{r} \varphi_m^r(\bar{q}_{ij})\theta_m^r$ for each point $(\bar{x}_{ij}, t_j)$. Thus, we have the same matrix as in Section 3.4(a), denoted $\mathbf{A}^d \in \mathbb{R}^{n_k(M-1)\times d}$, and a new matrix $\mathbf{A}^r \in \mathbb{R}^{n_k(M-1)\times r}$ whose row corresponding to the point $(\bar{x}_{ij}, t_j)$ is given by

$$\mathbf{a}_{ij}^r = \begin{bmatrix} \varphi_1^r(\bar{q}_{ij}) & \cdots & \varphi_r^r(\bar{q}_{ij}) \end{bmatrix} \in \mathbb{R}^{1\times r}, \tag{3.27}$$

so that the coefficient matrix $\mathbf{A}$ is now

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}^d & \mathbf{A}^r \end{bmatrix} \in \mathbb{R}^{n_k(M-1)\times(d+r)}. \tag{3.28}$$

The corresponding entry for the point $(\bar{x}_{ij}, t_j)$ in $\mathbf{b} \in \mathbb{R}^{n_k(M-1)\times 1}$ is $\partial\bar{q}_{ij}/\partial t$. Notice that this additional term in the PDE adds an extra block to the matrix without requiring a significant coupling with the existing equations from the simpler problem without proliferation. Thus, we estimate our coefficient vectors using the system

$$\mathbf{A}\boldsymbol{\theta} = \mathbf{b}, \quad \boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}^d \\ \boldsymbol{\theta}^r \end{bmatrix} \in \mathbb{R}^{(d+r)\times 1}. \tag{3.29}$$

We can take exactly the same stepwise procedure as in Section 3.4(a), except now the loss function (3.14) uses n_k, $\bar{q}_{ij}$, and $\bar{x}_{ij}$ rather than n, q_{ij}, and x_{ij}, respectively.

Accurate continuum limit

Let us now apply these ideas to our data from Figure 3.2, where we know that the continuum limit with $D(q) = 50/q^2$ and $R(q) = 0.15q - 0.01q^2$ is accurate. The expansions we use for $D(q)$ and $R(q)$ are given by

$$D(q) = \frac{\theta_1^d}{q} + \frac{\theta_2^d}{q^2} + \frac{\theta_3^d}{q^3}, \quad R(q) = \theta_1^r q + \theta_2^r q^2 + \theta_3^r q^3 + \theta_4^r q^4 + \theta_5^r q^5, \tag{3.30}$$

and we expect to learn $\boldsymbol{\theta}^d = (0, 50, 0)^\mathsf{T}$ and $\boldsymbol{\theta}^r = (0.15, -0.01, 0, 0, 0)^\mathsf{T}$. We average over 1000 identically-prepared realisations, saving the solutions at $M = 501$ equally spaced times between $t_1 = 0$ and $t_M = 50$ with $n_k = 50$ knots for averaging. For this problem, and for Case Study 4 discussed later, we find that working with 1000 identically-prepared realisations of the stochastic models leads to sufficiently smooth density profiles. As discussed in 3.F, the precise number of identically-prepared realisations is not important provided that the number is sufficiently large; when not enough realisations are taken, the results are inconsistent across different sets of realisations and will fail to identify the average behaviour from the learned model. We also use matrix pruning so that we only include points whose densities fall within the 10% and 90% density quantiles, as done in Section 3.4(a). The results we obtain are shown in Table 3.6, starting with all coefficients active.

Table 3.6: Stepwise equation learning results for Case Study 3: Fixed boundaries with proliferation, where the continuum limit is accurate, using the basis expansions (3.30), saving the results at $M = 501$ equally spaced times between $t_1 = 0$ and $t_M = 50$, averaging across 1000 realisations with $n_k = 50$ knots, pruning so that densities outside of the 10% and 90% density quantiles are not included, and starting with all diffusion and reaction coefficients active. Coefficients highlighted in blue show the coefficient chosen to be removed or added at the corresponding step.

Step	θ_1^d	θ_2^d	θ_3^d	θ_1^r	θ_2^r	θ_3^r $(\times 10^{-4})$	θ_4^r $(\times 10^{-5})$	θ_5^r $(\times 10^{-7})$	**Loss**
1	-11.66	147.43	**-191.51**	0.13	-0.00	-0.00	5.83	-11.30	∞
2	-2.24	60.86	0.00	0.13	-0.00	-5.72	2.62	-3.49	-0.71
3	**-2.25**	60.90	0.00	0.14	-0.01	0.00	-1.25	5.95	-1.92
4	0.00	52.95	0.00	0.14	-0.01	0.00	-1.36	6.49	-3.35
5	0.00	53.02	0.00	0.15	-0.01	0.00	0.00	0.32	-4.98
6	0.00	52.97	0.00	0.15	-0.01	0.00	0.00	0.00	-5.70

Table 3.6 shows that we find the vectors $\boldsymbol{\theta}^d = (0, 52.97, 0)^\mathsf{T}$ and $\boldsymbol{\theta}^r = (0.15, -0.010, 0, 0, 0)^\mathsf{T}$, which are both very close to the continuum limit. Figure 3.8 visualises these results, showing that the PDE solutions with the learned $D(q)$ and $R(q)$ match the discrete densities, and the mechanisms that we do learn are visually indistinguishable with the continuum limit functions (3.8) as shown in Figure 3.8(b)–(c).

Inaccurate continuum limit

We now extend the problem so that the continuum limit is no longer accurate, taking $k = 1/5$ to be consistent with Figure 3.3(a). Using the same

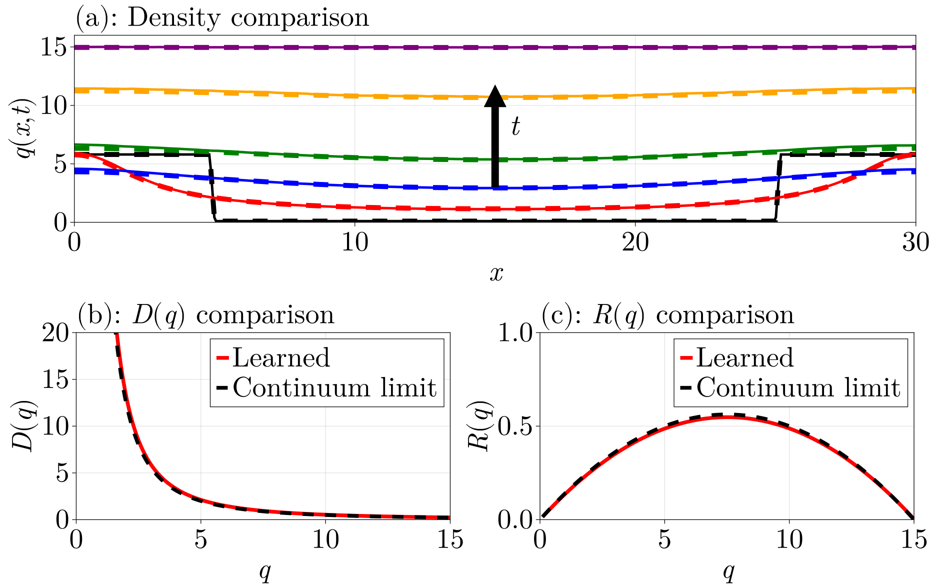

Figure 3.8: Stepwise equation learning results for Case Study 3: Fixed boundaries with proliferation, where the continuum limit is accurate. (a) Comparisons of the discrete density profiles (solid curves) with those learned from PDEs obtained from the results in Table 3.6 (dashed curves), plotted at the times $t = 0, 1, 5, 10, 20, 50$ in black, red, blue, green, orange, and purple, respectively. The arrow shows the direction of increasing time. (b)–(c) are comparisons of $D(q)$ and $R(q)$ with the forms from the continuum limit (3.8).

basis expansions in (3.30), we save the solution at $M = 751$ equally spaced times between $t_1 = 0$ and $t_M = 75$, averaging over 1000 realisations with $n_k = 200$. We find that we need to use the 25% and 75% density quantiles rather than the 10% and 90% density quantiles, as in the previous example, to obtain results in this case. With this configuration, the results we find are shown in Table 3.7 and Figure 3.9.

Results in Table 3.7 show $\boldsymbol{\theta}^d = (0, 0.12, 0)^{\mathsf{T}}$, which is reasonably close to the continuum limit with $(0, 0.2, 0)^{\mathsf{T}}$. The reaction vector, for which the continuum limit is $(0.15, -0.01, 0, 0, 0)^{\mathsf{T}}$ so that $R(q)$ is a quadratic, is now given by $\boldsymbol{\theta}^r = (0.16, -0.02, 7.49 \times 10^{-4}, -1.69 \times 10^{-5}, 0)^{\mathsf{T}}$, meaning the learned $R(q)$ is a quartic. Figure 3.9 compares the averaged discrete densities with the solution of the learned continuum limit model. Figure 3.9(c) compares the learned source term with the continuum limit. While both terms are visually indistinguishable at small densities, we see that the two source terms differ at high densities, with the learned carrying capacity

Table 3.7: Stepwise equation learning results for Case Study 3: Fixed boundaries with proliferation, where the continuum limit is inaccurate, using the basis expansions (3.30), saving the results at $M = 751$ equally spaced times between $t_1 = 0$ and $t_M = 75$, averaging across 1000 realisations with $n_k = 200$ knots, pruning so that densities outside of the 25% and 75% density quantiles are not included, and starting with all diffusion and reaction coefficients inactive. Coefficients highlighted in blue show the coefficient chosen to be removed or added at the corresponding step.

Step	θ_1^d	θ_2^d	θ_3^d	θ_1^r	θ_2^r	θ_3^r $(\times 10^{-4})$	θ_4^r $(\times 10^{-5})$	θ_5^r	**Loss**
1	0.00	0.00	0.00	**0.00**	0.00	0.00	0.00	0.00	-0.33
2	0.00	0.00	0.00	0.02	**0.00**	0.00	0.00	0.00	0.51
3	0.00	**0.00**	0.00	0.11	-0.01	0.00	0.00	0.00	0.20
4	0.00	0.11	0.00	0.11	-0.01	**0.00**	0.00	0.00	-0.04
5	0.00	0.12	0.00	0.13	-0.01	1.59	**0.00**	0.00	-0.46
6	0.00	0.12	0.00	0.16	-0.02	7.49	-1.69	0.00	-1.13

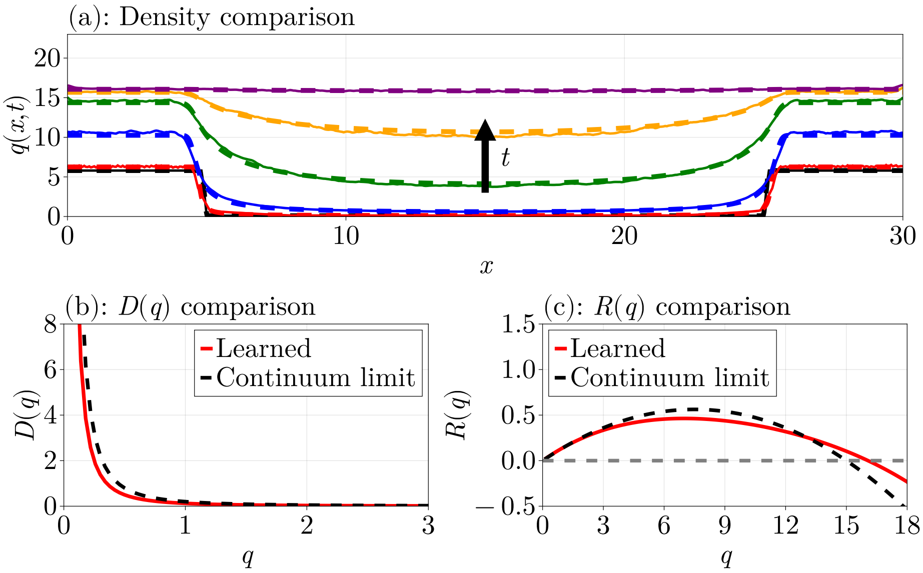

Figure 3.9: Stepwise equation learning results for Case Study 3: Fixed boundaries with proliferation, where the continuum limit is inaccurate. (a) Comparisons of the discrete density profiles (solid curves) with those learned from PDEs obtained from the results in Table 3.6 (dashed curves), plotted at the times $t = 0, 1, 10, 25, 40, 75$ in black, red, blue, green, orange, and purple, respectively. The arrow shows the direction of increasing time. (b)–(c) are comparisons of $D(q)$ and $R(q)$ with the forms from the continuum limit (3.8).

density, where $R(q) = 0$, reduced relative to the continuum limit. This is consistent with previous results [40].

3.4.4 Case Study 4: Free boundaries with proliferation

Case Study 4 is identical to Case Study 2 except that we now introduce proliferation into the discrete model so that $R(q) \neq 0$ in (3.7). First, as in Case Study 3 and as described in 3.D, we average our data across each realisation from our discrete model. This averaging provides us with points $\bar{x}_{ij}$ between $x = 0$ and $x = \bar{L}_j$ at each time t_j, $j = 1, \ldots, M$, where $\bar{L}_j$ is the average leading edge at $t = t_j$, with corresponding density values $\bar{q}_{ij}$, where $i = 1, \ldots, n_k$ and n_k is the number of knots to use for averaging. We expand the functions $D(q)$, $R(q)$, $H(q)$, and $E(q)$ as

$$\begin{aligned} D(q) &= \sum_{i=1}^{d} \theta_i^d \varphi_i^d(q), & R(q) &= \sum_{i=1}^{r} \theta_i^r \varphi_i^r(q), \\ H(q) &= \sum_{i=1}^{h} \theta_i^h \varphi_i^h(q), & E(q) &= \sum_{i=1}^{e} \theta_i^e \varphi_i^e(q), \end{aligned} \tag{3.31}$$

again restricting $D(q), E(q) \geq 0$. The function $E(q)$ is used in the moving boundary condition in (3.7), as in (3.17). The matrix $\mathbf{A}$ and vector $\mathbf{b}$ are given by

$$\begin{aligned} \mathbf{A} &= \operatorname{diag}(\mathbf{A}^{dr}, \mathbf{A}^h, \mathbf{A}^e) \in \mathbb{R}^{n_k(M-1)\times(d+r+h+e)}, \\ \mathbf{b} &= \begin{bmatrix} \mathbf{b}^{dr} \\ \mathbf{b}^h \\ \mathbf{b}^e \end{bmatrix} \in \mathbb{R}^{n_k(M-1)}, \end{aligned} \tag{3.32}$$

where $\mathbf{A}^{dr} = [\mathbf{A}^d\ \mathbf{A}^r]$ as defined in (3.28), $\mathbf{A}^h$ and $\mathbf{A}^e$ are the matrices from (3.20) and (3.21), respectively, and similarly for $\mathbf{b}^{dr} = \partial \mathbf{q}/\partial t$, $\mathbf{b}^h$, and $\mathbf{b}^e$ from (3.12), (3.20), and (3.21), respectively. Thus,

$$\mathbf{A}\boldsymbol{\theta} = \mathbf{b}, \quad \boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}^d \\ \boldsymbol{\theta}^r \\ \boldsymbol{\theta}^h \\ \boldsymbol{\theta}^e \end{bmatrix} \in \mathbb{R}^{(d+r+h+e)\times 1}. \tag{3.33}$$

Similar to Case Study 2, the coefficients for each mechanism are independent, except for $\boldsymbol{\theta}^d$ and $\boldsymbol{\theta}^r$. The loss function we use is the loss function from (3.24).

With this problem, it is difficult to learn all mechanisms simultaneously, especially as mechanical relaxation and proliferation occur on different time scales since mechanical relaxation dominates in the early part of the simulation, whereas both proliferation and mechanical relaxation play a role at later times. This means $D(q)$ and $R(q)$ cannot be estimated over the entire time range as was done in Case Study 3. To address this we take a sequential learning procedure to learn these four mechanisms using four distinct time intervals I^d, I^e, I^h, and I^r:

1. Fix $R(q) = H(q) = E(q) = 0$ and learn $\boldsymbol{\theta}^d$ over $t \in I^d$, solving $\mathbf{A}^d\boldsymbol{\theta}^d = \mathbf{b}^{dr}$.
2. Fix $R(q) = H(q) = 0$ and $\boldsymbol{\theta}^d$ and learn $\boldsymbol{\theta}^e$ over $t \in I^e$, solving $\mathbf{A}^e\boldsymbol{\theta}^e = \mathbf{b}^e$.
3. Fix $R(q) = 0$, $\boldsymbol{\theta}^d$, and $\boldsymbol{\theta}^e$ and learn $\boldsymbol{\theta}^h$ over $t \in I^h$, solving $\mathbf{A}^h\boldsymbol{\theta}^h = \mathbf{b}^h$.
4. Fix $\boldsymbol{\theta}^d$, $\boldsymbol{\theta}^e$, and $\boldsymbol{\theta}^h$ and learn $\boldsymbol{\theta}^r$ over $t \in I^r$, solving $\mathbf{A}^r\boldsymbol{\theta}^r = \mathbf{b}^{dr} - \mathbf{A}^d\boldsymbol{\theta}^d$.

In these steps, solving the system $\mathbf{A}\boldsymbol{\theta} = \mathbf{b}$ means to apply our stepwise procedure to this system; for these problems, we start each procedure with no active coefficients. The modularity of our approach makes this sequential learning approach straightforward to implement. For these steps, the interval I^d must be over sufficiently small times so that proliferation does not dominate, noting that fixing $R(q) = 0$ will not allow us to identify any proliferation effects when estimating the parameters. This is less relevant for I^h and I^e as the estimates of $\boldsymbol{\theta}^h$ and $\boldsymbol{\theta}^e$ impact the moving boundary only.

3.4.5 Accurate continuum limit

We apply this procedure to data from Figure 3.2, where the continuum limit is accurate with $D(q) = 50/q^2$, $R(q) = 0.15q - 0.01q^2$, $H(q) = 2q^2 - 0.4q^3$, and $E(q) = 50/q^2$. The expansions we use are

$$\begin{aligned} D(q) &= \frac{\theta_1^d}{q} + \frac{\theta_2^d}{q^2} + \frac{\theta_3^d}{q^3}, \\ R(q) &= \theta_1^r q + \theta_2^r q^2 + \theta_3^r q^3 + \theta_4^r q^4 + \theta_5^r q^5, \\ H(q) &= \theta_1^h q + \theta_2^h q^2 + \theta_3^h q^3 + \theta_4^h q^4 + \theta_5^h q^5, \\ E(q) &= \frac{\theta_1^e}{q} + \frac{\theta_2^e}{q^2} + \frac{\theta_3^e}{q^3}. \end{aligned} \tag{3.34}$$

With (3.34), we expect to learn $\boldsymbol{\theta}^d = (0, 50, 0)^{\mathsf{T}}$, $\boldsymbol{\theta}^r = (0.15, -0.01, 0, 0, 0)^{\mathsf{T}}$, $\boldsymbol{\theta}^h = (0, 2, -0.4, 0, 0, 0)^{\mathsf{T}}$, and $\boldsymbol{\theta}^e = (0, 50, 0)^{\mathsf{T}}$. We average the data over 1000 realisations. For saving the solution, the time intervals we use are $I^d = [0, 0.1]$, $I^e = [0, 5]$, $I^h = [5, 10]$, and $I^r = [10, 50]$, with 25, 50, 100, and 250 time points inside each time interval for saving. For interpolating the solution to obtain the averages, we use $n_k = 25$, $n_k = 50$, $n_k = 100$, and $n_k = 50$ over I^d, I^e, I^h, and I^r, respectively.

To now learn the mechanisms, we apply the sequential procedure described for learning them one at a time. For each problem, we apply pruning so that points outside of the 10% and 90% density quantiles or the 20% and 80% velocity quantiles are not included. We find that $\boldsymbol{\theta}^d = (0, 49.60, 0)^{\mathsf{T}}$, $\boldsymbol{\theta}^e = (0, 49.70, 0)^{\mathsf{T}}$, $\boldsymbol{\theta}^h = (-0.0084, 0, 0, -0.0011, 0)^{\mathsf{T}}$, and $\boldsymbol{\theta}^r = (0.15, -0.010, 0, 0, 0)^{\mathsf{T}}$. The results with all these learned mechanisms are shown in Figure 3.10. We see from the comparisons in Figure 3.10(a)–(b) that the PDE results from the learned mechanisms are nearly indistinguishable from the discrete densities. Similar to Case Study 2, $H(q)$ only matches the continuum limit at $q(L(t), t)$. Note also that the solutions in Figure 3.10(a) go up to $t = 100$, despite the stepwise procedure considering only times up to $t = 50$.

3.4.6 Inaccurate continuum limit

We now consider data from Figure 3.3(b) where the continuum limit is inaccurate. Here, $k = 1/5$ and the continuum limit vectors are $\boldsymbol{\theta}^d = (0, 0.2, 0)^{\mathsf{T}}$, $\boldsymbol{\theta}^r = (0.15, -0.01, 0, 0, 0)^{\mathsf{T}}$, $\boldsymbol{\theta}^h = (0, 2, -0.4, 0, 0, 0)^{\mathsf{T}}$, and $\boldsymbol{\theta}^e = (0, 0.2, 0)^{\mathsf{T}}$. Using the same procedures and expansions as Figure 3.10, we average the data over 1000 realisations. The time intervals we use are $I^d = [0, 2]$, $I^e = [2, 10]$, $I^h = [10, 20]$, and $I^r = [20, 50]$, using 20 time points for I^d and 200 time points for I^e, I^h, and I^r. We use $n_k = 50$ knots for averaging the solution over I^d, and $n_k = 100$ knots for averaging the solution over I^e, I^h, and I^r.

To apply the equation learning procedure we prune all matrices so that points outside of the 40% and 60% temporal quantiles are eliminated, where the *temporal quantiles* are the quantiles of $\partial q/\partial t$ from the averaged discrete data, and similarly for points outside of the 40% and 60% velocity quantiles. We find $\boldsymbol{\theta}^d = (0, 0.21, 0)^{\mathsf{T}}$, $\boldsymbol{\theta}^e = (0, 0.23, 0)^{\mathsf{T}}$, $\boldsymbol{\theta}^h = (-0.15, 0, 0, -0.0079, 0)^{\mathsf{T}}$, and $\boldsymbol{\theta}^r = (0.11, -0.0067, 0, 0, 0)^{\mathsf{T}}$. Inter-

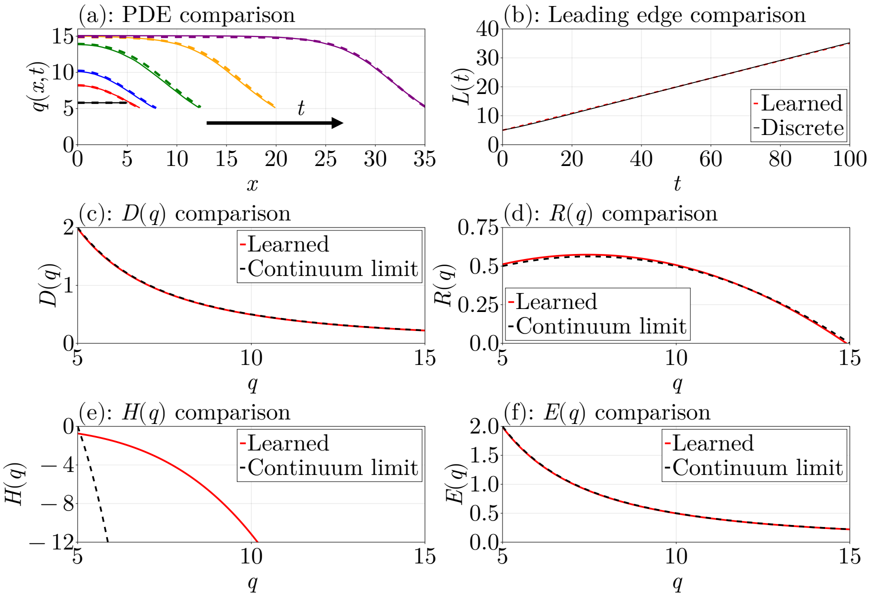

Figure 3.10: Stepwise equation learning results for Case Study 4: Free boundaries with proliferation, when the continuum limit is accurate, using the learned mechanisms with $\boldsymbol{\theta}^d = (0, 49.60, 0)^\mathsf{T}$, $\boldsymbol{\theta}^e = (0, 49.70, 0)^\mathsf{T}$, $\boldsymbol{\theta}^h = (-0.0084, 0, 0, -0.0011, 0)^\mathsf{T}$, and $\boldsymbol{\theta}^r = (0.15, -0.010, 0, 0, 0)^\mathsf{T}$. (a) Comparisons of the discrete density profiles (solid curves) with those learned from PDEs with the given $\boldsymbol{\theta}^d$, $\boldsymbol{\theta}^e$, $\boldsymbol{\theta}^h$, and $\boldsymbol{\theta}^r$ (dashed curves), plotted at the times $t = 0, 5, 10, 25, 50, 100$ in black, red, blue, green, orange, and purple, respectively. The arrow shows the direction of increasing time. (b) As in (a), except comparing the leading edges. (c)–(f) are comparisons of the learned forms of $D(q)$, $R(q)$, $H(q)$, and $E(q)$ with the forms from the continuum limit (3.8).

estingly, here we learn $R(q)$ is quadratic with coefficients that differ from the continuum limit. The results with all these learned mechanisms are shown in Figure 3.11. We see from the comparisons in Figure 3.11 that the PDE results from the learned mechanisms are visually indistinguishable from the discrete densities. Moreover, as in Figure 3.10, the learned $H(q)$ and $E(q)$ match the continuum results at $q(L(t), t)$ which confirms that the learned continuum limit conserves mass, as expected. Note also that the solutions in Figure 3.11(a) go up to $t = 250$, despite the stepwise procedure considering only times up to $t = 50$, demonstrating the extrapolation power of our method.

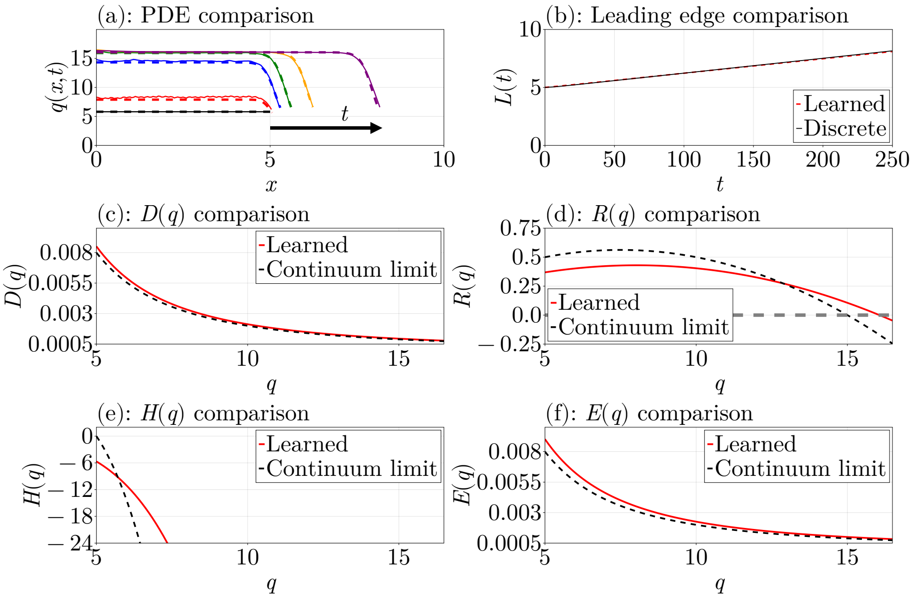

Figure 3.11: Stepwise equation learning results for Case Study 4: Free boundaries with proliferation, when the continuum limit is inaccurate, using the learned mechanisms with $\boldsymbol{\theta}^d = (0, 0.21, 0)^\mathsf{T}$, $\boldsymbol{\theta}^e = (0, 0.23, 0)^\mathsf{T}$, $\boldsymbol{\theta}^h = (-0.15, 0, 0, -0.0079, 0)^\mathsf{T}$, and $\boldsymbol{\theta}^r = (0.11, -0.0067, 0, 0, 0)^\mathsf{T}$. (a) Comparisons of the discrete density profiles (solid curves) with those learned from PDEs with the given $\boldsymbol{\theta}^d$, $\boldsymbol{\theta}^e$, $\boldsymbol{\theta}^h$, and $\boldsymbol{\theta}^r$ (dashed curves), plotted at the times $t = 0, 5, 25, 50, 100, 250$ in black, red, blue, green, orange, and purple, respectively. The arrow shows the direction of increasing time. (b) As in (a), except comparing the leading edges. (c)–(f) are comparisons of the learned forms of $D(q)$, $R(q)$, $H(q)$, and $E(q)$ with the forms from the continuum limit (3.8).

3.5 Conclusion and discussion

In this work, we presented a stepwise equation learning framework for learning continuum descriptions of discrete models describing population biology phenomena. Our approach provides accurate continuum approximations when standard coarse-grained approximations are inaccurate. The framework is simple to implement, efficient, easily parallelisable, and modular, allowing for additional components to be added into a model with minimal changes required to accommodate them into an existing procedure. In contrast to other approaches, like neural networks [49] or linear regression approaches [52], results from our procedure are interpretable in terms of the underlying discrete process. The coefficients incorporated or removed at each stage of our procedure give a sense of the influence each model term contributes to the model, giving a greater interpretation of the results, highlighting an advantage of the stepwise approach over traditional sparse regression methods [47, 48, 51]. The learned continuum descriptions from our procedure enable the discovery of new mechanisms and equations describing the data from the discrete model. For example, the discovered form of $D(q)$ can be interpreted relative to the discrete model, describing the interaction forces between neighbouring cells. In addition, we found in Case Study 4 that, when $k = 1/5$ so that the continuum limit is inaccurate, the positive root of the quadratic form of the source term $R(q)$ is greater than the mean field carrying capacity density K, as seen in Figure 3.11. This increase suggests that, when the rate of mechanical relaxation is small relative to the proliferation rates, the mean field carrying capacity density in the continuum description can be different from that in the discrete model.

We demonstrated our approach using a series of four biologically motivated case studies that incrementally build on each other, studying a discrete individual-based mechanical free boundary model of epithelial cells [38, 40–42]. In the first two case studies, we demonstrated that we can easily rediscover the continuum limit models derived by Baker et al. [38], including the equations describing the evolution of the free boundary. The last two case studies demonstrate that, when the coarse-grained models are inaccurate, our approach can learn an accurate continuum approximation. The last case study was the most complicated, with four mechanisms needing to be learned, but the modularity of our approach made it simple to apply a sequential procedure to learning the mechanisms, applying the procedure to each mechanism in sequence. Our procedure was able to recover terms that conserved mass, despite not enforcing conservation of mass explicitly. The procedure as we have described does have some limitations, such as assuming that the mechanisms are linear combinations of basis functions, which could be handled more generally by instead using nonlinear least squares [50]. The procedure may also be sensitive to the quality of the data points included in the matrices, and thus to the parameters used for the procedure. In 3.F, we discuss a parameter sensitivity study that investigates this in greater detail. In this parameter sensitivity study, we find that the most important parameters to choose are the pruning parameters. These parameters can be easily tuned thanks to the efficiency of our method, modifying each parameter in sequence and using trial and error to determine suitable parameter values.

There are many avenues for future work based on our approach. Firstly, two-dimensional extensions of our discrete model could be considered [34, 134], which would follow the same approach except the continuum problems would have to be solved using a more detailed numerical approximation [135–137]. Another avenue for exploration would be to consider applying the discrete model on a curved interface which is more realistic than considering an epithelial sheet on a flat substrate [138, 139]. Working with heterogeneous populations of cells, where parameters in the discrete model can vary between individuals in the population, is also another interesting option for future exploration [45]. Uncertainty quantification could also be considered using bootstrapping [50] or Bayesian inference [140]. Allowing for uncertainty quantification would also allow for noisy data sets to be modelled, unlike the idealised, noise-free data used in this work. Lastly, another interesting possibility for future work is to consider fitting data sets from multiple parameter sets simultaneously, including parameter sets where the continuum limit is accurate, so that a more general macroscopic model could be obtained that has the continuum limit as a special case. We emphasise that, regardless of the approach taken for future work, we believe that our flexible stepwise learning framework can form the basis of these potential future studies.

Chapter 3: Supplementary material

3.A Confidence bands for inaccurate continuum limits

In Figure 3.3, we show a series of curves for Case Study 3 and Case Study 4 with $k = 1/5$, finding that the solution to the continuum limit is no longer a good match to the data from the discrete model. Figure 3.A.1 shows the confidence bands around each of these curves, showing how the uncertainty evolves over time.

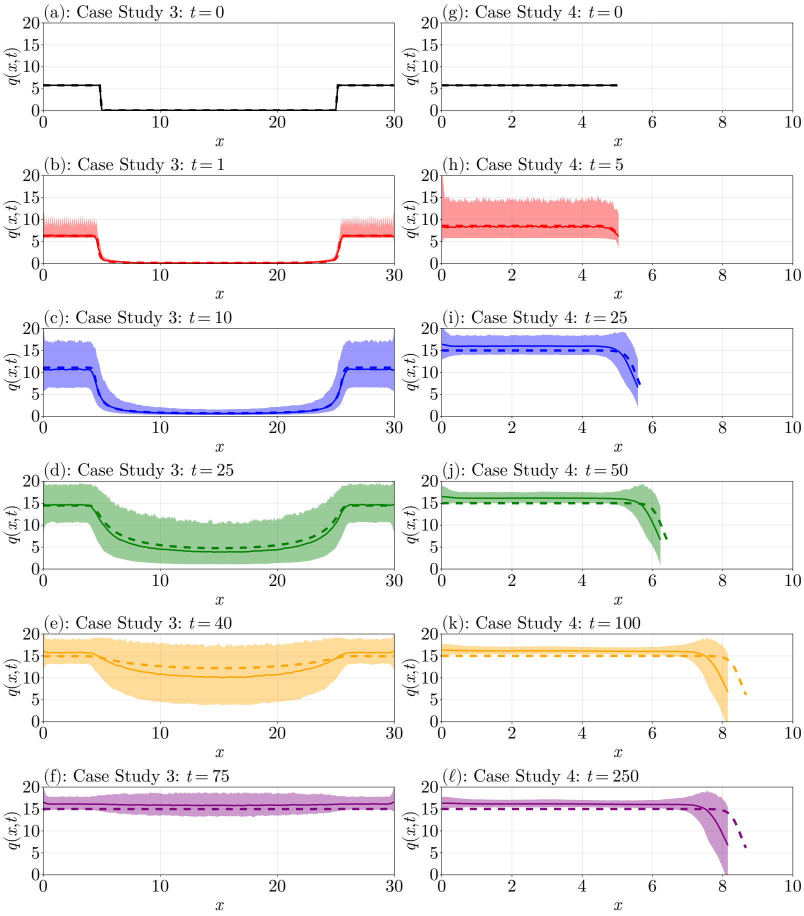

Figure 3.A.1: Complementary figure to Figure 3.3, showing inaccurate continuum limits for Case Study 3 (left column) and Case Study 4 (right column). The solid curves are the discrete densities from Equation (3.5) and the dashes curves are solutions to the continuum limit problem in Equation (3.7). The shaded regions show 95% confidence bands from the mean discrete curves at each time shown.

3.B Discrete densities at the boundaries

In (3.6), we give the following formulae for computing the cell densities from our discrete model at the boundary:

$$\begin{aligned} q_1(t) &= \frac{2}{x_2(t) - x_1(t)} - \frac{2}{x_3(t) - x_1(t)}, \\ q_n(t) &= \frac{2}{x_n(t) - x_{n-1}(t)} - \frac{2}{x_n(t) - x_{n-2}(t)}, \end{aligned} \tag{3.35}$$

noting also that $x_1(t) = 0$ in this work. In this section, we derive the expressions for $q_1(t)$ and $q_n(t)$ and show the need for these complicated expressions over those from Baker et al. [38], namely $q_1(t) = 1/(x_2(t) - x_1(t))$ and $q_n(t) = 1/(x_n(t) - x_{n-1}(t))$, through an example.

3.B.1 Derivation

We give the derivation for $q_n(t)$ only, as $q_1(t)$ is derived in the same way. We follow the idea from Baker et al. [38], relating the cell index i to the density q according to

$$i(x, t) = 1 + \int_0^x q(y, t)\,\mathrm{d}y.$$

Baker et al. [38] use $1 = n - (n-1)$ together with this relationship to write

$$1 = \int_{x_{n-1}(t)}^{x_n(t)} q(y, t)\,\mathrm{d}y,$$

and Baker et al. [38] then use a right endpoint rule to approximate $q_n(t)$. If we instead use a trapezoidal rule, then

$$1 = \int_{x_{n-1}(t)}^{x_n(t)} q(y, t)\,\mathrm{d}y \approx \left(\frac{x_n(t) - x_{n-1}(t)}{2}\right)\left(q_n(t) + q_{n-1}(t)\right). \tag{3.36}$$

We use this expression to solve for $q_n(t)$:

$$q_n(t) = \frac{2}{x_n(t) - x_{n-1}(t)} - q_{n-1}(t) = \frac{2}{x_n(t) - x_{n-1}(t)} - \frac{2}{x_n(t) - x_{n-2}(t)},$$

which is exactly the formula in (3.35). We note that an alternative derivation of this formula is to use linear extrapolation, treating the density $1/(x_n(t) - x_{n-1}(t))$ as if it were placed at the cell midpoint $(x_{n-1}(t) +$

$x_n(t))/2$ rather than $x_n(t)$.

3.B.2 Motivation

Let us now give the motivation for why we need the modifications to the boundary densities in (3.35) compared to those given in Baker et al. [38]. Consider a mechanical relaxation problem, starting with 30 equally spaced nodes in $0 \leq x \leq 5$, taking the parameters $k = 50$, $s = 1/5$, $\eta = 1$ and leaving the right boundary free. Let us compare the discrete densities at $t = 2$ to those from the continuum limit, as well as estimates of the gradient $\partial q/\partial x$ at the right boundary.

Figure 3.B.1 shows our comparisons. Focusing on the densities at the right boundary of Figure 3.B.1(a) gives Figure 3.B.1(b), where we can see a clear difference in the slopes of each curve. The curve obtained using the approach of Baker et al. [38], using $q_n(t) = 1/(x_n(t) - x_{n-1}(t))$, has a different slope from the continuum limit, whereas the red curve, using $q_n(t) = 2/(x_n(t) - x_{n-1}(t)) - 2/(x_n(t) - x_{n-2}(t))$, has a slope that is much closer to the slope of the continuum limit model at this point. These issues become more apparent when we try to estimate $\partial q/\partial x$ at the boundary for each time, as we would have to do in our equation learning procedure. Shown in Figure 3.B.1(c), we see that the estimates of $\partial q/\partial x$ that use $q_n(t) = 1/(x_n(t) - x_{n-1}(t))$ do not resemble what we expect in the continuum limit, namely $\partial q/\partial x = H(q) = 2q^2(1 - qs)$ (using $q = q(x_n, t)$, where $q(x, t)$ is the solution from the continuum limit partial differential equation (PDE)). Our new expression for $q_n(t)$ gives estimates for $\partial q/\partial x$ that are much closer to $H(q)$, with $H(q)$ passing directly through the center of these estimates across the entire time domain. Thus, our revised formulae (3.35) are necessary if we want to obtain accurate estimates for the boundary gradients.

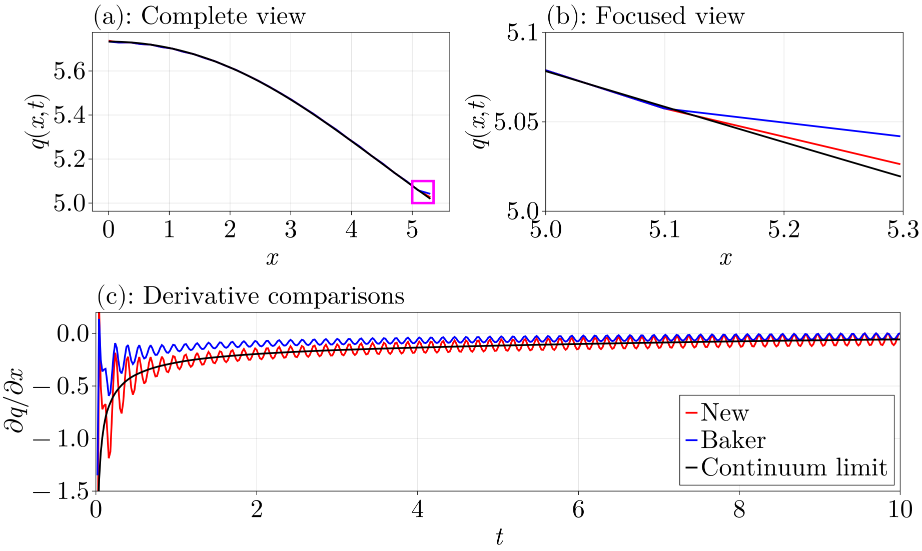

Figure 3.B.1: Comparison of the density definitions from Baker et al. [38] to those in (3.35), using data from a mechanical relaxation problem as an example. (a) Comparing the definitions at $t = 2$ together with densities from the continuum limit PDE. The magenta rectangle shows the region that is zoomed in on in (b). (b) Zooming in on the magenta rectangle from (a) at the right boundary. (c) Comparing estimates of $\partial q/\partial x$ at the right boundary using each definition along with the continuum limit boundary condition $\partial q/\partial x = 2q^2(1 - qs)$.

3.C Numerical methods

In this section we give the details involved in solving the PDEs on the fixed and moving domains numerically using the finite volume method [83]. We have provided Julia packages `FiniteVolumeMethod1D.jl` and `MovingBoundaryProblems1D.jl` to implement these methods for the fixed and moving domains, respectively.

3.C.1 Fixed domain

We start by considering the fixed domain problem. The PDE we consider is

$$\begin{aligned} \frac{\partial q}{\partial t} &= \frac{\partial}{\partial x}\left(D(q)\frac{\partial q}{\partial x}\right) + R(q) \quad 0 < x < L,\, t > 0, \\ \frac{\partial q}{\partial x} &= 0 \qquad\qquad\qquad\qquad\qquad\, x \in \{0, L\},\, t > 0, \end{aligned} \tag{3.37}$$

where L is the length of the domain, $D(q)$ is the nonlinear diffusivity function, and $R(q)$ is the source term. To discretise (3.37), define a grid $0 = x_1 < x_2 < \cdots < x_n = L$ with $x_i = (n-1)\Delta x$ and $\Delta x = L/(n-1)$. This grid enables us to define control volumes $\Omega_i = [w_i, e_i]$ for each i, where

$$\begin{aligned} w_i &= \begin{cases} x_1 & i = 1, \\ \dfrac{1}{2}\left(x_{i-1} + x_i\right) & i = 2, \ldots, n, \end{cases} \\ e_i &= \begin{cases} \dfrac{1}{2}\left(x_i + x_{i+1}\right) & i = 1, \ldots, n-1, \\ x_n & i = n. \end{cases} \end{aligned} \tag{3.38}$$

The volumes of these control volumes are denoted $V_i = e_i - w_i$, $i = 1, \ldots, n$. We then integrate (3.37) over a single Ω_i to give

$$\frac{\mathrm{d}\bar{q}_i}{\mathrm{d}t} = \frac{1}{V_i}\left\{D\left(q(e_i, t)\right)\frac{\partial q(e_i, t)}{\partial x} - D\left(q(w_i, t)\right)\frac{\partial q(w_i, t)}{\partial x}\right\} + \bar{R}_i, \tag{3.39}$$

where

$$\bar{q}_i = \frac{1}{V_i}\int_{w_i}^{e_i} q(x, t)\,\mathrm{d}x \quad \text{and} \quad \bar{R}_i = \frac{1}{V_i}\int_{w_i}^{e_i} R[q(x, t)]\,\mathrm{d}x.$$

To proceed, let $q_i = q(x_i, t)$, $D_i = D(q_i)$, $R_i = R(q_i)$ and define the following approximations:

$$\begin{aligned}
\bar{q}_i &= q_i & i &= 1, \ldots, n, \\
\bar{R}_i &= R_i & i &= 1, \ldots, n, \\
D\left(q(e_i, t)\right) &= \frac{1}{2}\left(D_i + D_{i+1}\right) & i &= 1, \ldots, n-1, \\
D\left(q(w_i, t)\right) &= \frac{1}{2}\left(D_{i-1} + D_i\right) & i &= 2, \ldots, n, \\
\frac{\partial q(e_i, t)}{\partial x} &= \frac{q_{i+1} - q_i}{\Delta x} & i &= 1, \ldots, n-1, \\
\frac{\partial q(w_i, t)}{\partial x} &= \frac{q_i - q_{i-1}}{\Delta x} & i &= 2, \ldots, n.
\end{aligned} \tag{3.40}$$

Using the approximations in (3.40), (3.39) becomes

$$\frac{\mathrm{d}q_i}{\mathrm{d}t} = \frac{1}{V_i}\left[\left(\frac{D_i + D_{i+1}}{2}\right)\left(\frac{q_{i+1} - q_i}{\Delta x}\right) - \left(\frac{D_{i-1} + D_i}{2}\right)\left(\frac{q_i - q_{i-1}}{\Delta x}\right)\right] + R_i, \tag{3.41}$$

for $i = 2, \ldots, n-1$. The boundary conditions are $x = 0$ and $x = L$ are incorporated by simply setting the associated derivative term in (3.39) to zero, giving

$$\frac{\mathrm{d}q_1}{\mathrm{d}t} = \frac{1}{2V_1\Delta x}\left(D_1 + D_2\right)\left(q_2 - q_1\right) + R_1, \tag{3.42}$$

$$\frac{\mathrm{d}q_n}{\mathrm{d}t} = -\frac{1}{2V_n\Delta x}\left(D_{n-1} + D_n\right)\left(q_n - q_{n-1}\right) + R_n. \tag{3.43}$$

The system of ordinary differential equations (ODEs) is thus given by (3.41)–(3.43) and defines the numerical solution to (3.37). In particular, letting $\mathbf{q}^n = (q_1(t_n), \ldots, q_n(t_n))^\mathsf{T}$ for some time t_n, we start with $\mathbf{q}^0 = (q_0(x_1), \ldots, q_0(x_n))^\mathsf{T}$ using the initial condition $q(x, 0) = q_0(x)$, and then integrate forward in time via (3.41)–(3.43). This procedure is implemented in the Julia package `FiniteVolumeMethod1D.jl` which makes use of `DifferentialEquations.jl` to solve the system of ODEs with the `TRBDF2(linsolve = KLUFactorization())` algorithm [85–87].

3.C.2 Moving boundary problem

We now describe how we solve the PDEs for a moving boundary problem. The PDE we consider is

$$\begin{aligned}
\frac{\partial q}{\partial t} &= \frac{\partial}{\partial x}\left(D(q)\frac{\partial q}{\partial x}\right) + R(q) && 0 < x < L(t),\, t > 0,\\
\frac{\partial q}{\partial x} &= 0 && x = 0,\, t > 0,\\
\frac{\partial q}{\partial x} &= H(q) && x = L(t),\, t > 0,\\
q\frac{\mathrm{d}L}{\mathrm{d}t} &= -E(q)\frac{\partial q}{\partial x} && x = L(t),\, t > 0.
\end{aligned} \tag{3.44}$$

We assume that $L(t) > 0$ for $t \geq 0$. The discretisation starts by transforming onto a fixed domain using the Landau transform $\xi = x/L(t)$ [38, 141, 142]. With this change of variable, (3.44) becomes

$$\begin{aligned}
\frac{\partial q}{\partial t} &= \frac{\xi}{L}\frac{\mathrm{d}L}{\mathrm{d}t}\frac{\partial q}{\partial \xi} + \frac{1}{L^2}\frac{\partial}{\partial \xi}\left(D(q)\frac{\partial q}{\partial \xi}\right) + R(q) && 0 < \xi < 1,\, t > 0,\\
\frac{\partial q}{\partial \xi} &= 0 && \xi = 0,\, t > 0,\\
\frac{\partial q}{\partial \xi} &= LH(q) && \xi = 1,\, t > 0,\\
q\frac{\mathrm{d}L}{\mathrm{d}t} &= -\frac{E(q)}{L}\frac{\partial q}{\partial \xi} && \xi = 1,\, t > 0.
\end{aligned} \tag{3.45}$$

To now discretise (3.45), define $\xi_i = (i-1)\Delta\xi$ for $i = 1, \ldots, n$, where $\Delta\xi = 1/(n-1)$, and then let

$$\begin{aligned}
w_i &= \begin{cases} \xi_1 & i = 1,\\ \dfrac{1}{2}\left(\xi_{i-1} + \xi_i\right) & i = 2, \ldots, n, \end{cases}\\
e_i &= \begin{cases} \dfrac{1}{2}\left(\xi_i + \xi_{i+1}\right) & i = 1, \ldots, n-1,\\ \xi_n & i = n. \end{cases}
\end{aligned} \tag{3.46}$$

We then define a control volume to be the interval $\Omega_i = [w_i, e_i]$ with volume $V_i = e_i - w_i$, $i = 1, \ldots, n$. Next, the PDE in (3.45) is integrated over this control volume to give

$$\int_{w_i}^{e_i} \frac{\partial q}{\partial t}\,\mathrm{d}\xi = \frac{1}{L}\frac{\mathrm{d}L}{\mathrm{d}t}\int_{w_i}^{e_i} \xi \frac{\partial q}{\partial \xi}\,\mathrm{d}\xi + \int_{w_i}^{e_i} R(q)\,\mathrm{d}\xi \\ + \frac{1}{L^2}\left[D\left(q(e_i,t)\right)\frac{\partial q(e_i,t)}{\partial \xi} - D\left(q(w_i,t)\right)\frac{\partial q(w_i,t)}{\partial \xi}\right]. \qquad (3.47)$$

Using integration by parts, the first integral on the right-hand side of (3.47) is simply

$$\int_{w_i}^{e_i} \xi \frac{\partial q}{\partial \xi}\,\mathrm{d}\xi = e_i q(e_i,t) - w_i q(w_i,t) - \int_{w_i}^{e_i} q\,\mathrm{d}\xi.$$

Next, define the control volume averages

$$\bar{q}_i = \frac{1}{V_i}\int_{w_i}^{e_i} q\,\mathrm{d}\xi, \qquad \bar{R}_i = \frac{1}{V_i}\int_{w_i}^{e_i} R\,\mathrm{d}\xi,$$

and set $q_i = q(\xi_i,t)$, $D_i = D(q_i)$, and $R_i = R(q_i)$. With this notation, we define the following set of approximations:

$$\begin{aligned}
\bar{q}_i &= q_i & i &= 1,\ldots,n,\\
\bar{R}_i &= R_i & i &= 1,\ldots,n,\\
q(e_i,t) &= \frac{1}{2}\left(q_i + q_{i+1}\right) & i &= 1,\ldots,n-1,\\
q(w_i,t) &= \frac{1}{2}\left(q_{i-1} + q_i\right) & i &= 2,\ldots,n,\\
D\left(q(e_i,t)\right) &= \frac{1}{2}\left(D_i + D_{i+1}\right) & i &= 1,\ldots,n-1,\\
D\left(q(w_i,t)\right) &= \frac{1}{2}\left(D_{i-1} + D_i\right) & i &= 2,\ldots,n,\\
\frac{\partial q(e_i,t)}{\partial \xi} &= \frac{q_{i+1} - q_i}{\Delta\xi} & i &= 1,\ldots,n-1,\\
\frac{\partial q(w_i,t)}{\partial \xi} &= \frac{q_i - q_{i-1}}{\Delta\xi} & i &= 2,\ldots,n.
\end{aligned} \qquad (3.48)$$

Using these approximations, (3.47) becomes

$$\frac{\mathrm{d}q_i}{\mathrm{d}t} = \frac{1}{V_i L}\frac{\mathrm{d}L}{\mathrm{d}t}\left[e_i\left(\frac{q_i + q_{i+1}}{2}\right) - w_i\left(\frac{q_{i-1} + q_i}{2}\right)\right] - \frac{1}{L}\frac{\mathrm{d}L}{\mathrm{d}t}q_i + R_i \\ + \frac{1}{V_i L^2}\left[\left(\frac{D_i + D_{i+1}}{2}\right)\left(\frac{q_{i+1} - q_i}{\Delta\xi}\right) - \left(\frac{D_{i-1} + D_i}{2}\right)\left(\frac{q_i - q_{i-1}}{\Delta\xi}\right)\right]. \qquad (3.49)$$

The last component to handle are the boundary conditions. Since

$\partial q/\partial\xi = 0$ at $\xi = 0$, and since $w_1 = \xi_1 = 0$, our discretisation at $\xi = 0$ becomes

$$\frac{\mathrm{d}q_1}{\mathrm{d}t} = \frac{1}{V_1 L}\frac{\mathrm{d}L}{\mathrm{d}t}e_1\left(\frac{q_1+q_2}{2}\right) - \frac{1}{L}\frac{\mathrm{d}L}{\mathrm{d}t}q_1 + R_1 + \frac{1}{V_1 L^2}\left(\frac{D_1+D_2}{2}\right)\left(\frac{q_2-q_1}{\Delta\xi}\right). \tag{3.50}$$

The boundary condition at $\xi = 1$ is $\partial q/\partial\xi = LH(q)$, thus

$$\begin{aligned}\frac{\mathrm{d}q_n}{\mathrm{d}t} &= \frac{1}{V_n L}\frac{\mathrm{d}L}{\mathrm{d}t}\left[q_n - w_n\left(\frac{q_{n-1}+q_n}{2}\right)\right] - \frac{1}{L}\frac{\mathrm{d}L}{\mathrm{d}t}q_n + R_n \\ &\quad + \frac{1}{V_n L^2}\left[D_n LH(q_n) - \left(\frac{D_{n-1}+D_n}{2}\right)\left(\frac{q_n-q_{n-1}}{\Delta\xi}\right)\right]. \end{aligned} \tag{3.51}$$

The remaining boundary condition is the moving boundary condition, $q\mathrm{d}L/\mathrm{d}t = -[E(q)/L]\partial q/\partial\xi$. Since $\partial q/\partial\xi = LH(q)$, we can write $q_n\mathrm{d}L/\mathrm{d}t = -[E(q_n)/L]LH(q_n) = -E(q_n)H(q_n)$, giving

$$q_n\frac{\mathrm{d}L}{\mathrm{d}t} = -E(q_n)H(q_n). \tag{3.52}$$

The system of ODEs (3.49)–(3.52), together with the initial conditions $q_i(0) = q_0(\xi_i L(0))$ for $i = 1, \ldots, n$ and $L(0) = L_0$, where $q_0(x)$ and L_0 are the initial conditions, define our complete discretisation. Solving these ODEs over time give values for $q(\xi_i, t_j)$, for some t_j, which gets translated back in terms of x via $x_i = \xi_i L(t_j)$. As in the fixed domain case, we solve these ODEs using `DifferentialEquations.jl` together with the `TRBDF2(linsolve = KLUFactorization())` algorithm [85–87]. We provide our implementation of this procedure in a separate Julia package, `MovingBoundaryProblems1D.jl`.

3.D Additional stepwise equation learning details

In this section, we give some extra details for our stepwise equation learning procedure.

3.D.1 Discrete mechanism averaging

We start by discussing how we take multiple stochastic realisations from our discrete cell simulations and average them into a single density function.

The discrete simulations give us n_s identically prepared realisations that can be averaged over to estimate the mean density curve. This average can be estimated using a linear interpolant across each time and for each simulation. In particular, let n_k be the number of knots to use for the interpolant at each time. Then, for a given time t_j, let the knots be given by $\bar{x}_{ij}$ for $i = 1, \ldots, n_k$. These knots are equally spaced with $\bar{x}_{1j} = 0$ and $\bar{x}_{n_k j} = (1/n_s)\sum_{\ell=1}^{n_s} L_j^{(\ell)}$, where $L_j^{(\ell)}$ is the leading edge at the time t_j from the ℓth simulation. Then, letting $q^{(\ell)}(x, t_j)$ denote the linear interpolant of the density data at the time t_j from the ℓth simulation, we define

$$\bar{q}_{ij} = \frac{1}{n_s}\sum_{\ell=1}^{n_s} q^{(\ell)}(\bar{x}_{ij}, t_j), \tag{3.53}$$

for $i = 1, \ldots, n_k$ and $j = 1, \ldots, M$. If $q^{(\ell)}(\bar{x}_{ij}, t_j) < 0$ for a given ℓ, then we set $q^{(\ell)}(\bar{x}_{ij}, t_j) = 0$. This density data is used for computing the system $(\mathbf{A}, \mathbf{b})$ for equation learning when proliferation is involved.

3.D.2 Derivative estimation

The equation learning system $(\mathbf{A}, \mathbf{b})$ requires estimates for the derivatives $\partial q_{ij}/\partial t$, $\partial q_{ij}/\partial x$, $\partial^2 q_{ij}/\partial x^2$, and $\mathrm{d}L_j/\mathrm{d}t$. To give a formula for an estimate of these derivatives, suppose we have three function values $\{f_1, f_2, f_3\}$ for some function $f(x)$ at the points $\{x_1, x_2, x_3\}$, where $f_i = f(x_i)$ for $i = 1, 2, 3$. These points do not need to be equally spaced. The Lagrange interpolating polynomial through this data is given by

$$g(x) = \frac{(x-x_2)(x-x_3)}{(x_1-x_2)(x_1-x_3)}f_1 + \frac{(x-x_1)(x-x_3)}{(x_2-x_1)(x_2-x_3)}f_2 + \frac{(x-x_1)(x-x_2)}{(x_3-x_1)(x_3-x_2)}f_3,$$

which can be used to estimate the derivatives via $f'(x_i) \approx g'(x_i)$, $i = 1, 2, 3$, and similarly for $f''(x)$. Using this approximation, we write

$$\begin{aligned} f'(x_1) \approx{}& \left(\frac{1}{x_1 - x_2} + \frac{1}{x_1 - x_3}\right) f_1 - \frac{x_1 - x_3}{(x_1 - x_2)(x_2 - x_3)} f_2 \\ &+ \frac{x_1 - x_2}{(x_1 - x_3)(x_2 - x_3)} f_3, \end{aligned} \tag{3.54}$$

$$\begin{aligned} f'(x_2) \approx{}& \frac{x_2 - x_3}{(x_1 - x_2)(x_1 - x_3)} f_1 + \left(\frac{1}{x_2 - x_3} - \frac{1}{x_1 - x_2}\right) f_2 \\ &+ \frac{x_2 - x_1}{(x_1 - x_3)(x_2 - x_3)} f_3, \end{aligned} \tag{3.55}$$

$$\begin{aligned} f'(x_3) \approx{}& \frac{x_3 - x_2}{(x_1 - x_2)(x_1 - x_3)} f_1 + \frac{x_1 - x_3}{(x_1 - x_2)(x_2 - x_3)} f_2 \\ &- \left(\frac{1}{x_1 - x_3} + \frac{1}{x_2 - x_3}\right) f_3, \end{aligned} \tag{3.56}$$

$$\begin{aligned} f''(x_i) \approx{}& \frac{2}{(x_1 - x_2)(x_1 - x_3)} f_1 - \frac{2}{(x_1 - x_2)(x_2 - x_3)} f_2 \\ &+ \frac{2}{(x_1 - x_3)(x_2 - x_3)} f_3, \end{aligned} \tag{3.57}$$

where (3.57) is valid for $i = 1, 2, 3$.

We can use the formulae (3.54)–(3.57) to approximate our required derivatives. For example, taking $\{x_1, x_2, x_3\} = \{t_{j-1}, t_j, t_{j+1}\}$ and $\{f_1, f_2, f_3\} = \{L_{j-1}, L_j, L_{j+1}\}$ gives

$$\frac{\mathrm{d}L_j}{\mathrm{d}t} \approx \frac{L_{j+1} - L_{j-1}}{h}, \quad j = 2, \dots, M-1, \tag{3.58}$$

assuming the times are equally spaced with spacing h. The estimate for $\mathrm{d}L_M/\mathrm{d}t$ is obtained by taking $\{x_1, x_2, x_3\} = \{t_{M-2}, t_{M-1}, t_M\}$ and $\{f_1, f_2, f_3\} = \{L_{M-2}, L_{M-1}, L_M\}$, giving

$$\frac{\mathrm{d}L_M}{\mathrm{d}t} \approx \frac{3L_M - 4L_{M-1} + L_{M-2}}{2h}. \tag{3.59}$$

Similarly, taking the points $\{x_1, x_2, x_3\} = \{x_{i-1,j}, x_{ij}, x_{i+1,j}\}$ and function values $\{f_1, f_2, f_3\} = \{q_{i-1,j}, q_{ij}, q_{i+1,j}\}$ gives

$$\begin{aligned}\frac{\partial^2 q_{ij}}{\partial x^2} \approx & \frac{2}{(x_{i-1,j}-x_{i,j})(x_{i-1,j}-x_{i+1,j})}q_{i-1,j} \\ & - \frac{2}{(x_{i-1,j}-x_{ij})(x_{ij}-x_{i+1,j})}q_{ij} \\ & + \frac{2}{(x_{i-1,j}-x_{i+1,j})(x_{ij}-x_{i+1,j})},\end{aligned} \tag{3.60}$$

for $i = 2, \ldots, n_j - 1$ and $j = 1, \ldots, M$, where n_j is the number of nodes at $t = t_j$. The remaining derivatives can be obtained similarly, ensuring that the appropriate finite difference (backward, central, or forward) is taken for the given point.

The only exception to these rules are for $\partial q/\partial x$ at the boundaries. We find that using simple forward and backward differences there gives better results than with (3.54) and (3.55), so we use

$$\frac{\partial q_{1j}}{\partial x} \approx \frac{q_{2j}-q_{1j}}{x_{2j}-x_{1j}}, \quad \frac{\partial q_{n_j j}}{\partial x} \approx \frac{q_{n_j j}-q_{n_j-1,j}}{x_{n_j j}-x_{n_j-1,j}}. \tag{3.61}$$

3.D.3 Matrix pruning

We now discuss our approach to *matrix pruning*, wherein we discard points from our equation learning matrix $\mathbf{A}$ that do not help to improve our estimates for $\boldsymbol{\theta}$. The approach we take is inspired from the data thresholding idea from VandenHeuvel et al. [50].

To start with our approach, let $\mathbf{q} = (q_{12}, \ldots, q_{n_M M})^{\mathsf{T}}$ be the vector of all discrete densities, letting n_j be the number of nodes at the time $t = t_j$, excluding the densities from the initial condition. Then, take the *threshold tolerance* $0 \leq \tau_q < 1/2$ and compute the interval $(\mathcal{Q}^{\mathbf{q}}_{\tau_q}, \mathcal{Q}^{\mathbf{q}}_{1-\tau_q})$, where $\mathcal{Q}^{\mathbf{y}}_{\tau}$ denotes the $100\tau\%$ quantile of the vector $\mathbf{y}$. With these intervals, we only include a row in the matrix $\mathbf{A}$ from a given point (x_{ij}, t_j) if $\mathcal{Q}^{\mathbf{q}}_{\tau_q} \leq q_{ij} \leq \mathcal{Q}^{\mathbf{q}}_{1-\tau_q}$.

By choosing the threshold τ_q appropriately, we can significantly improve the estimates for $\boldsymbol{\theta}$ as we only include the most relevant data for estimation, excluding all points with relatively low or high density. Similar thresholds can be defined for the other quantities $|\partial\mathbf{q}/\partial x|, |\partial^2\mathbf{q}/\partial x^2|, |\partial\mathbf{q}/\partial t|$, and $|\mathrm{d}\mathbf{L}/\mathrm{d}t|$, defining these vectors similarly to $\mathbf{q}$, for example $|\partial_t\mathbf{q}| = (|\partial_t q_{12}|, \ldots, |\partial_t q_{n_M M}|)^{\mathsf{T}}$, with respective threshold tolerances satisfying $0 \leq \tau_{\partial q/\partial x}, \tau_{\partial^2 q/\partial x^2}, \tau_{\partial q/\partial t}, \tau_{\mathrm{d}L/\mathrm{d}t} < 1/2$.

3.E Additional examples

In this section, we give some additional case studies to further demonstrate our method, exploring different force law and proliferation laws, and enforcing conservation of mass together with a discussion about enforcing equality constraints in general.

3.E.1 Enforcing conservation of mass

In the main chapter, we discussed at the end of Case Study 2 that it could be possible to enforce mass conservation to fix the issue with $D(q) \neq E(q)$, noting that mass conservation requires $D(q(L(t), t)) = E(q(L(t), t))$. In this section, we consider the results when we fix $D(q) = E(q)$ so that mass is conserved from the outset.

This change $D(q) = E(q)$ is reasonably straightforward to implement in the algorithm, simply replacing the boundary condition (3.17) so that

$$q(L(t), t) \frac{\mathrm{d}L(t)}{\mathrm{d}t} = -D\left(q(L(t), t)\right) \frac{\partial q(L(t), t)}{\partial x}. \tag{3.62}$$

This constraint $D(q) = E(q)$ also needs to be reflected in the matrix $\mathbf{A}$. This is simple to do in this case. Previously, our matrix system took the block diagonal form

$$\begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_3 \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}^d \\ \boldsymbol{\theta}^h \\ \boldsymbol{\theta}^e \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \mathbf{b}_3 \end{bmatrix}. \tag{3.63}$$

With the constraint $D(q) = E(q)$, (3.63) becomes

$$\begin{bmatrix} \mathbf{A}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 \\ \mathbf{A}_3 & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}^d \\ \boldsymbol{\theta}^h \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \mathbf{b}_3 \end{bmatrix}. \tag{3.64}$$

We note that, if we wanted to enforce this constraint in Case Study 4, where $\mathbf{A}_1 = [\mathbf{A}^d \ \mathbf{A}^r]$, with $\mathbf{A}^d$ and $\mathbf{A}^r$ defined from (3.28), then we instead have

$$\begin{bmatrix} \mathbf{A}^d & \mathbf{A}^r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_2 \\ \mathbf{A}_3 & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}^d \\ \boldsymbol{\theta}^r \\ \boldsymbol{\theta}^h \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \mathbf{b}_3 \end{bmatrix}. \tag{3.65}$$

Table 3.E.1: Stepwise equation learning results for Case Study 2, using the basis expansions (3.25), saving the results at $M = 200$ equally spaced times between $t_1 = 0$ and $t_M = 15$, pruning with $\tau_q = 0.35$ and $\tau_{\mathrm{d}L/\mathrm{d}t} = 0.1$, starting with all terms inactive, and enforcing conservation of mass with $D(q) = E(q)$. Coefficients highlighted in blue show the coefficient chosen to be removed or added at the corresponding step.

Step	θ_1^d	θ_2^d	θ_3^d	θ_1^h	θ_2^h	θ_3^h	θ_4^h	θ_5^h	**Loss**
1	0.000	0.000	0.000	0.000	**0.000**	0.000	0.000	0.000	-3.371
2	0.000	**0.000**	0.000	0.000	-0.025	0.000	0.000	0.000	-2.371
3	0.000	47.413	0.000	0.000	-0.025	0.000	0.000	**0.000**	-1.706
4	0.000	47.413	0.000	0.000	0.443	0.000	0.000	-0.004	-0.688

Let us now consider the results with mass conservation. We use the same parameters that were used to produce the results in Figure 3.7. In particular, we save the solution at $M = 200$ equally spaced times between $t_1 = 0$ and $t_M = 15$, $\tau_q = 0.35$, $\tau_{\mathrm{d}L/\mathrm{d}t} = 0.1$, and we start with all coefficients initially inactive. The results we obtain are shown in Table 3.E.1 and Figure 3.E.1. We see that the form we learn for $D(q)$, and hence for $E(q)$ also, is close to the continuum limit $50/q^2$, and similarly $H(q)$ is a good match; note that $H(q)$ is only evaluated at the boundary densities, which is approximately 5 for $t > 0$, so indeed $H(q)$ matches the continuum limit. Looking to Figure 3.E.1(a)–(b), the results are indistinguishable from the continuum limit, which is also what we found in Figure 3.7 before we enforced conservation of mass.

Imposing linear equality constraints generically

We note that this approach to implementing the constraint $D(q) = E(q)$ requiring such a significant change to the matrix system, giving (3.64), and to the boundary condition (3.62), might suggest that the modularity of our approach weakens here. This does not need to be the case, and so let us briefly remark about how constraints such as $D(q) = E(q)$, or any other linear constraints, could be alternatively implemented in our approach seamlessly, further demonstrating the modularity.

Suppose we take our system $\mathbf{A}\boldsymbol{\theta} = \mathbf{b}$, with $\mathbf{A} \in \mathbb{R}^{m \times p}$, $\boldsymbol{\theta} \in \mathbb{R}^p$, and $\mathbf{b} \in \mathbb{R}^m$, and suppose we have constraints of the form $\mathbf{Q}^{\mathsf{T}}\boldsymbol{\theta} = \mathbf{c}$ where $\mathbf{Q} \in \mathbb{R}^{p \times c}$ and $\mathbf{c} \in \mathbb{R}^c$, where $c < p$ and $\mathbf{Q}$ has full rank. The constrained least squares estimator for $\boldsymbol{\theta}$ subject to these constraints, denoted $\hat{\boldsymbol{\theta}}^c$, is

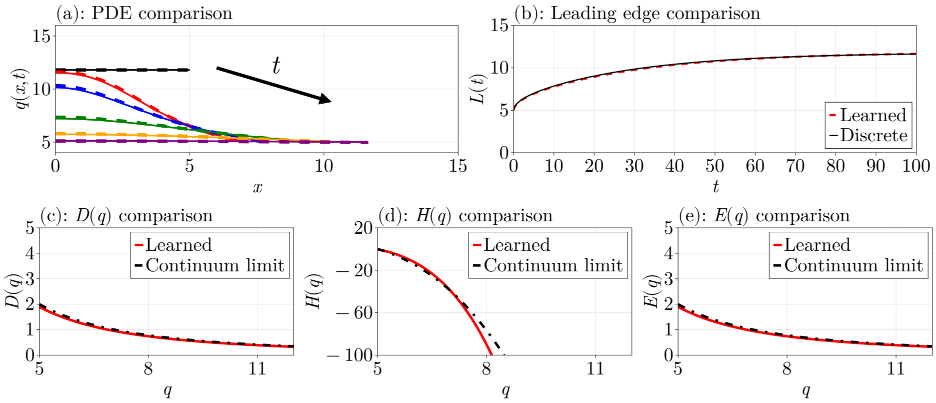

Figure 3.E.1: Stepwise equation learning results from Table 3.E.1. (a) Comparisons of the discrete density profiles (solid curves) with those from the learned PDE (dashed curves), plotted at the times $t = 0, 5, 10, 25, 50, 100$ in black, red, blue, green, orange, and purple, respectively. (b) As in (a), except comparing the leading edges. (c)–(e) are comparisons of the learned forms of $D(q)$, $H(q)$, and $E(q)$ compared to the forms from the continuum limit.

then given by

$$\hat{\boldsymbol{\theta}}^c = \hat{\boldsymbol{\theta}} - \left(\mathbf{A}^\mathsf{T}\mathbf{A}\right)^{-1}\mathbf{Q}\left[\mathbf{Q}^\mathsf{T}\left(\mathbf{A}^\mathsf{T}\mathbf{A}\right)^{-1}\mathbf{Q}\right]^{-1}\left(\mathbf{Q}^\mathsf{T}\hat{\boldsymbol{\theta}} - \mathbf{c}\right), \tag{3.66}$$

where $\hat{\boldsymbol{\theta}} = (\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}\mathbf{b}$ is the unconstrained least squares estimator for $\mathbf{A}\boldsymbol{\theta} = \mathbf{b}$ [143]. Using this formulation, imposing $D(q) = E(q)$ is simple to enforce without changing the boundary condition or the matrix $\mathbf{A}$, simply using $\mathbf{c} = \mathbf{0}_{3\times 1}$ and

$$\mathbf{Q} = \begin{array}{c} \boldsymbol{\theta}^d \\ \boldsymbol{\theta}^h \\ \boldsymbol{\theta}^e \end{array} \begin{bmatrix} \mathbf{I}_3 \\ \mathbf{0}_{5\times 3} \\ -\mathbf{I}_3 \end{bmatrix} = \begin{array}{c} \theta_1^d \\ \theta_2^d \\ \theta_3^d \\ \theta_1^h \\ \theta_2^h \\ \theta_3^h \\ \theta_4^h \\ \theta_5^h \\ \theta_1^e \\ \theta_2^e \\ \theta_3^e \end{array} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix},$$

where $\mathbf{I}_n$ and $\mathbf{0}_{m\times n}$ denote the n-square identity matrix and $m \times n$ zero matrix, respectively. This does not solve the problem entirely, though, since we also have coefficients that we force to zero throughout the stepwise procedure. These zeros constraints can also be imposed by including additional columns of $\mathbf{Q}$. For example, if θ_1^h and θ_2^d are inactive, then $\mathbf{Q}$ becomes

$$\mathbf{Q} = \begin{array}{c} \boldsymbol{\theta}^d \\ \boldsymbol{\theta}^h \\ \boldsymbol{\theta}^e \end{array} \begin{bmatrix} \mathbf{I}_3 & \mathbf{e}_2^d & \mathbf{0}_{3\times 1} \\ \mathbf{0}_{5\times 3} & \mathbf{0}_{5\times 1} & \mathbf{e}_1^h \\ -\mathbf{I}_3 & \mathbf{0}_{3\times 1} & \mathbf{0}_{3\times 1} \end{bmatrix}, \tag{3.67}$$

where $\mathbf{e}_2^d = (0,1,0)^\mathsf{T}$ and $\mathbf{e}_1^h = (1,0,0,0,0)^\mathsf{T}$. In particular, each inactive coefficient θ_i corresponds to a new column with a one in the row corresponding to that coefficient. Note that $\mathbf{Q}$ in (3.67) can be further written as $\mathbf{Q} = [\mathbf{Q}_1\ \mathbf{Q}_2]$, where $\mathbf{Q}_1$ are the user-provided constraints $D(q) = E(q)$ and $\mathbf{Q}_2$ are the constraints imposed by the inactive coefficients, making it easy to incorporate constraints in this manner. Additional care is required to ensure that there are no redundant constraints represented by $\mathbf{Q}_1$ and $\mathbf{Q}_2$ as $\mathbf{Q}$ must be full rank. For example, imposing $\theta_1^d = 0$ and $\theta_1^e = 0$ together with the constraint $\theta_1^d = \theta_1^e$ from $D(q) = E(q)$ can be represented using only two constraints rather than three, and the associated matrix

$$\mathbf{Q} = \begin{array}{c@{}c} \begin{array}{c} \boldsymbol{\theta}^d \\ \boldsymbol{\theta}^h \\ \boldsymbol{\theta}^e \end{array} & \begin{bmatrix} \mathbf{I}_3 & \mathbf{e}_1^d & \mathbf{0}_{3\times 1} \\ \mathbf{0}_{5\times 3} & \mathbf{0}_{5\times 1} & \mathbf{0}_{5\times 1} \\ -\mathbf{I}_3 & \mathbf{0}_{3\times 1} & \mathbf{e}_1^h \end{bmatrix} \end{array}, \tag{3.68}$$

where $\mathbf{e}_1^h = (1,0,0)^{\mathsf{T}}$, only has rank 4 rather than the full rank 5. This could be dealt with by finding a basis for the column space of $\mathbf{Q}$, replacing $\mathbf{Q}$ with the corresponding matrix of basis vectors.

To summarise this discussion, it is straightforward to implement our procedure with the ability to enforce linear equality constraints, allowing for additional constraints, such as conservation of mass, to be enforced. This is easy to code without breaking the modularity of the approach and requiring a significant change to the procedure that would be cumbersome to implement by increasing the complexity of the corresponding code.

3.E.2 A piecewise proliferation law

In this section, we consider the problem described in Section 3.3 of Murphy et al. [40]. This problem given by Murphy et al. [40] is used to demonstrate a case where the solution of the continuum limit no longer gives a good match with averaged data from the discrete model, as the value of k used is too low relative to the proliferation rate. Here, we show how our method can learn an accurate continuum model in this case.

The example is as follows. We consider $F(\ell_i) = k(s - \ell_i)$ as usual, taking $k = 10^{-4}$ and $s = 0$, but our proliferation law is now given by

$$G(\ell_i) = \begin{cases} 0 & 0 \leq \ell_i < \ell_p, \\ \beta & \ell_i \geq \ell_p, \end{cases} \tag{3.69}$$

where $\ell_p = 0.2$ is the proliferation threshold and $\beta = 10^{-2}$. We use $\Delta t = 10^{-2}$ for the proliferation events. The initial condition places $n = 41$ equally spaced nodes in $[0, 10]$ so that $\ell_i = 0.25$ at $t = 0$ for each of the 40 cells. In Figure 3.E.2, we show a comparison of the discrete data from this problem with the solution of the continuum limit. We also compare the cell numbers $N(t)$, where the cell numbers from the PDE $q(x, t)$ are obtained via $N(t) = \int_0^{10} q(x,t)\,\mathrm{d}x$. We see that the densities from the solution of the continuum limit reach a capacity at 50 cells, while the discrete model instead reaches 80 cells. Note that the densities appear jagged in Figure 3.B.1 due to the combination of the averaging procedure from Section 3.D.1 with the variance of the densities for moderate t; a better averaging method could be to build the knots at each time t based on the node positions themselves, but we do not consider that here as it does not impact the results.

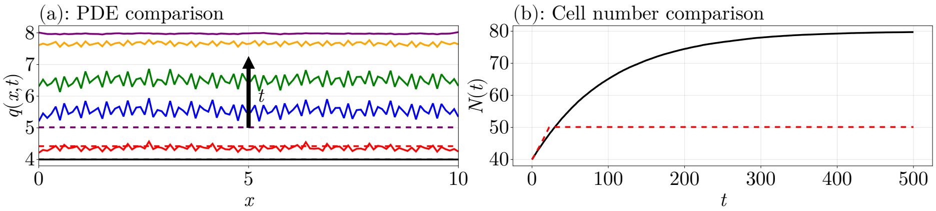

Figure 3.E.2: Comparison of the solution of the piecewise proliferation law problem with the solution of continuum limit, where $F(\ell_i) = k(s-\ell_i)$ and $G(\ell_i) = \beta$ for $\ell_i \geq \ell_p$ and $G(\ell_i) = 0$ otherwise, using $k = 10^{-4}$, $s = 0$, $\ell_p = 0.2$, $\eta = 1$, $\beta = 10^{-2}$, and $\Delta t = 10^{-2}$. (a) The solid curves are the discrete densities, and the dashed curves are the densities from the solution of the continuum limit. The arrow shows the direction of increasing time. The density profiles are shown at the times $t = 0, 10, 50, 100, 250, 500$ in black, red, blue, green, orange, and purple, respectively. (b) Comparison of the number of cells from the discrete model with that computed from the solution of the continuum limit, using $N(t) = \int_0^{10} q(x,t)\,\mathrm{d}x$ for the continuum limit case. In (a)–(b), the discrete results are averaged over 1000 identically prepared realisations, using $n_k = 100$ knots for the averaging procedure described in Section 3.D.1.

The continuum limit for this problem is

$$D(q) = \frac{10^{-4}}{q^2} \quad \text{and} \quad R(q) = \begin{cases} 0 & q > 1/\ell_p, \\ 10^{-2}q & q \leq 1/\ell_p. \end{cases}$$

This suggests one possible basis expansion to use for $R(q)$ in our equation learning procedure, with the aim to learn an appropriate continuum approximation to the results in Figure 3.E.2, could be

$$R(q) = \left[\theta_0^r + \theta_1^r q + \theta_2^r q^2 + \theta_3^r q^3\right] \mathbb{I}\left(q \leq \frac{1}{\ell_p}\right),$$

where $\mathbb{I}(A)$ is the indicator function for the set A. We find that this does not lead to any improved model for this problem, and so we instead consider a polynomial model:

$$R(q) = \theta_0^r + \theta_1^r q + \theta_2^r q^2 + \theta_3^r q^3 + \theta_4^r q^4 + \theta_5^r q^5. \tag{3.70}$$

For $D(q)$, this mechanism does not appear to be relevant in this example, with the results that follow all giving visually indistinguishable regardless of whether $D(q) = 0$ or $D(q) = 10^{-4}/q^2$. Thus, we do not bother learning it in this case, simply fixing $D(q) = 10^{-4}/q^2$; if we do not fix $D(q)$, we just end up learning $D(q) = 0$ in the results that follow. With (3.70) and $D(q) = 10^{-4}/q^2$, the results we obtain are shown in Table 3.E.2 and Figure 3.E.3.

Step	θ_1^r	θ_2^r	θ_3^r	θ_4^r	θ_5^r	θ_6^r	**Loss**
1	**0.00**	0.00	0.00	0.00	0.00	0.00	-1.63
2	0.00	**0.00**	0.00	0.00	0.00	0.00	-1.53
3	0.077	-0.0096	0.00	0.00	0.00	0.00	-6.22

Table 3.E.2: Equation learning results for the piecewise proliferation law problem in Figure 3.E.2, fixing $D(q) = 10^{-4}/q^2$ and using the expansion of $R(q)$ in (3.70). The discrete data is averaged over 1000 identically prepared realisations with $n_k = 100$ knots for interpolating, and the solution is saved every 0.1 units of time between $t = 0$ and $t = 500$.

The results in Table 3.E.2 and Figure 3.E.3 show that we have learned

$$R(q) = 0.077 - 0.0096q. \tag{3.71}$$

The results in Figure 3.E.3(a)–(b) show a good match between the discrete data and the learned PDE solution. Most interestingly, 3.E.3(c), we see that this learned $R(q)$ connects the endpoints of the continuum limit form continuously. In particular, $R(q) \approx \beta(K - q) = \beta K(1 - q/K)$, where $K = 8$ is the maximum density from the averaged discrete data. We have thus learned an accurate continuum model to describe this problem, originally from Murphy et al. [40], showing that the piecewise continuum limit form of $R(q)$ is more appropriately described by a simple linear model that connects the jumps in $R(q)$.

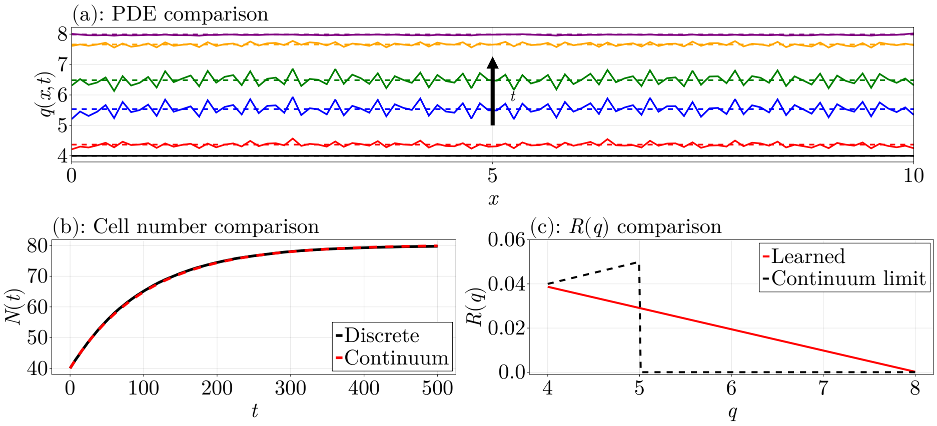

Figure 3.E.3: Equation learning results for the piecewise proliferation law problem in Figure 3.E.2, using the results from Table 3.E.2. (a) Comparison of the averaged discrete densities (solid curves) with the solution of the learned PDE (dashed). The arrow shows the direction of increasing time. The arrow shows the direction of increasing time. The density profiles are shown at the times $t = 0, 10, 50, 100, 250, 500$ in black, red, blue, green, orange, and purple, respectively.(b) Comparison of the cell numbers. (c) Comparison of the learned form of $R(q)$ with the continuum limit form of $R(q)$.

3.E.3 Linear diffusion

In this section, we consider an example where we consider a force law that leads to linear diffusion, namely

$$F(\ell_i) = k\left(\frac{1}{\ell_i} - s\right), \tag{3.72}$$

We use $k = 20$ and $s = 1$. For the initial condition, we consider a Gaussian initial density $q_0(x)$ with variance three centered at $x = L_0/2$ over $0 \leq x \leq L_0$ with $L_0 = 10$, and scaled so that the initial number of cells is 40, meaning $40 = \int_0^{10} q_0(x)\,\mathrm{d}x$. This leads to

$$q_0(x) = \frac{A}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2}\left(\frac{x - L_0/2}{\sigma}\right)^2\right\}, \quad A = \left[\operatorname{erf}\left(\frac{L_0\sqrt{2}}{4\sigma}\right)\right]^{-1} N(0), \tag{3.73}$$

where $N(0) = 40$, $\sigma^2 = 3$, and erf is the error function. To convert this density into a set of initial cell positions, we consider a set of nodes $x_1, \ldots, x_{41}$ with $x_1 = 0$ and $x_{41} = L_0$. The interior nodes $\tilde{\mathbf{x}}(0) = (x_2(0), \ldots, x_{40}(0))^{\mathsf{T}}$

are obtained by solving the optimisation problem

$$\tilde{\mathbf{x}}(0) = \underset{\tilde{\mathbf{x}}\in\mathbb{R}^{39}}{\operatorname{argmin}} \sum_{i=1}^{41} \left(q_0\left(x_i(0)\right) - q_i\right)^2$$

subject to the constraint $0 < x_2(0) < \cdots < x_{40}(0) < L_0$, where q_i is the density at x_i using our piecewise formulae. This problem is solved using `NLopt.jl` [91, 92]. The discrete densities we obtain over $0 \leq t \leq 100$ are shown in Figure 3.E.4, where we also compare the data to the solution of the continuum limit.

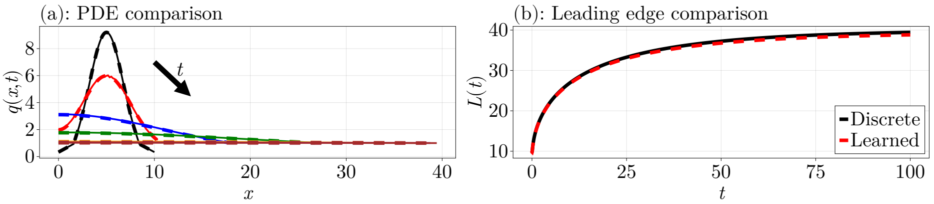

Figure 3.E.4: Comparison of the linear diffusion problem with its continuum limit, where $F(\ell_i) = k(a/\ell_i - s)$ with $k = 20$, $s = 1$, $\eta = 1$, and a Gaussian initial density. (a) The solid curves are the discrete densities, and the dashed curves are the densities from the solution of the continuum limit. The arrow shows the direction of increasing time. The density profiles are shown at the times $t = 0, 0.1, 2, 10, 50, 75, 100$ in black, red, blue, green, orange, purple, and brown, respectively. (b) Like in (a), except comparing the leading edges.

To apply the equation learning procedure to this problem, we note that we expect $D(q) = E(q) = 20$, and $H(q) = 2q - 2q^2$. We thus consider

$$\begin{aligned}
D(q) &= \frac{\theta^d_{-2}}{q^2} + \frac{\theta^d_{-1}}{q} + \theta^d_0 + \theta^d_1 q + \theta^d_2 q^2, \\
H(q) &= \theta^h_1 q + \theta^h_2 q^2 + \theta^h_3 q^3 + \theta^h_4 q^4 + \theta^h_5 q^5, \\
E(q) &= \frac{\theta^e_{-2}}{q^2} + \frac{\theta^e_{-1}}{q} + \theta^e_0 + \theta^e_1 q + \theta^e_2 q^2.
\end{aligned}$$

Saving the solution between $t = 0$ and $t = 100$ every 0.01 units of time and pruning with $\tau_q = 0.3$ and $\tau_{\mathrm{d}L/\mathrm{d}t} = 0.2$, we obtain the results in Table 3.E.3 and Figure 3.E.5, showing a good match between the solution of the learned model and the discrete data.

Step	θ^d_{-2}	θ^d_{-1}	θ^d_0	θ^d_1	θ^d_2	θ^h_1	θ^h_2	θ^h_3	θ^h_4	θ^h_5	θ^e_{-2}	θ^e_{-1}	θ^e_0	θ^e_1	θ^e_2	**Loss**
1	0.00	0.00	**0.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.76
2	0.00	0.00	19.18	0.00	0.00	0.00	0.00	0.00	0.00	**0.00**	0.00	0.00	0.00	0.00	0.00	1.14
3	0.00	0.00	19.18	0.00	0.00	0.00	0.00	0.00	0.00	-0.02	0.00	0.00	**0.00**	0.00	0.00	0.98
4	0.00	0.00	19.18	0.00	0.00	0.00	**0.00**	0.00	0.00	-0.02	0.00	0.00	20.05	0.00	0.00	-5.05
5	0.00	0.00	19.18	0.00	0.00	0.00	0.42	0.00	0.00	-0.42	0.00	0.00	20.05	0.00	0.00	-10.73

Table 3.E.3: Equation learning results for the linear diffusion problem in Figure 3.E.4. The solution is saved every 10^{-2} units of time between $t = 0$ and $t = 100$, and matrix pruning is used with $\tau_q = 0.3$ and $\tau_{\mathrm{d}L/\mathrm{d}t} = 0.2$.

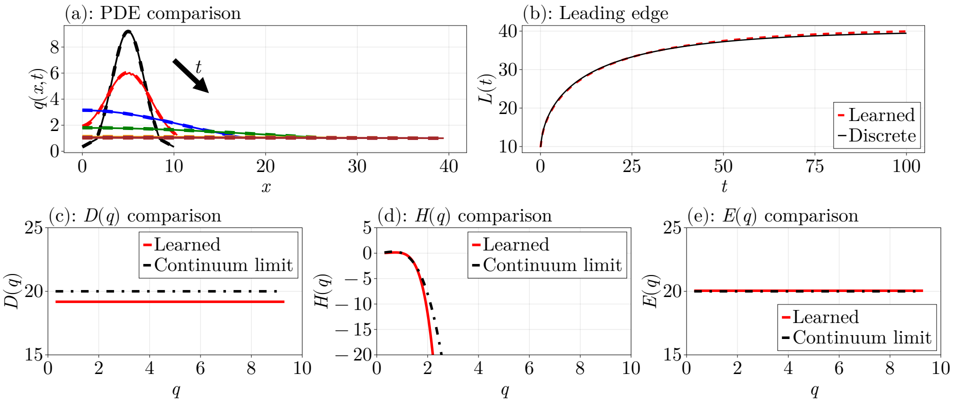

Figure 3.E.5: Equation learning results for the linear diffusion problem in Figure 3.E.2, using the results from Table 3.E.2. (a) Comparison of the discrete densities (solid curves) with the solution of the learned PDE (dashed). The arrow shows the direction of increasing time. The density profiles are shown at the times $t = 0, 0.1, 2, 10, 50, 75, 100$ in black, red, blue, green, orange, purple, and brown, respectively. (b) Line in (a), except comparing the leading edges. (c)–(e) shows comparisons of the learned mechanisms with the forms from the continuum limit.

3.F Parameter sensitivity study

In this appendix, we provide a brief parameter sensitivity study, exploring the impact of parameters such as the pruning parameters and the number of time points on the results of our stepwise learning framework. We use Case Study 3 for this purpose, taking the case $k = 1/5$ so that the continuum limit is inaccurate. The parameters we consider are h, the duration between time points; n_s, the number of identically-prepared realisations; t_M, the final time, noting that $t_1 = 0$; n_k, the number of knots used for averaging; and τ_q, the pruning parameter for the density quantiles. We only vary each parameter one at a time, so that the default values for each parameter are $h = 0.1$, $n_s = 1000$, $n_k = 200$, $t_M = 75$, and $\tau_q = 0.25$ while a given parameter is being varied.

To assess the results for each set of parameters we use the loss of the learned model, $\mathcal{L}(\hat{\boldsymbol{\theta}})$. To further examine the results, we divide the results into two categories: those that learn $D(q) = 0$, and those that learn $D(q) \neq 0$. The results of the study are shown in Figure 3.F.1. We see that there is little dependence of the results on h, or equivalently on the number of time points. Figure 3.F.1(b) shows that n_s needs to be sufficiently large, around $n_s > 500$, in order for any diffusion terms to be selected, although the loss does not change significantly once $D(q)$ terms are identified. The final time is important, where only final times in $50 \leq t_M \leq 75$ give reasonable results. The number of knots is not too important according to Figure 3.F.1(d), so long as there are not too many or too few. The most impactful parameter is τ_q, where we need $\tau_q \approx 0.2$ to obtain an adequate learned model; for other case studies which involve other pruning parameters, such as on the derivatives or on the leading edge, we also find that these parameters are the most influential.

Overall, Figure 3.F.1 shows that τ_q and t_M are the most important parameters for this problem. This is consistent with what we have found for the other case studies, where the choice of pruning parameters is crucial and the time horizon needs to be carefully chosen so that $D(q)$ can be identified. Choosing these parameters can be quite difficult, and trial and error is needed to identify appropriate terms, as well as understanding why a certain model is failing to give good results. For example, in Case Study 2 we determined that we had to shrink the time interval used for learning the results, and that we needed to use velocity quantiles, by determining

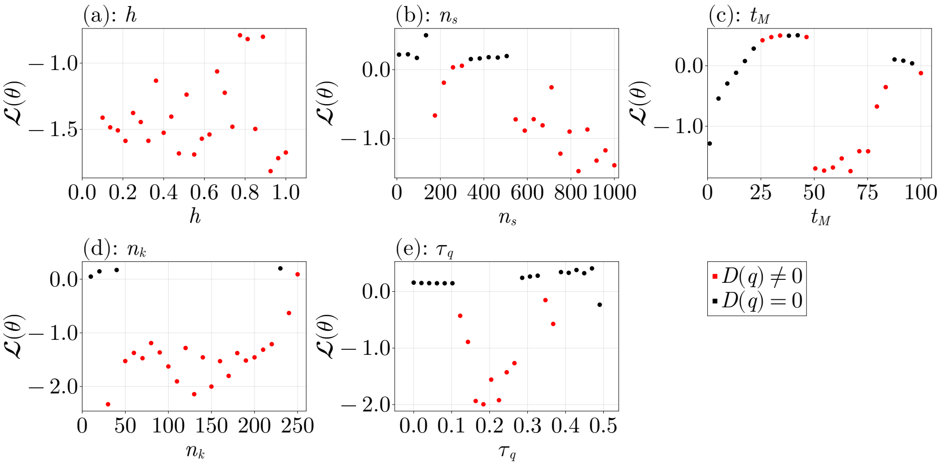

Figure 3.F.1: Dependence of $\mathcal{L}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the vector combined the learned $\boldsymbol{\theta}^d$ and $\boldsymbol{\theta}^r$, on the parameters h, n_s, t_M, n_K, and τ_q. For each parameter, as it is varied the other parameters are held at their default values $h = 0.1$, $n_s = 1000$, $t_M = 75$, $n_k = 200$, and $\tau_q = 0.25$.

what mechanisms are failing to be learned and seeing where the model fails to extrapolate. The values that we used for these parameters, though, had to be chosen with trial and error. Our procedure is efficient enough for this trial and error procedure to be performed quickly, but future work could examine these issues in more detail to simplify the selection of these parameters.

Chapter 4

Conclusion and Future Work

In this thesis, we apply mathematical and statistical methods to problems in tissue engineering. There are two primary outcomes of this research. Firstly, we combine mathematical modelling with a likelihood-based uncertainty framework to demonstrate that the cellular mechanisms driving tissue growth are independent of pore geometry. While previous research has explored tissue growth in various pore geometries [13, 21–26, 55, 56, 58–62], this is the first work that has directly examined the cellular mechanisms between geometries and provided evidence of this independence. Lastly, we develop an equation learning framework that can learn models describing tissue growth experiments, like those examined in the first half of this thesis, enabling partial differential equation models to be derived for experiments more complicated than those considered in this thesis.

Our findings in Chapter 2 have several important implications. Firstly, it is possible to make predictions of tissue growth experiments, and thus of quantities such as the bridging time, on pore geometries using models calibrated from other pore geometries. This can be useful as a screening tool, where experiments on new geometries can be simulated before even fabricating the scaffolds; these simulations do not replace experimental verification, but they will significantly assist with planning and interpreting experiments effectively. Moreover, the results imply that observed curvature effects, such as those discussed by Callens et al. [58] who find that cells generally prefer concave regions, are due to space availability rather than cellular mechanisms. This insight is useful as it may assist in how future experiments are designed to further study this phenomena, and to focus research into mechanisms other than the cellular mechanisms for understanding curvature effects further. These findings do have some limitations that are important to recognise. Firstly, we only considered the MC3T3-E1 cell line [27] for these results. Secondly, if we were to apply our procedure to other more complicated experiments, where effects such as cell adhesion may be relevant, the model used may need to be more complex. This complexity may increase the need for more data so that the additional model parameters can be estimated [28, 76].

The equation learning framework in Chapter 3 has shown that we can learn continuum descriptions of discrete individual-based models beyond their coarse-grained continuum limit, and in particular of models that describe tissue growth experiments, especially for parameters outside of regimes where known models are no longer accurate. Our framework can be especially useful when considering more complicated experiments for which a mathematical model, like the one in Chapter 2, is not known. Results from our procedure could give further insight into cellular mechanisms, or other mechanisms more generally, since the learned models can be interpreted relative to these mechanisms.

The research in this thesis could be extended in many ways. Firstly, it would be of interest to apply our likelihood-based inference methods of Chapter 2 to tissue growth experiments on other pore geometries or with different cell lines, as this could be explored without any modification to our approach. Secondly, a further examination of how curvature effects influence tissue growth would be worthwhile, using our approach of predicting results on new experiments to systematically vary the curvature on a geometry while visualising the density profiles for each geometry. The equation learning work in Chapter 3 could be extended into higher dimensions of space, where the main change to our framework would be the need for a more complex method for solving the PDEs numerically [135–137]. Applying the framework to heterogeneous cell populations would also be of interest [45], where the main difficulty would be in changing the nonlinear diffusion function to vary in space to allow for varying rates of mechanical relaxation. Lastly, it would be useful to consider applying the equation learning framework in Chapter 3 to experimental data like from the experiments in Chapter 2, which would also involve exploring how uncertainty quantification can be incorporated into the framework, for example using bootstrapping [50] or Bayesian inference [140] together with methods from multimodel inference [144, 145]. Similarly, for such an application it may be necessary to borrow ideas from recent work in equation learning to extend the framework, such as using denoising methods for estimating finite differences from noisy data or other extensions [140, 146–153]. Applying the framework in Chapter 3 to experimental data will also require values for the discrete model parameters, which may be taken from the literature [34] or calibrated to the experimental data directly [154–162].

This thesis shows the importance of applying ideas from mathematical and statistical modelling to problems in tissue engineering. We have used modelling to examine images from a tissue growth experiment to test the question of whether cellular mechanisms are independent of pore geometry, and further used statistical methods to demonstrate how experimental design and experimenting planning can be improved for new geometries. The procedure we use for this is highly general, and can be used more broadly in tissue engineering. We have shown how, even in cases where a model describing the experiments is not known, equation learning can be applied to learn such a continuum model that can then be used for analysis. From the above discussion, we can see that the results in this thesis can be extended in many directions.

Appendix A

Equation learning example

This appendix serves to provide a didactic example of how equation learning is traditionally applied, in the spirit of Brunton et al. [47]. We will conclude this appendix by clarifying similarities between this traditional approach to how we apply equation learning in Chapter 3.

A.1 Learning an ordinary differential equation

We give an example where we learn data from an ordinary differential equation (ODE). The data we consider is simulated from a weak Allee effect model, given by [163]

$$\frac{\mathrm{d}C(t)}{\mathrm{d}t} = rC(t)\left(1 - \frac{C(t)}{K}\right)\left(1 + \frac{C(t)}{A}\right), \quad 0 < t \leq T, \tag{A.1}$$

where $C(t)$ is the population density with $C(0) = 1/24$, $K = 1$ is the carrying capacity density, $A = 50$ defines a deviation away from logistic growth via $1 + C(t)/A$, $r = 1/10$ is the proliferation rate, and $T = 100$ is the final time. The data we will be using for this exercise is shown in Figure A.1. For the purposes of this exercise, we do not consider adding any noise to the data, similar to what we have in Chapter 3.

The data we show in Figure A.1 shows a time series with data given by $\{(t_j, C_j)\}_{j=1}^{M}$, where $M = 1001$, $t_j = (j-1)\Delta t$ with $\Delta t = 1/10$, and $C_j = C(t_j)$. Our aim is to use this data to learn the original equation

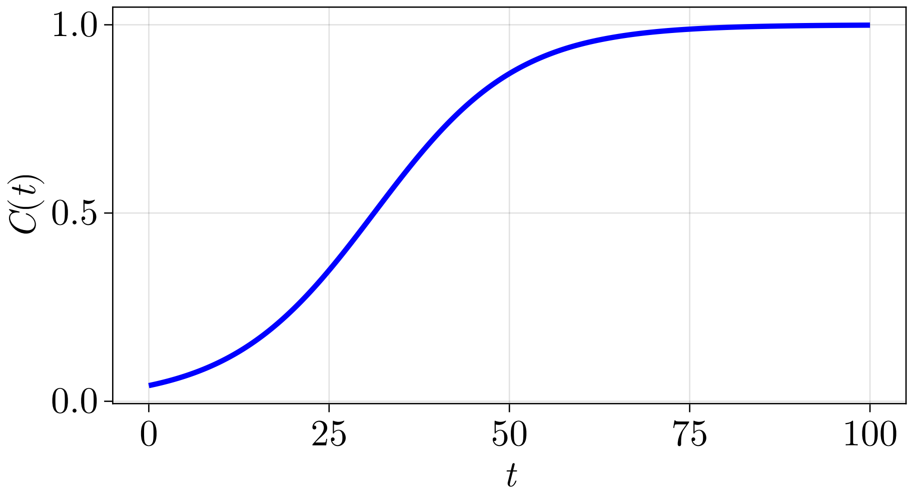

Figure A.1: Simulated data from the weak Allee effect model (A.1) with $C(0) = 1/24$, $K = 1$, $B = 1/2$, $r = 1$, and $T = 100$.

(A.1). To learn these equations, the procedure starts by writing

$$\mathbf{C} = \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_M \end{bmatrix} \in \mathbb{R}^{M\times 1}, \quad \dot{\mathbf{C}} = \begin{bmatrix} \mathrm{d}C_1/\mathrm{d}t \\ \mathrm{d}C_2/\mathrm{d}t \\ \vdots \\ \mathrm{d}C_M/\mathrm{d}t \end{bmatrix} \in \mathbb{R}^{M\times 1}, \tag{A.2}$$

where the derivatives $\mathrm{d}C_j/\mathrm{d}t$ are estimated using finite differences, $j = 1, \ldots, M$. Next, we define a library $\boldsymbol{\Theta}(\mathbf{C}) \in \mathbb{R}^{M\times p}$ which contains terms that we might expect to appear in the learned model. For this problem, we suppose that the learned model might contain terms up to quartic order so that

$$\boldsymbol{\Theta}(\mathbf{C}) = \begin{bmatrix} 1 & C_1 & C_1^2 & C_1^3 & C_1^4 \\ 1 & C_2 & C_2^2 & C_2^3 & C_2^4 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & C_M & C_M^2 & C_M^3 & C_M^4 \end{bmatrix} \in \mathbb{R}^{M\times 5}. \tag{A.3}$$

Thus, if we suppose that the model takes the form

$$\frac{\mathrm{d}C}{\mathrm{d}t} = \xi_0 + \xi_1 C + \xi_2 C^2 + \xi_3 C^3 + \xi_4 C^4, \tag{A.4}$$

then the library (A.3) gives the system

$$\dot{\mathbf{C}} = \boldsymbol{\Theta}(\mathbf{C})\boldsymbol{\xi}, \tag{A.5}$$

where $\boldsymbol{\xi} = (\xi_0, \xi_1, \xi_2, \xi_3, \xi_4)^{\mathsf{T}}$. In the notation from Chapter 1, where we wrote $\partial q/\partial t = \mathcal{N}(q, \mathcal{D}, \boldsymbol{\theta})$, we have $q = C$, $\mathcal{D} = \emptyset$, $\boldsymbol{\theta} = \boldsymbol{\xi}$, and $\mathcal{N}(q, \mathcal{D}, \boldsymbol{\theta}) = \xi_0 + \xi_1 C + \xi_2 C^2 + \xi_3 C^3 + \xi_4 C^4$.

To solve (A.5), sparse regression is used. For this problem, we use LASSO regression so that [47]

$$\boldsymbol{\xi} = \operatorname*{argmin}_{\boldsymbol{\xi}'} \left\{ \left\| \dot{\mathbf{C}} - \boldsymbol{\Theta}(\boldsymbol{\xi})\boldsymbol{\xi}' \right\|_2 + \lambda \|\boldsymbol{\xi}'\|_1 \right\}, \tag{A.6}$$

which simultaneously balances goodness of fit with model complexity. For $\boldsymbol{\xi}$ to match (A.1), we need $\boldsymbol{\xi} = (0, 0.1, -0.098, -0.002, 0)^{\mathsf{T}}$. Optimising (A.6) with $\lambda = 1/50$ gives

$$\boldsymbol{\xi} = (3.24 \times 10^{-6}, 0.1, -0.098, -0.00191, -8.13 \times 10^{-5})^{\mathsf{T}}, \tag{A.7}$$

which has a relative error of 0.087% compared to the true $\boldsymbol{\xi}$; sequential thresholded least squares could be used to find a $\boldsymbol{\xi}$ which iteratively zeros out components [47]. We show in Figure A.2 the comparison between the data and the solution to the learned ODE.

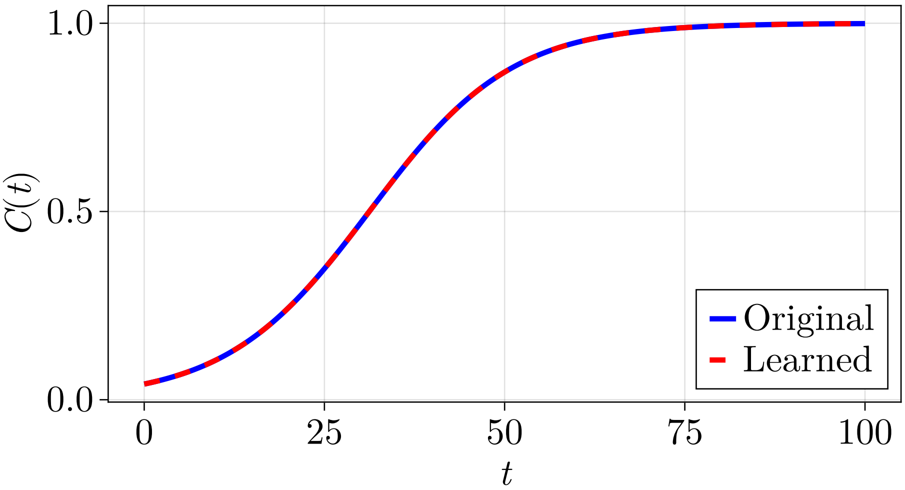

Figure A.2: Comparison of the learned ODE (A.4) with $\boldsymbol{\xi}$ given by (A.7) to the original data from Figure A.1.

A.2 Similarities with the approach in Chapter 3

The example we give is useful as it shows some similarities with our approach from Chapter 3. Firstly, both of these problems lead to a matrix problem for the coefficients to be estimated, although the method used for solving it differs. Secondly, both methods simultaneously balance model complexity with the goodness of fit, a feature common to all equation learning methods. Lastly, both methods use time series data and finite differences as input, although in our case we have data not of density but of position that then gets converted into the appropriate variable for the differential equation.

There are also some differences. While the method of [47] learns the entire problem at once, our method instead learns a constrained problem so that it takes the form of a conservation law. In particular, we use

$$\mathcal{N}(q, \mathcal{D}, \boldsymbol{\theta}) = \frac{\partial}{\partial x}\left(D(q)\frac{\partial q}{\partial x}\right) + R(q) \tag{A.8}$$

so that $\mathcal{D} = \{\partial/\partial x\}$ and $\boldsymbol{\theta}$ are the coefficients parametrising $D(q)$ and $R(q)$. Additionally, rather than using (A.6) to estimate the parameters, we use a least squares problem to estimate the parameters and then use a loss function similar to (A.6) for iteratively refining the parameter vector.

Bibliography

[1] VandenHeuvel D, Devlin B, Buenzli P, Woodruff M, Simpson M. 2023 New computational tools and experiments reveal how geometry affects tissue growth in 3D printed scaffolds. *Chemical Engineering Journal* **475**, 145776. x, 8, 9.

[2] VandenHeuvel DJ, Buenzli PR, Simpson MJ. 2023 Pushing coarse-grained models beyond the continuum limit using equation learning. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* p. 20230619. x, 8, 9.

[3] Ikada Y. 2006 Challenges in tissue engineering. *Journal of the Royal Society Interface* **3**, 589–601. 1.

[4] Khademhosseini A, Langer R. 2016 A decade of progress in tissue engineering. *Nature Protocols* **11**, 1775–1781. 1, 14.

[5] Lysaght M, Reyes J. 2004 The growth of tissue engineering. *Tissue Engineering* **7**, 485–493. 1, 14.

[6] Evans ND, Oreffo ROC, Healy E, Thurner PJ, Man YH. 2013 Epithelial mechanobiology, skin wound healing, and the stem cell niche. *Journal of the Mechanical Behavior of Biomedical Materials* **28**, 397–409. 1.

[7] Dehkordi AN, Babaheydari FM, Chehelgerdi M, Dehkordi SR. 2019 Skin tissue engineering: wound healing basd on stem-cell-based therapeutic strategies. *Stem Cell Research & Therapy* **10**. 1.

[8] Ho J, Walsh C, Yue D, Dardik A, Cheema U. 2017 Current advancements and strategies in tissue engineering for wound healing: A comprehensive review. *Advances in Wound Care* **6**, 191–209. 1.

[9] Hu MS, Maan ZN, Wu JC, Rennert RC, Hong WX, Lai TS, Cheung ATM, Walmsley GG, Chung MT, McArdle A, Longaker MT, Lorenz HP. 2014 Tissue engineering and regenerative repair in wound healing. *Annals of Biomedical Engineering* **42**, 1494–1507. 1.

[10] Do AV, Khorsand B, Geary S, Salem A. 2018 3D printing of scaffolds for tissue regeneration applications. *Advanced Healthcare Materials* **4**, 1742–1762. 1, 14.

[11] Zieliński P, Gudeti P, Rikmanspoel T, MK Włodarczyk-Biegun. 2022 3D printing of bio-instructive materials: Toward directing the cell. *Bioactive Materials* **19**, 292–327. 1, 14.

[12] Mani M, Sadia M, Jaganathan S, Khudzari A, Supriyanto E, Saidin S, Ramakrishna S, Ismail A, Faudzi A. 2022 A review on 3D printing in tissue engineering applications. *Journal of Polymer Engineering* **42**, 243–265. 1, 14, 16.

[13] Buenzli P, Lanaro M, Wong C, McLaughlin M, Allenby M, Woodruff M, Simpson M. 2020 Cell proliferation and migration explain pore bridging dynamics in 3D printed scaffolds of different pore size. *Acta Biomaterialia* **114**, 285–295. 1, 2, 14, 15, 16, 20, 22, 28, 51, 122.

[14] Lanaro M, Mclaughlin M, Simpson M, Buenzli P, Wong C, Allenby M, Woodruff M. 2021 A quantitative analysis of cell bridging kinetics on a scaffold using computer vision algorithms. *Acta Biomaterialia* **136**, 429–440. 1, 15, 16, 18, 19.

[15] Kilian D, von Witzleben M, Lanaro M, Wong CS, Vater C, Lode A, Allenby MC, Woodruff MA, Gelinsky M. 2022 3D plotting of calcium phosphate cement and melt electrowriting of polycaprolactone microfibers in one scaffold: A hybrid additive manufacturing process. *Journal of Functional Biomaterials* **13**, 75. 1.

[16] Forrestal D, Klein T, Woodruff M. 2016 Challenges in engineering large customized bone constructs. *Biotechnology and Bioengineering* **114**, 1129–1139. 1, 14, 16.

[17] Hrynevich A, Elçi B, Haigh J, McMaster R, Youssef A, Blum C, Blunk T, Hochleitner G, Groll J, Dalton P. 2018 Dimension-based design of melt electrowritten scaffolds. *Small* **22**, 1800232. 1, 14.

[18] Paxton N, Ren J, Ainsworth M, Solanki A, Jones J, Allenby M, Stevens M, Woodruff M. 2019 Rheological characterization of biomaterials directs additive manufacturing of strontium-substituted bioactive glass/polycaprolactone microfibers. *Macromolecular Rapid Communications* **40**, 1900019. 1, 14.

[19] Dzobo K, Thomford N, Senthebane D, Shipange H, Rowe A, Dandara C, Pillay M, Motaung K. 2018 Advances in regenerative medicine and tissue engineering: Innovation and transformation of medicine. *Stem Cells International* **2018**, 2495848. 1, 14.

[20] Hollister S, Flanagan C, Zopf D, Morrison R, Nasser H, Patel J, Ebramzadeh E, Sangiorgio S, Wheeler M, Green G. 2015 Design control for clinical translation of 3D printed modular scaffolds. *Annals of Biomedical Engineering* **43**, 774–786. 1, 14.

[21] Nelson C, Jean R, Tan J, Liu W, Sniadecki N, Spector A, Chen C. 2005 Emergent patterns of growth controlled by multicellular form and mechanics. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 11594–11599. 1, 14, 122.

[22] Bidan C, Kommareddy K, Rumpler M, Kollmannsberger P, Bréchet Y, Fratzi P, Dunlop J. 2012 How linear tension converts to curvature: Geometric control of bone tissue growth. *PLOS ONE* **7**, e36336. 1, 14, 122.

[23] Rumpler M, Woesz A, Dunlop J, van Dongen J, Fratzl P. 2008 The effect of geometry on three-dimensional tissue growth. *Journal of the Royal Society Interface* **5**, 1173–180. 1, 14, 122.

[24] Browning A, Maclaren O, Buenzli P, Lanaro M, Allenby MC, Woodruff M, Simpson M. 2021 Model-based data analysis of tissue growth in thin 3D printed scaffolds. *Journal of Theoretical Biology* **528**, 110852. 1, 3, 5, 14, 15, 16, 20, 21, 22, 24, 28, 36, 41, 122.

[25] Jin W, Lo KY, Chou SE, McCue S, Simpson M. 2018 The role of initial geometry in experimental models of wound closing. *Chemical Engineering Science* **179**, 221–226. 1, 15, 122.

[26] Buenzli P, Simpson M. 2022 Curvature dependences of wave propagation in reaction-diffusion models. *Proceedings of the Royal Society A* **478**, 20220582. 1, 15, 40, 122.

[27] Yan XZ, Yang W, Yang F, Kersten-Niessen M, Jansen J, Both S. 2014 Effects of continuous passaging on minearlization of MC3T3-E1 cells with improved osteogenic culture protocol. *Tissue Engineering* **20**, 198–204. 2, 16, 18, 123.

[28] Simpson M, Maclaren O. 2023 A profile likelihood-based workflow for identifiability analysis, estimation, and prediction with mechanistic mathematical models. *bioRxiv*. 2, 22, 24, 25, 40, 50, 53, 123.

[29] Murray JD. 2002 *Mathematical Biology I*. New York: Springer 3rd edition. 3, 4.

[30] Maini P, McElwain D, Leavesley D. 2004 Traveling wave model to interpret a wound-healing cell migration assay for human peritoneal mesothelial cells. *Tissue Engineering* **10**, 475–482. 3, 21, 22, 28, 50.

[31] Jin W, Shah ET, Penington CJ, McCue SW, Maini PK, Simpson MJ. 2017 Logistic proliferation of cells in scratch assay is delayed. *Bulletin of Mathematical Biology* **79**, 1028–1050. 3.

[32] Sherratt J, Murray J. 1990 Models of epidermal wound healing. *Proceedings of the Royal Society: Series B* **241**, 29–36. 3, 5, 17, 20.

[33] Evans DJ, Morriss G. 2008 *Statistical mechanics of nonequilibrium liquids*. Cambridge: Cambridge University Press. 4, 6, 64.

[34] Osborne JM, Fletcher AG, Pitt-Francis JM, Maini PK, Gavaghan DJ. 2017 Comparing individual-based approaches to modelling the self-organization of multicellular tissues. *PLoS Computational Biology* **13**, e1005387. 5, 96, 124.

[35] Fletcher AG, Breward CJW, Chapman SJ. 2012 Mathematical modeling of monoclonal conversion in the colonic crypt. *Journal of Theoretical Biology* **300**, 118–133. 5.

[36] Pathmanathan P, Cooper J, Fletcher A, Mirams G, Murray P, Osborne J, Pitt-Francis J, Walter A, Chapman SJ. 2009 A computa-

tional study of discrete mechanical tissue models. *Physical Biology* **6**, 036001. 5.

[37] van Leeuwen IMM, Mirams GR, Walter A, Fletcher A, Murray P, Osborne J, Varma S, Young SJ, Cooper J, Doyle B, Pitt-Francis J, Momtahan L, Pathmanathan P, Whiteley JP, Chapman SJ, Gavaghan DJ, Jensen OE, King JR, Maini PK, Waters SL, Byrne HM. 2009 An integrative computational model for intestinal tissue renewal. *Cell Proliferation* **42**, 617–636. 5.

[38] Baker RE, Parker A, Simpson MJ. 2019 A free boundary model of epithelial dynamics. *Journal of Theoretical Biology* **481**, 61–74. 5, 6, 63, 66, 68, 69, 95, 99, 100, 101, 104.

[39] Azuaje F. 2011 Computational discrete models of tissue growth and regeneration. *Briefings in Bioinformatics* **12**, 64–77. 5.

[40] Murphy RJ, Buenzli PR, Baker RE, Simpson MJ. 2020 Mechanical cell competition in heterogeneous epithelial tissues. *Bulletin of Mathematical Biology* **82**, 130. 6, 63, 66, 70, 89, 95, 114, 116.

[41] Murray PJ, Edwards CM, Tindall MJ, Maini PK. 2009 From a discrete to a continuum model of cell dynamics in one dimension. *Physical Review E* **80**, 031912. 6, 63, 66, 68, 69, 95.

[42] Murray PJ, Edwards CM, Tindall MJ, Maini PK. 2012 Classifying general nonlinear force laws in cell-based models via the continuum limit. *Physical Review E* **85**, 021921. 6, 63, 66, 95.

[43] Hufton PG, Lin YT, Galla T. 2019 Model reduction methods for population dynamics with fast-switching environments: Reduced master equations, stochastic differential equations, and applications. *Physical Review E* **99**, 032122. 6, 63.

[44] Ihle T. 2016 Chapman-Enskog expansion for the Vicsek model of self-propelled particles. *Journal of Statistical Mechanics: Theory and Experiment* p. 083205. 6, 63.

[45] Murphy RJ, Buenzli PR, Baker RE, Simpson MJ. 2019 A one-dimensional individual-based mechanical model of cell movement in

heterogeneous tissues and its coarse-grained approximation. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **475**, 20180838. 6, 63, 66, 68, 96, 123.

[46] Chopard B, Droz M. 1998 *Cellular automata modeling of physical systems*. Cambridge: Cambridge University Press. 6, 64.

[47] Brunton SL, Proctor JL, Kutz JN. 2016 Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 3932–3937. 6, 7, 64, 65, 75, 76, 77, 95, 125, 127, 128.

[48] Nardini JT, Baker RE, Simpson MJ, Flores KB. 2021 Learning differential equation models from stochastic agent-based model simulations. *Journal of the Royal Society Interface* **18**, 20200987. 6, 7, 63, 65, 77, 95.

[49] Lagergren JH, Nardini JT, Baker RE, Simpson MJ, Flores KB. 2020 Biologically-informed neural networks guide mechanistic modelling from sparse experimental data. *PLoS Computational Biology* **16**, e1008462. 6, 7, 64, 65, 94.

[50] VandenHeuvel DJ, Drovandi C, Simpson MJ. 2022 Computationally efficient mechanism discovery for cell invasion with uncertainty quantification. *PLoS Computational Biology* **18**, e1010599. 6, 7, 65, 74, 79, 95, 96, 109, 123.

[51] Rudy SH, Brunton SL, Proctor JL, Kutz JN. 2017 Data-driven discovery of partial differential equations. *Science Advances* **3**, e1602614. 7, 64, 65, 75, 76, 77, 95.

[52] Simpson MJ, Baker RE, Buenzli PR, Nicholson R, Maclaren OJ. 2022 Reliable and efficient parameter estimation using approximation continuum limit descriptions of stochastic models. *Journal of Theoretical Biology* **549**, 111201. 7, 65, 94.

[53] Johnston ST, Ross JV, Binder BJ, McElwain DLS, Haridas P, Simpson MJ. 2016 Quantifying the effect of experimental design choices for *in vitro* scratch assays. *Journal of Theoretical Biology* **400**, 19–31. 14.

[54] Egan P, Ferguson S, Shea K. 2017 Design of hierarchical three-dimensional printed scaffolds considering mechanical and biological factors for bone tissue engineering. *Journal of Mechanical Design* **139**, 061401. 14.

[55] Paris M, Götz A, Hettrich I, Bidan C, Dunlop J, Razi H, Zizak I, Hutmacher D, Fratzl P, Buda G, Wagermaier W, Ciptria A. 2017 Scaffold curvature-mediated novel biomineralization process originates a continuous soft tissue-to-bone interface. *Acta Biomaterialia* **60**, 64–80. 14, 122.

[56] Callens S, Uyttendaele R, Fratila-Apachitei L, Zadpoor A. 2020 Substrate curvature as a cue to guide spatiotemporal cell and tissue organization. *Biomaterials* **232**, 119739. 14, 122.

[57] Bidan C, Kommareddy K, Rumpler M, Kollmannsberger P, Fratzl P, Dunlop J. 2012 Geometry as a factor for tissue growth: Towards shape optimization of tissue engineering scaffolds. *Advanced Healthcare Materials* **2**, 186–194. 14.

[58] Callens S, Fan D, van Hengel I, Minneboo M, Díaz-Payno P, Stevens M, Fratila-Apachitei L, Zadpoor A. 2023 Emergent collective organization of bone cells in complex curvature fields. *Nature Communicatons* p. 855. 14, 122.

[59] Alias M, Buenzli P. 2018 Osteoblasts infill irregular pores under curvature and porosity controls: a hypothesis-testing analysis of cell behaviours. *Biomechanics and Modeling in Mechanobiology* **17**, 1357–1371. 14, 122.

[60] Alias M, Buenzli P. 2017 Modeling the effect of curvature on the collective behavior of cells growing new tissue. *Biophysical Journal* **112**, 193–204. 14, 122.

[61] Alias M, Buenzli P. 2019 A level-set method for the evolution of cells and tissue during curvature-controlled growth. *International Journal for Numerical Methods in Biomedical Engineering* **36**, e3279. 14, 122.

[62] Hegarty-Cremer S, Simpson M, Andersen T, Buenzli P. 2021 Modelling cell guidance and curvature control in evolving biological tissues. *Journal of Theoretical Biology* **520**, 110658. 14, 122.

[63] Pawitan Y. 2001 *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford: Oxford University Press. 17, 23, 24, 25.

[64] Royston P. 2007 Profile likelihood for estimation and confidence intervals. *The Stata Journal* **7**, 376–387. 17.

[65] Simpson M, Walker S, Studerus E, McCue S, Murphy R, Maclaren O. 2023 Profile likelihood-based parameter and predictive interval analysis guides model choice for ecological population dynamics. *Mathematical Biosciences* **355**, 108950. 17, 22, 24, 25.

[66] Villaverde A, Tsiantis N, Banga J. 2019 Full observability and estimation of unknown inputs, states and parameters of nonlinear biological models. *Journal of the Royal Society Interface* **16**, 20190043. 17.

[67] Simpson M. 2009 Depth-averaging errors in reactive transport modelling. *Water Resources Research* **45**, W02505. 20.

[68] Browning A, Haridas P, Simpson M. 2018 A Bayesian sequential learning framework to parameterise continuum models of melanoma invasion into human skin. *Bulletin of Mathematical Biology* **81**, 676–698. 21.

[69] Treloar K, Simpson M. 2013 Sensitivity of edge detection methods for quantifying cell migration assays. *PLoS ONE* **8**, e67389. 21.

[70] Vittadello S, McCue S, Gunasingh G, Hass N, Simpson M. 2018 Mathematical models for cell migration with real-time cell cycle dynamics. *Biophysical Journal* **114**, 1241–1253. 21.

[71] Simpson M, Murphy R, Maclaren O. 2023 Modelling count data with partial differential equation models in biology. *bioRxiv*. 22.

[72] Simpson M, Browning A, Warne D, Maclaren O, Baker R. 2022 Parameter identifiability and model selection for sigmoid population growth models. *Journal of Theoretical Biology* **535**, 110998. 22.

[73] Casella G, Berger R. 2002 *Statistical Inference*. Pacific Grove, CA: Duxbury 2nd edition. 23, 25.

[74] Wasserman L. 2013 *All of Statistics: A Concise Course in Statistical Inference*. New York: Springer Science & Business Media. 24.

[75] Browning A, Sharp J, Murphy R, Gunasingh G, Lawson B, Burrage K, Haass N, Simpson M. 2021 Quantitative analysis of tumour spheroid structure. *eLife* **10**, e73020. 24.

[76] Raue A, Kreutz C, Maiwald T, Klingmüller U, Timmer J. 2011 Addressing parameter identifiability by model-based experimentation. *IET Systems Biology* **5**, 120–130. 24, 40, 123.

[77] VandenHeuvel D. 2023 ProfileLikelihood.jl: Methods for profile likelihood analysis. https://doi.org/10.5281/zenodo.7827704. Version 0.2.3. 25, 26, 52.

[78] Byrne S. 2014 KernelDensity.jl: Kernel density estimators for Julia. https://github.com/JuliaStats/KernelDensity.jl. Version 0.6.5. 25.

[79] Jin W, Shah E, Penington C, McCue S, Chopin L, Simpson M. 2016 Reproducibility of scratch assays is affected by the initial degree of confluence: Experiments, modelling and model selection. *Journal of Theoretical Biology* **390**, 136–145. 28.

[80] Treloar K, Simpson M, McElwain D, Baker R. 2014 Are in vitro estimates of cell diffusivity and cell proliferation rate sensitive to assay geometry?. *Journal of Theoretical Biology* **356**, 71–84. 39.

[81] Bezanson J, Edelman A, Karpinski S, Shah V. 2017 Julia: A fresh approach to numerical computing. *SIAM Review* **59**, 65–98. 41.

[82] Rokne J. 1991 The area of a simple polygon. In Arvo J, editor, *Graphics Gems II* pp. 5–6. San Diego, California: Morgan Kaufmann. 42.

[83] Versteeg HK, Malalasekera W. 2007 *An introduction to computational fluid mechanics*. Harlow: Prentice Hall 2nd edition. 44, 102.

[84] VandenHeuvel D. 2023 DelaunayTriangulation.jl: Delaunay triangulation and Voronoi tessellations of planar point sets. https://doi.org/10.5281/zenodo.7964424. Version 0.7.2. 44.

[85] Rackauckas C, Nie Q. 2017 DifferentialEquations.jl — A performant and feature-rich ecosystem for solving differential equations in Julia. *Journal of Open Research Software* **5**, 15. 46, 53, 103, 106.

[86] Davis TA, Palamadai Natarajan E. 2010 Algorithm 907: KLU, A direct sparse solver for circuit simulation problems. *ACM Transactions on Mathematical Software* **37**, 1–17. 46, 103, 106.

[87] Hosea ME, Shampine LF. 1996 Analysis and implementation of TR-BDF2. *Applied Numerical Mathematics* **20**, 21–37. 46, 103, 106.

[88] VandenHeuvel D. 2023 FiniteVolumeMethod.jl: Solver for two-dimensional conservation equations using the finite volume method. https://doi.org/10.5281/zenodo.7950651. Version v0.4.7. 46.

[89] Stagner L. 2017 ConcaveHull.jl: Julia package for calculating 2D concave/convex hulls. https://github.com/lstagner/ConcaveHull.jl. Version 1.2.0. 48.

[90] Moreira A, Santos M. 2007 Concave hull: A k-nearest neighbours approach for the computation of the region occupied by a set of points. In Braz J, Vázquez PP, Madeiras Pereira J, editors, *Proceedings of the 2nd International Conference on Computer Graphics Theory and Applications (GRAPP 2007)* vol. GM/R pp. 61–68. INSTICC - Institute for Systems and Technologies of Information, Control and Communication. 48.

[91] Johnson SG. 2013 NLopt.jl: Package to call the NLopt nonlinear-optimization library from the Julia language. https://github.com/JuliaOpt/NLopt.jl. Version 0.6.5. 52, 118.

[92] Johnson SG. 2010 The NLopt nonlinear-optimization package. https://github.com/stevengj/nlopt. Version 2.7.1. 52, 118.

[93] Powell M. 2009 The BOBYQA algorithm for bound constrained optimization without derivatives. Technical report Department of Applied Mathematics and Theoretical Physics Cambridge, England. 52.

[94] Boiger R, Hasenauer J, Hroß S, Kaltenbacher B. 2016 Integration based profile likelihood calculation for PDE constrained parameter estimation problems. *Inverse Problems* **32**, 125009. 52.

[95] Maclaren OJ, Byrne HM, Fletcher AG, Maini PK. 2015 Models, measurement and inference in epithelial tissue dynamics. *Mathematical Biosciences and Engineering* **12**, 1321–1340. 63.

[96] Turner S, Sherratt JA, Painter KJ, Savill NJ. 2004 From a discrete to a continuous model of biological cell movement. *Physical Review E* **69**, 021910. 63.

[97] Alber M, Chen N, Lushnikov PM, Newman SA. 2007 Continuous macroscopic limit of a discrete stochastic model for interaction of living cells. *Physical Review Letters* **99**, 168102. 63.

[98] Zmurchok C, Bhaskar D, Edelstein-Keshet L. 2018 Coupling mechanical tension and GTPase signaling to generate cell and tissue dynamics. *Physical Biology* **15**, 046004. 63.

[99] Marée AFM, Grieneisen VA, Edelstein-Keshet L. 2012 How cells integrate complex stimuli: The effect of feedback from phosphoinositides and cell shape on cell polarization and motility. *PLoS Computational Biology* **8**, e1002402. 63.

[100] Mason J, Jack RL, Bruna M. 2023 Macroscopic behaviour in a two-species exclusion process via the method of matched asymptotics. *Journal of Statistical Physics* **190**. 63.

[101] Bruna M, Chapman SJ, Schmidtchen M. 2023 Derivation of a macroscopic model for Brownian hard needles. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **479**, 20230076. 63.

[102] Bruna M, Chapman SJ. 2012 Diffusion of multiple species with excluded-volume effects. *The Journal of Chemical Physics* **137**, 204116. 63.

[103] Bruna M, Chapman SJ. 2012 Excluded-volume effects in the diffusion of hard spheres. *Physical Review E* **85**, 011103. 63.

[104] Tambyah TA, Murphy RJ, Buenzli PR, Simpson MJ. 2021 A free boundary mechanobiological model of epithelial tissues. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **476**, 20200528. 63, 66.

[105] Murphy RJ, Buenzli PR, Tambyah TA, Thompson EW, Hugo HJ, Baker RE, Simpson MJ. 2021 The role of mechanical interactions in EMT. *Physical Biology* **18**, 046001. 63, 66.

[106] Fozard JA, Byrne HM, Jensen OE, King JR. 2010 Continuum approximations of individual-based models for epithelial monolayers. *Mathematical Medicine and Biology: A Journal of the IMA* **27**, 39–74. 63, 66.

[107] Supekar R, Song B, Hastewell A, Choi GPT, Mietke A, Dunkel J. 2023 Learning hydrodynamic equations for active matter from particle simulations and experiments. *Proceedings of the National Academy of Sciences of the United States of America* **120**, e2206994120. 63.

[108] Español P. 2004 Statistical mechanics of coarse-graining. *Novel Methods in Soft Matter Computing* **640**, 69–115. 63.

[109] Middleton AM, Fleck C, Grima R. 2014 A continuum approximation to an off-lattice individual-cell based model of cell migration and adhesion. *Journal of Theoretical Biology* **359**, 220–232. 63.

[110] Jeon J, Quaranta V, Cummings PT. 2010 An off-lattice hybrid discrete-continuum model of tumor growth and invasion. *Biophysical Journal* **98**, 37–47. 63.

[111] Osborne JM, Walker A, Kershaw SK, Mirams GR, Fletcher AG, Pathmanathan P, Gavaghan D, Jensen OE, Maini PK, Byrne HM. 2010 A hybrid approach to multi-scale modelling of cancer. *Philosophical Transactions of the Royal Society A* **368**, 5013–5028. 63.

[112] Buttenschön A, Edelstein-Keshet L. 2020 Bridging from single to collective cell migration: A review of models and links to experiments. *PLoS Computational Biology* **16**, e1008411. 63.

[113] Surendran A, Plank MJ, Simpson MJ. 2020 Population dynamics with spatial structure and an Allee effect. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **476**, 20200501. 63.

[114] Lorenzi T, Murray PJ, Ptashnyk M. 2020 From individual-based mechanical models of multicellular systems to free-boundary problems. *Interfaces and Free Boundaries: Mathematical Analysis, Computation and Applications* **22**, 205–244. 63.

[115] Romeo N, Hastewell A, Mietke A, Dunkel J. 2021 Learning developmental mode dynamics from single-cell trajectories. *eLife* **10**, e68679. 63.

[116] Pughe-Sanford JL, Quinn S, Balabanski T, Grigoriev RO. 2023 Computing chaotic time-averages from a small number of periodic orbits. (https://arxiv.org/abs/2307.09626v1). 63.

[117] Vo BN, Drovandi CC, Pettitt AN, Simpson MJ. 2015 Quantifying uncertainty in parameter estimates for stochastic models of collective cell spreading using approximate Bayesian computation. *Mathematical Biosciences* **263**, 133–142. 63.

[118] Codling EA, Plank MJ, Benhamou S. 2008 Random walk models in biology. *Journal of the Royal Society Interface* **5**, 813–834. 63.

[119] Simpson MJ, Landman KA, Hughes BD. 2010 Cell invasion with proliferation mechanisms motivated by time-lapse data. *Physica A: Statistical Mechanics and its Applications* **389**, 3779–3790. 63, 64, 66.

[120] Hughes BD. 1995 *Random walks and random environments: Random walks*. Oxford: Clarendon Press. 63, 64.

[121] Nardini JT, Lagergren JH, Hawkins-Daarud A, Curtin L, Morris B, Rutter EM, Swanson KR, Flores KB. 2020 Learning equations from biological data with limited time samples. *Bulletin of Mathematical Biology* **82**, 119. 64, 65.

[122] Lagergren JH, Nardini JT, Lavigne GM, Rutter EM, Flores KB. 2020 Learning partial differential equations for biological transport models from noisy spatio-temporal data. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **476**, 20190800. 64.

[123] Yamashita T, Yamashita K, Kamimura R. 2007 A stepwise AIC method for variable selection in linear regression. *Communications in Statistics - Theory and Methods* **36**, 2395–2403. 65, 76.

[124] Guillot C, Lecuit T. 2013 Mechanics of epithelial tissue homeostasis and morphogenesis. *Science* **340**, 1185–1189. 66.

[125] Bragulla HH, Homberger DG. 2009 Structure and functions of keratin proteins in simple, stratified, keratinized and cornified epithelia. *Journal of Anatomy* **214**, 516–559. 66.

[126] Paster I, Stojadinovic O, Yin NC, Ramirez H, Nusbaum AG, Saway A, Patel SB, Khalid L, Isseroff RR, Tomic-Canic M. 2014 Epithelialization in wound healing: A comprehensive review. *Advances in Wound Care* **3**, 445–464. 66.

[127] Begnaud S, Chen T, Delacour D, Mège RM, Ladoux B. 2016 Mechanics of epithelial tissues during gap closure. *Current Opinion in Cell Biology* **42**, 52–62. 66.

[128] Paredes J, Figueiredo J, Albergaria A, Oliveira P, Carvalho J, Ribeiro AS, Caldeira J, Costa AM, Simoões-Correia J, Oliveira MJ, Pinheiro H, Pinho SS, Mateus R, Reis CA, Leite M, Fernandes MS, Schmitt F, Carneiro F, Figueiredo C, Oliveira C, Seruca R. 2012 Epithelial E-and P-cadherins: Role and clinical significance in cancer. *Biochimica et Biophysica Acta* **1826**, 297–311. 66.

[129] Hittelman WN. 2006 Genetic instability in epithelial tissues at risk for cancer. *Annals of the New York Academy of Sciences* **952**, 1–12. 66.

[130] Bezanson J, Edelman A, Karpinski S, Shah VB. 2017 Julia: A fresh approach to numerical computing. *SIAM Review* **59**, 65–98. 66.

[131] Witelski TP. 1995 Shocks in nonlinear diffusion. *Applied Mathematics Letters* **8**, 27–32. 74.

[132] Simpson MJ, Landman KA, Hughes BD, Fernando AE. 2010 A model for mesoscale patterns in motile populations. *Physica A: Statistical Mechanics and its Applications* **389**, 1412–1424. 74.

[133] Johnston ST, Baker RE, McElwain DLS, Simpson MJ. 2017 Co-operation, competition and crowding: A discrete frameworking linking allee kinetics, nonlinear diffusion, shocks and sharp-fronted travelling waves. *Scientific Reports* **7**, 42134. 74.

[134] Smith AM, Baker RE, Kay D, Maini PK. 2012 Incorporating chemical signalling factors into cell-based models of growing epithelial tissues. *Journal of Mathematical Biology* **65**, 441–463. 96.

[135] Tam AKY, Simpson MJ. 2023 Pattern formation and front stability for a moving-boundary model of biological invasion and recession. *Physica D: Nonlinear Phenomena* **444**, 133593. 96, 123.

[136] Sethian JA. 1999 *Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science.* Cambridge: Cambridge University Press. 96, 123.

[137] Macklin P, Lowengrub JS. 2008 A new ghost cell/level set method for moving boundary problems: Application to tumor growth. *Journal of Scientific Computing* **35**, 266–299. 96, 123.

[138] Morris RG, Rao M. 2019 Active morphogenesis of epithelial monolayers. *Physical Review E* **100**, 022413. 96.

[139] Chang YW, Cruz-Acuña R, Tennenbaum M, Fragkopoulos AA, García AJ, Fernández-Nieves A. 2022 Quantifying epithelial cell proliferation on curved surfaces. *Frontiers in Physics* **10**. 96.

[140] Martina-Perez S, Simpson MJ, Baker RE. 2021 Bayesian uncertainty quantification for data-driven equation learning. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **477**, 20210426. 96, 123, 124.

[141] Landau HG. 1950 Heat conduction in a melting solid. *The Quarterly Journal of Mechanics & Applied Mathematics* **8**, 81–94. 104.

[142] Furzeland RM. 1980 A comparative study of numerical methods for moving boundary problems. *IMA Journal of Applied Mathematics* **26**, 411–429. 104.

[143] Amemiya T. 1985 *Advanced econometrics.* Cambridge, MA: Harvard University Press. 112.

[144] Symonds MRE, Moussalli A. 2011 A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. *Behavioral Ecology and Sociobiology* **65**, 13–21. 124.

[145] Burnham KP, Anderson DR. 2002 *Model selection and multimodel inference.* New York: Springer 2nd edition. 124.

[146] Chartrand R. 2011 Numerical differentiation of noisy, nonsmooth data. *International Scholarly Research Notices* **2011**, 164564. 124.

[147] Kaheman K, Brunton SL, Kutz JN. 2022 Automatic differentiation to simultaneously identify nonlinear dynamics and extract noise probability distributions from data. *Machine Learning: Science and Technology* **3**, 015031. 124.

[148] Fasel U, Kutz JN, Brunton BW, Brunton SL. 2022 Ensemble-SINDy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **478**, 20210904. 124.

[149] Champion K, Lusch B, Kutz JN, Brunton SL. 2019 Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 22445–22451. 124.

[150] Xu H, Chang H, Zhang D. 2021 DL-PDE: Deep-learning based data-driven discovery of partial differential equations from discrete and noisy data. *Communications in Computational Physics* **29**, 698–728. 124.

[151] Messenger DA, Bortz DM. 2021 Weak SINDy for partial differential equations. *Journal of Computational Physics* **443**, 110525. 124.

[152] Wentz J, Doostan A. 2023 Derivative-based SINDy (DSINDy): Addressing the challenge of discovering governing equations from noisy data. *Computer Methods in Applied Mechanics and Engineering* **413**, 116096. 124.

[153] Hokanson JM, Iaccarino G, Doostan A. 2023 Simultaneous identification and denoising of dynamical systems. *SIAM Journal on Scientific Computing* **45**, A1413–A1437. 124.

[154] van der Vaart E, Beaumont MA, Johnston ASA, Sibly RM. 2015 Calibration and evaluation of individual-based models using Approximate Bayesian Computation. *Ecological Modelling* **312**, 182–190. 124.

[155] Hazelbag CM, Dushoff J, Dominic EM, Mthombothi ZE, Delva W. 2020 Calibration of individual-based models to epidemiological data: A systematic review. *PLoS Computational Biology* **16**, e1007893. 124.

[156] Browning AP, McCue SW, Binny RN, Plank MJ, Shah ET, Simpson MJ. 2018 Inferring parameters for a lattice-free model of cell migration and proliferation using experimental data. *Journal of Theoretical Biology* **437**, 251–260. 124.

[157] McCulloch J, Ge J, Ward JA, Heppenstall A, Polhill JG, Malleson N. 2022 Calibrating agent-based models using uncertainty quantification methods. *Journal of Artifical Societies and Social Simulation* **25**, 1. 124.

[158] Cess CG, Finley SD. 2023 Calibrating agent-based models to tumor images using representation learning. *PLoS Computational Biology* **19**, e1011070. 124.

[159] Lima EABF, Faghihi D, Philley R, Yang J, Virostko J, Phillips CM, Yankeelov TE. 2021 Bayesian calibration of a stochastic, multiscale agent-based model for predicting *in vitro* tumor growth. *PLoS Computational Biology* **17**, e1008845. 124.

[160] Van Liedekerke P, Palm MM, Jagiella N, Draso D. 2015 Simulating tissue mechanics with agent-based models: concepts, perspectives and some novel results. *Computational Particle Mechanics* **2**, 401–444. 124.

[161] Schneider KM, Giehl K, Baeurle SA. 2023 Development and application of an agent-based model for the simulation of the extravasation process of circulating tumor growth. *International Journal for Numerical Methods in Biomedical Engineering* **39**, e3679. 124.

[162] Simpson MJ, Treloar KK, Binder BJ, Haridas P, Manton KJ, Leavesley DI, McElwain DLS, Baker RE. 2013 Quantifying the role of cell motility and cell proliferation in a circular barrier assay. *Journal of the Royal Society Interface* **10**, 20130007. 124.

[163] Fadai NT, Simpson MJ. 2020 Population dynamics with threshold effects give rise to a diverse family of Allee effects. *Bulletin of Mathematical Biology* **82**, 74. 125.