University of Sheffield

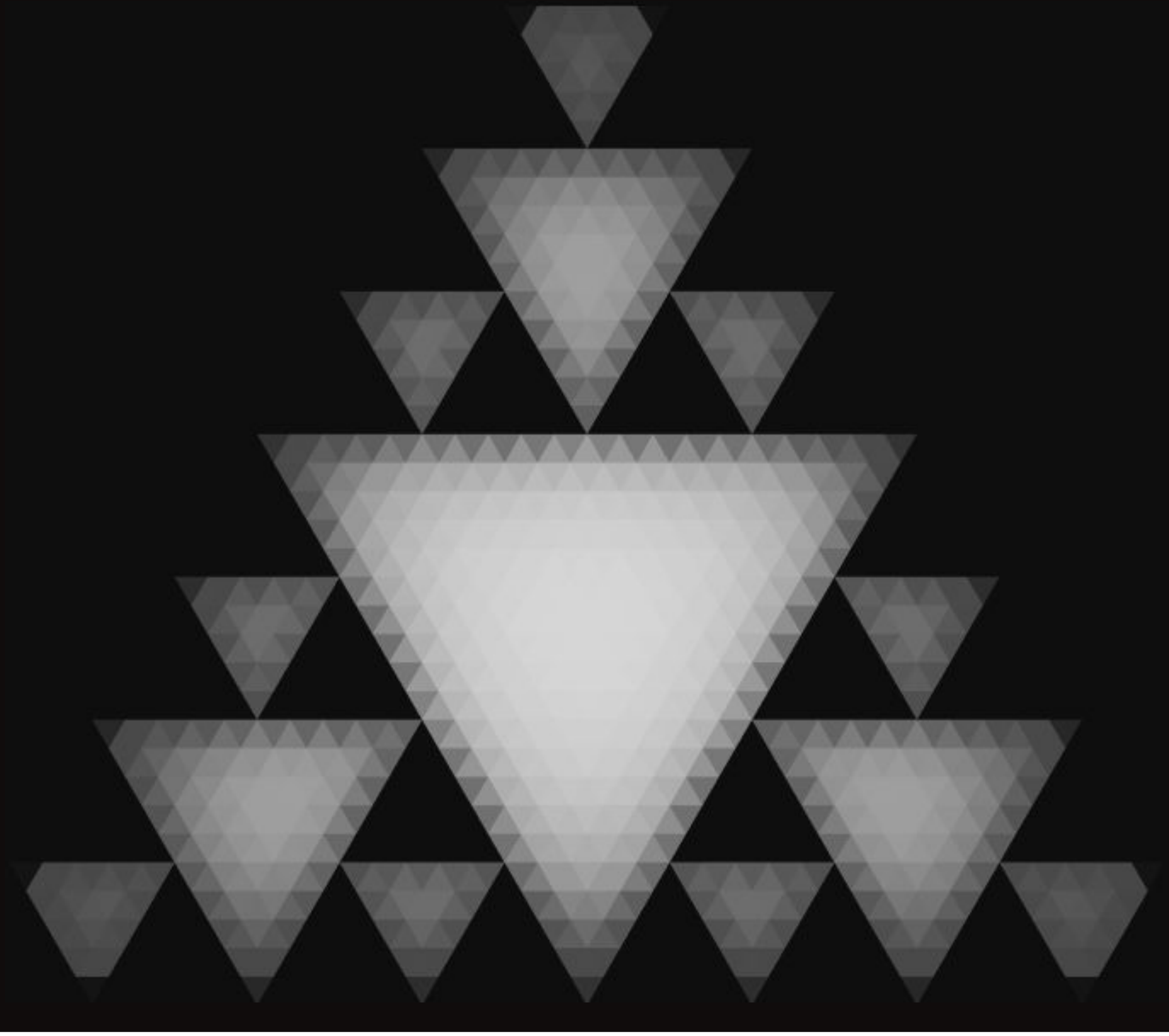
Novel Single-Cell Mathematical Modelling Approaches and Analysis of Human Stem Cell Populations

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

By Samantha Ivings

Department of Automatic Control and Systems Engineering

September 2023



Acknowledgements

I would firstly like to thank my supervisor, Prof. Daniel Coca, for his academic guidance over the course of my project. Although I was only formally supervised for a year by Dr. Giuliano Punzo, his enthusiasm and support prior to and throughout this time have been infectious and invaluable. Additionally, my thanks go to Dr. Paul Gokhale, whose conversations have never failed to generate interesting ideas and insights.

Whilst I undertook my lab work, I could not be more grateful to Dr. Larissa Butler, who kindly gave her time to help me design and conduct experiments at the level of a trained research biologist. Dr. Veronica Biga also offered support, travelling from Manchester to offer her expertise and advice.

I believe I could not have accomplished this thesis without the continued support of my mentor, Helen Shaw. Her contributions to my well-being over the past few years have gone above and beyond her duties as a mentor, and our meetings have always led to a greatly improved sense of clarity and groundedness.

A huge thanks go to Elliott and Clèm, as both their delightful company and home-cooked meals have warmed my heart when I needed it, even though they steal all my worms in Pikominos. Thanks greatly to Ben - your advice and support has been so appreciated - and to my mum, whose support has been unwavering and unconditional for as long as I can remember. Lastly, thank you to my sweet kitten Yuna, whose soft fur and sharp claws remind me to take time to think about the consequences before diving into the fluffy belly of life.

Abstract

The purpose of this thesis is to develop and apply novel modelling and analytical tools for human pluripotent stem cell (hPSC) populations. Two major types of observed heterogeneity across cells in culture are dealt with: (1) in the varied expression levels of several key genetic markers both before and after differentiation, and (2) in specific genetic mutations at the chromosomal level which confer advantages for the cells. The aim of (1) is to enable the novel development of efficient and low-cost differentiation protocols, while (2) seeks to improve quality control in stem cell research as unwanted genetic mutations are challenging to automatically detect.

A novel spatial analysis is developed with the aim of understanding how continuous single-cell Nanog expression levels are linked to the local cellular environment. Nanog expression per cell is thus shown to be inversely proportional to the cell's neighbour count. These results drive the formulation of a new model for predicting continuous single-cell gene expression, incorporating effects of both diffusive and juxtacrine (contact-based) signalling. It is shown that cells expressing low Nanog levels are less responsive to signalling cues from their environment.

Following this, cell fate selection upon differentiation is modelled through juxtacrine signalling. A novel paradigm for modelling node symmetries of an undirected graph is proposed, which is applied to hPSCs by modelling single cells as nodes, assigning edges between cells that are physically contacting in culture. Importantly, nodes possess internal dynamics. To model cell differentiation, internal dynamics are given by a novel dynamical model for the concentration levels of the differentiating morphogen at each cell. Symmetric cells have symmetrical functions for their internal dynamics, thereby sharing identical equilibria. Model equilibria qualitatively recapitulate experimental fate patterning results, suggesting that geometric symmetries of hPSC cultures may orchestrate fate selection.

Next, a local density analysis is applied to time-lapse images of wildtype and variant hPSC cultures. This helps to uncover that decreasing wildtype cell membrane resilience, thus lowering survival rates at high densities, is a crucial strategy in mechanical cell competition. Lastly, a new ensemble model is developed as an automated tool for quickly and accurately identifying variant cells in culture. Importantly, the model does not rely on cell tagging, meaning that it may be applied to wildtype hPSC cultures without knowing *a priori* if variant cells are present.

Contents

Nomenclature

Abbreviations

AB	Antibody
BMP4	Bone morphogenetic protein 4
BSA	Bovine serum albumin
CO_2	Carbon dioxide
Cy3/5	Cyanine3/5
DAPI	4’,6-diamidino-2-phenylindole
DNA	Deoxyribonucleic acid
E8	Essential 8 culture media
FCS	Fetal calf serum
FITC	Fluorescein
GFP	Green fluorescent protein
GTX	Geltrex
GRN	Gene regulatory network
HCl	Hydrochloric acid
hESC	Human embryonic stem cell
hPSC	Human pluripotent stem cell
iPSC	Induced pluripotent stem cell
LDA	Linear discriminant analysis
LR	Logistic regression
MI	Moran’s Index
mESC	Mouse embryonic stem cell
MPE	Mean percentage error
MSD	Mean squared displacement
(N)RMSE	(Normalised) root mean squared error
Oct4	Octamer-binding transcription factor 4

PFA	Paraformaldehyde
PBS	Phosphate-buffered saline
(q)PCR	(quantitative) Polymerase chain reaction
RNA	Ribonucleic acid
RFP	Red fluorescent protein
rpm	Revolutions per minute
Rock	Rho-associated coiled-coil kinase
SE8	Stable Essential 8 culture media
Sox2/10/17	Sex-determining region Y-box 2/10/17
SSEA3	Stage-specific embryonic antigen 3
SVM	Support vector machine
T	Brachyury
UV	Ultraviolet
VTN	Vitronectin

Operators

$\|x\|$	Absolute value of x
$\in$	Belongs to set
r	Correlation coefficient
$DT(A)$	Delaunay triangulation of set of points A
$det(A)$	Determinant of matrix A
$d()$	Euclidean distance
$\|\|x\|\|$	Euclidean norm of x
$dx/dt, \dot{x}$	First-order derivative of x
$\partial u/\partial x$	First-order partial derivative of u with repsect to x
x^*	Equilibrium of x
$\backsim$	Equivalence in set theory
Δ	Finite difference
∇	Gradient
$\cap$	Intersection in set theory
$\bar{x}$	Mean of x
$\ln$	Natural logarithm
$s()$	Step distance (length of shortest path) between nodes
$\subset, \subseteq$	Subset of set (proper and improper)

$\sum$	Summation
$tr(A)$	Trace of matrix A
$\cup$	Union in set theory

Notations

bmp	BMP4
$\mathbb{Z}$	Domain of integers
$\mathbb{N}$	Domain of natural numbers
$\mathbb{R}$	Domain of real numbers
$\forall$	For all
nog	Noggin
$\hat{y}$	Predicted value of y
J	Jacobian matrix
ξ	Stochastic Gaussian noise with zero mean
$\exists$	There exists
$****$	A between-distribution p-value < 0.001 from a statistical test

Chapter 1

Introduction

1.1 Background

1.1.1 Human Stem Cells and Pluripotency

Stem cells are cells in the body that can self-renew indefinitely, and have more than one option for which specialised cell type to become [1]. The former of these two properties means that stem cells can continuously generate more of themselves, both within the body and when kept in laboratory culture. To become a specialised cell type, stem cells must undergo the process of differentiation, whereby each cell commits to a specific fate. There are different types of stem cell, possessing different differentiation capabilities. Multipotent stem cells can be found in the human brain, heart, liver, bone marrow, blood, and skin [2, 3, 4]. However, the specialised fates of multipotent cells are limited depending on the cells' location and function. Conversely, pluripotent stem cells are able to differentiate into all cell types in the body [5]. As human pluripotent stem cells (hPSCs) have indefinite self-renewal capabilities *in vitro*, they retain pluripotency in culture [6]. Many researchers are therefore interested in hPSCs for their potential range and impact of applications if they are properly understood.

Importantly, there are two types of hPSCs. Those derived from a human embryo at the blastocyst stage are known as human embryonic stem cells (hESCs), which were first isolated *in vitro* in 1998 [7]. Human somatic cells that have been artificially reprogrammed to a state of pluripotency are known as induced pluripotent stem cells (iPSCs). The first published discovery of iPSCs was in 2006 by Yamanaka and Takahashi [8]. However, in parallel with this research, Yu *et al.* also derived iPSCs from somatic cells using slightly different methodology [9]. Both hESCs and iPSCs have advantages and disadvantages. An advantage of hESCs is that they provide a more accurate model for the events of embryogenesis, whereas iPSCs can possess variable differentiation potential [10, 11, 12]. On the other hand, the ethics of hESCs is still questioned in some countries, whereas iPSCs are widely accepted for medical use [11].

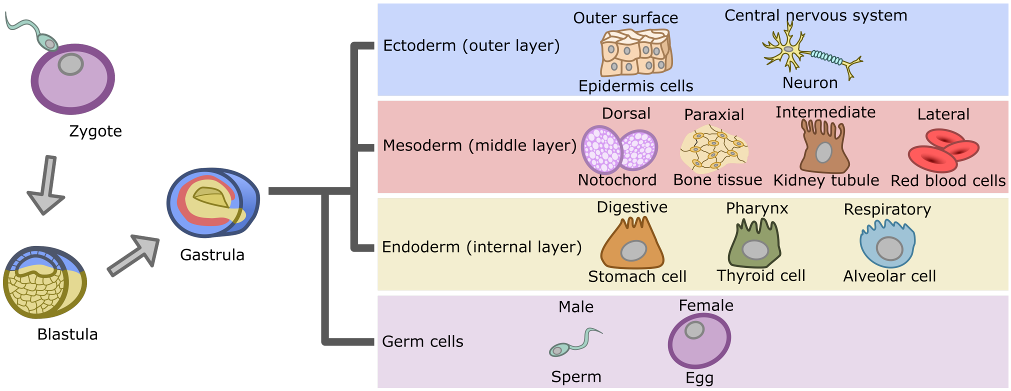
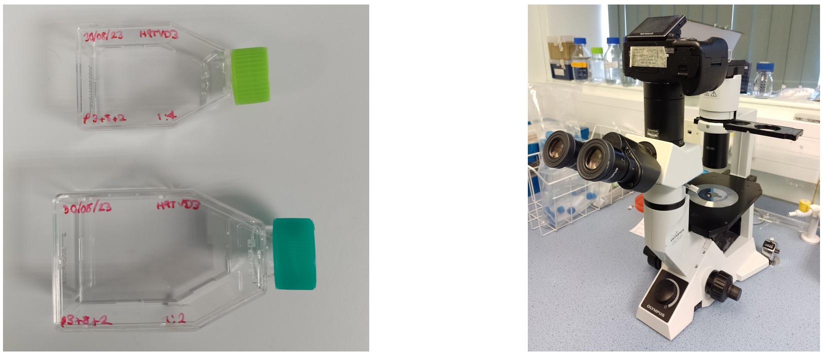


Figure 1.1: Stem cell lineage tree.

1.1.2 Application of Stem Cells to Regenerative Medicine

Regenerative medicine is a field that uses human cells to treat disease [13]. Due to their unique properties, hPSCs hold great promise for the future of regenerative medicine. Specifically, hPSCs derived from a patient could treat diseases which are otherwise regarded as untreatable by other methods [14]. For example, hPSCs taken from a patient suffering from Parkinson's disease could be differentiated into dopaminergic neurons, which are lost as a result of the illness [15, 16, 17]. Type I diabetes may even be able to be reversed by differentiating hPSCs into cells that produce insulin, then transplanting these cells into diabetic patients [18, 19]. Recovery from spinal cord injuries has been shown to be improved in mice by implanting a polymer scaffold seeded with neural stem cells into the cord, which researchers are working to apply to humans [20].

Importantly, therapeutic methods that utilise hPSCs derived from a patient provide fully tailored therapies. This avoids problems associated with donor transplants, such as limited availability and the risk of tissue rejection [21, 22]. For regenerative therapies, iPSCs are generally preferred, as these can be derived from patients' own somatic cells. Although hESCs cannot be taken from a patient, it is important to consider that research into the safe use of iPSCs is very much in development. In particular, research into how the naïve state of pluripotency is regulated has been known to yield previously unexpected results [23], with more of such research needed. Moreover, as of 2023, protocols for iPSC induction still need to be optimised to preserve long-term genomic integrity of cells [24]. However, hESCs already naturally possess the full range of differentiation potentials. Therefore, hESCs offer a more complete understanding of pluripotency than iPSCs, providing an essential framework for effective modern medicine offering personalised regenerative therapies.



(a) T12.5 and T25 flasks.

(b) Olympus CKX41.

Figure 1.2: Images of equipment commonly used for hPSC culture: (a) flasks used for culture growth and maintenance, and (b) a standard microscope.

1.1.3 Differentiation to Specialised Cell Types

As hPSCs differentiate, they lose pluripotency and become committed to a specialised lineage. One of the earliest lineage decisions that hPSCs make is differentiation into one of the three primary germ layers: ectoderm, mesoderm or endoderm. These germ layers are the basis of all somatic cell types in the body [25]. Understanding this crucial choice made by hPSCs attracts a great deal of research, as it is still not fully clear how this happens at the single-cell level. A schematic lineage tree showing specialised cell types yielded by each of the primary germ layers is given in Figure 1.1.

1.2 Fundamentals of Stem Cell Research

1.2.1 Maintenance of Stem Cells in Culture

Maintaining stocks of cell lines through laboratory culture is an essential aspect of stem cell research. Culture flasks are used to grow hPSCs; generally, these are T12.5 flasks, as shown in Figure 1.2(a). As cells are cultured to adhere only to a single 2D surface, T12.5 flasks are named as such because the internal area of the surface designed for cell adherence is 12.5cm^2. Flasks need to be coated with a substrate to allow cells to adhere. Traditionally, human stem cells have been cultured using substrate layers made from mouse embryonic fibroblasts. However, modern culture methods have removed the need for murine cells, reducing the risk of hPSC exposure to animal pathogens [26]. Two major substrates now used for hPSCs are Vitronectin (VTN) and Geltrex (GTX). A standard microscope used for culture is shown in Figure 1.2(b).

When cells have successfully attached to the flask, culture media must be supplied regularly to provide hPSCs with necessary growth and survival factors. Popular media supplements are Essential 8 (E8) and mTeSR; when using either of these, old media must be aspirated and new media supplied to the cells every 24h. More recently, Stable E8 (SE8) was formulated, which must be resupplied every 48h. Flasks of hSPCs cultured with growth media are stored in incubators kept at 37°C with 5% CO_2 in the air, as detailed in Appendix A.1.1.

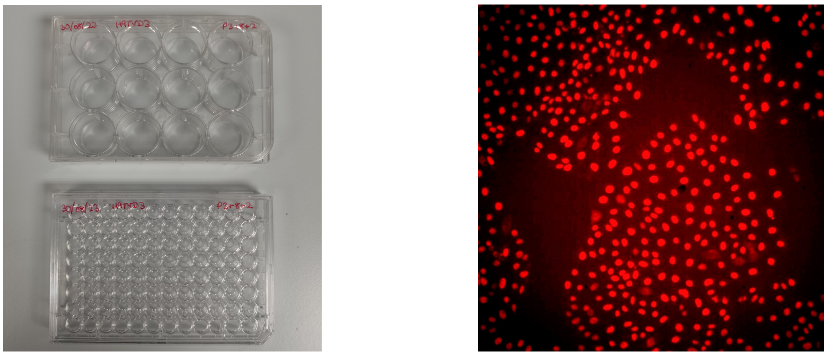
As cells grow, it is a necessary part of culture maintenance to re-plate cells onto new flasks, which must be done approximately 2-3 times per week for wildtype hESC lines. This process is known as 'passaging'. For this, an enzyme is administered to the flask which gently detaches cell colonies. Using a wide stripette so as not to break up colonies, cells are transferred to two or more newly prepared flasks at the desired splitting ratio. For example, if cells are highly confluent, a splitting ratio of 1:4 may be used, where one flask of cells is distributed between four new flasks. As cell survival is heavily reliant on neighbours, it is important to retain colonies during passaging, and not choose a splitting ratio that yields very sparse new cultures. The passaging protocol used throughout this thesis is given in Appendix A.1.2.

1.2.2 Stem Cell Surface Markers

Stem cell surface markers are specialised receptor proteins, called antigens, that can selectively adhere to signalling molecules [27]. To our advantage, fluorescent molecules called fluorophores can be attached to signalling molecules, making cell nuclei or membranes highly visible under specific microscopes. Specifically, fluorophores emit light when excited by short-wavelength light such as UV or lasers. This 'cell staining' method works by using antibodies to detect the epitope of the corresponding antigen on the cell surface. This is particularly useful as the result shows both the localisation and relative expression levels of the antigen being detected, allowing single-cell comparisons across the culture [28].

For the work conducted within this thesis, expression levels of the following genetic markers are considered:

- Nanog (pluripotency marker),
- Octamer-binding transcription factor 4 (Oct4) (pluripotency marker),
- Sex-determining region Y-box 2 (Sox2) (pluripotency & endoderm marker),
- Sex-determining region Y-box 17 (Sox17) (endoderm maker),
- Brachyury (mesoderm marker),
- CDX2 (mesoderm marker),
- Sex-determining region Y-box 10 (Sox10) (neuro-ectoderm marker).



(a) 12-well and 96-well Costar® plates.

(b) Wildtype hESC nuclei stained with Cy5.

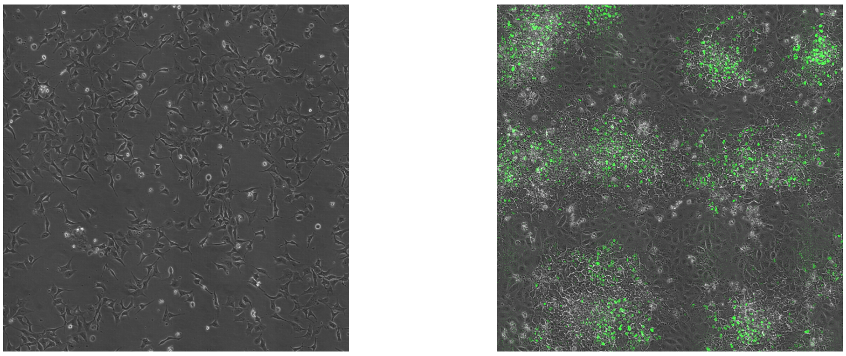
Figure 1.3: Images associated with the identification of cell surface markers: (a) plates commonly used for research consisting of different culture/experimental conditions per well, and (b) a wildtype hESC culture that has been stained with Cy5 fluorophore to identify nuclear expression levels of Nanog.

Notably, the above are all transcription factors, and as such are produced within cell nuclei.

To correctly fit within imaging devices, cells are grown in a plate prior to cell staining, such as the 12-well or 96-well plates shown in Figure 1.3(a). Prior to cell staining, cells must be fixed to immobilise target antigens and preserve cell morphology [29] (protocol given in Appendix A.2.1). Then, after washing the cell culture, a primary antibody is used to bind to the antigen being investigated. Also, to prevent non-specific anti-body binding, a 'blocking' step is conducted by applying a particular solution to the cell culture. Cells are incubated in the dark for 1h at room temperature with the primary antigen applied. Then, the primary antigen is aspirated and a secondary antibody is administered to bind to the primary; cells at this stage are kept at 4°C in the dark. For this step, it is often useful to gently rock the cell culture to facilitate homogeneous application of the secondary antibody. Detailed protocols for cell staining procedures used in this thesis are given in Appendix A.2.2.

Importantly, multiple stains can be made upon a single cell culture. However, limitations arise for two reasons: (1) antibodies derived from different animal species may cross-react, leading to a need for optimisation of staining protocols, and (2) 'bleed-through' may occur, whereby fluorescence bandwidths of two or more fluorophores overlap. For cell nucleus identification and/or segmentation, the nuclear stain Hoechst (fluorophore name DAPI) is most commonly used. For identifying other markers, commonly used fluorophores include Cy2, Cy5, FITC, and green fluorescent protein (GFP). The latter two are often used interchangeably.

Administration of one or more fluorophores facilitates immunofluorescence imaging. For imaging, an InCell Analyzer was used, which can emit light of the desired wave-



(a) Brightfield. **(b)** GFP reporter.

Figure 1.4: Images of H9TVD3 hESCs, generated as part of a time-lapse microscopy experiment using a BioStation. GFP shows nuclear Brachyury expression.

length to visualise the fluorophores used for staining. An example of a hESC culture stained with Cy5 is given in Figure 1.3(b).

1.2.3 Time-Lapse Microscopy

Time-lapse images are most commonly used to generate a video of live cell cultures over a specific time frame. As for immunofluorescence microscopy, cells are generally grown in a plate. There are two major reasons why cells cannot be continuously videoed, but rather must be time-lapse imaged at intervals: (1) lasers used for imaging are photo-toxic to live hPSCs, and (2) limited equipment availability imposes constraints on individual researchers within the same laboratory. The combination of these two factors forces important decisions to be made regarding imaging frequency and exposure intensity depending on the requirements of the experiment.

The BioStation is a piece of machinery used for time-lapse imaging. Brightfield and immunofluorescence images can be obtained from a BioStation, as shown in Figure 1.4. As cells are live, traditional immuno-histochemistry cannot be used. Instead, a hPSC nucleus may be 'tagged' with a fluorescent dye, which has expression directly proportionally to that of a particular genetic marker. Cell lines that have been tagged in this way are referred to as reporter lines. PhaseContrast is another imaging tool for time-lapse microscopy, which differs to the BioStation by showing high contrast variations across cell surfaces. This makes computational object segmentation highly challenging. An example of a brightfield PhaseContrast image is given in Figure 1.5.

All time-lapse microscopy data independently generated for this thesis was obtained using a BioStation; protocols for this are provided in Appendix A.3. There are multiple parts of this thesis that deal with time-lapse data generated by other researchers, for which the PhaseContrast imaging tool was used.

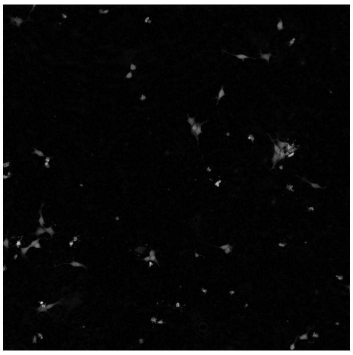


Figure 1.5: Brightfield microscopy image of genetically variant hPSCs obtained using a PhaseContrast microscope. Image by Dr. C.J. Price, University of Sheffield.

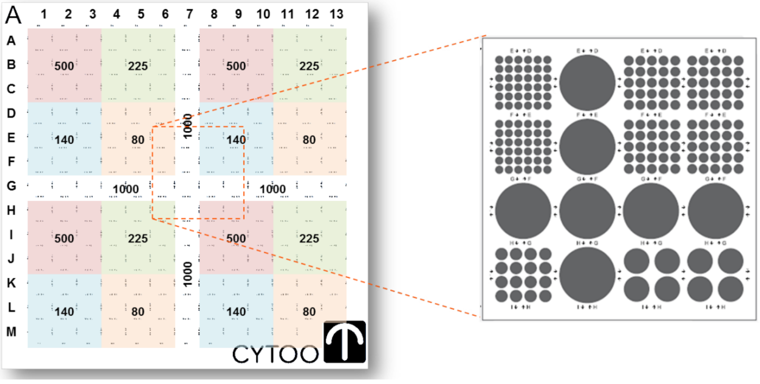


Figure 1.6: Layout of the CYTOOchip™Arena micro-pattern, provided by CYTOO.

1.2.4 CYTOOchips for Micro-Patterned Cultures

In traditional cell culture, stem cell populations grown in flasks or plates are spatially unconstrained, other than by the overall shape of the vessel they are grown in. As such, cell colonies are extremely likely to form in irregular shapes and sizes.

An increasingly popular method in stem cell research is constraining cell cultures to a particular geometric domain. Geometric constraints upon hPSC cultures have been shown to yield repeatable patterns in cell fate upon differentiation. For this work and in highly influential published work [30, 31], the CYTOOchip™Arena is used to grow hPSCs on disc micro-patterns. CYTOOchip™Arena micro-patterns are provided as a thin glass insert for cell culture vessels. The micro-pattern is square with sides of length 19.5mm; therefore, 6-well plates are the only type of culture plate that can fit the micro-pattern into one of its wells. Discs features on the chip are of 80, 140, 225, 500, and 1000μm diameter. The layout of an individual CYTOOchip™Arena micro-pattern is shown in Figure 1.6.

1.3 Motivation and Objectives

1.3.1 Current Research Limitations within the Field

Before hPSCs can be commonly used in medicinal practice, the overarching question that needs to be addressed is how exactly the cells acquire lineage - i.e. become specialised - and thus make fate decisions. This is challenging to answer, primarily because colonies of hPSCs display heterogeneity *in vitro* which is not fully accounted for [32]. This means that, for cells within the same culture, the state of one cell at any given time may be different to the state of another cell. A cell's state is quantifiable by its gene expression levels. As a cell's state influences its lineage trajectory, single-cell fates become incredibly difficult to predict due to this heterogeneous patterning [33, 34]. There are protocols for pushing hPSCs towards a specific cell type, but these are limited, expensive, and may yield variable results [35, 36, 37].

An additional challenge in stem cell research is quality control. It is well known that culturing hPSCs for a prolonged time leads to a substantial risk of some cells acquiring unwanted genetic mutations [38]. These occur at the chromosomal level, and as such are challenging to detect without disruptive and costly methods for genetically screening the entire population [39, 40, 41, 42, 43]. In many cases, particular mutations lead to 'super-competition' *in vitro*, whereby wildtype (normal) hPSCs are completely out-competed by variants. Both of these cases may be greatly dangerous and/or costly, threatening hPSC research quality. The mechanisms of stem cell super-competition are poorly understood at present, but such understanding is needed to ensure the safety of cell therapies.

1.3.2 'State of the Art' for Mathematical and Computational Modelling in Stem Cell Biology

Mathematical and computational models of biological processes can help to optimise the efficiency of laboratory protocols, and predict outcomes prior to conducting experiments. This can reduce the high costs of laboratory work, providing indispensable savings to medical research facilities.

Several studies have been published that explicitly focus on mathematically modelling hPSC fate. Nearly all of these models were inspired by a groundbreaking study in 2014, which revealed that by confining the adherence of cells to a disc shaped micro-pattern, hPSC fate no longer appears heterogeneous after inducing differentiation [31]. In 2018, this resulted in the development of a reaction-diffusion model which accurately mimicked the evolution of two coupled chemical concentrations, one of which initiates hPSC differentiation, whilst the other directly inhibits this chemical [30]. However, this model, as with all traditional reaction-diffusion models, is highly sensitive to its initial conditions. Contrarily, the process of embryonic formation is extremely robust, as it is able to overcome morphological changes and perturbations to chemical inputs [44].

A study on mouse ESCs in 2018 suggested that focusing on cell-to-cell interactions, which can be controlled through culture geometry, is key to capturing the robustness of embryonic formation [45]. However, whilst the local cellular neighbourhood was analysed, no model describing cell-to-cell interactions was offered. In more recent years, the problem of robustness for hPSC models is addressed; much of the literature is now tending towards models that describe how cells interact with neighbours to orchestrate fate patterning.

In 2019, an extended Cellular Potts model was developed which reproduced hPSC fates seen on disc micro-patterns [46]. Additional micro-patterned geometries were explored in 2020 [47], with resultant hPSC fates explained through the mechanical tension that cells were subject to at different micro-pattern locations. A reaction-contraction-diffusion model proposed in 2021 hence incorporated mechanical tension into a traditional reaction-diffusion system [48]; however, simulations were only given for cell cultures on disc domains, so it is unclear whether the model is applicable to other geometries. Moreover, the model does not account for edge-sensing mechanisms present for hPSC cultures [49]. In early 2023, a highly interesting and multi-layered mathematical model for hPSC fate patterning was published [50]. An agent-based model is given, where agents are not explicitly cells, but rather one of several genes that are used as fate markers. Thus, the model uses a feedback system to trigger gene regulatory networks (GRNs) (activatory/inhibitory relationships are well known) and reaction-diffusion. Notably, the GRNs used are very basic; dedicated GRN models from 2020 [51] and 2021 [52] thoroughly explore dynamical system properties, such as bifurcation points and switching behaviour. Nevertheless, the multi-layer model produces patterns of gene expression which correspond well to experimental cell fates [50]. Again however, edge-sensing is not accounted for, and the model is only simulated for cells on a disc. There is yet to be proposed a predictive model for hPSC fate patterning that demonstrates generalisability to the various micro-patterned geometries that have been explored experimentally.

Increasingly more research is being conducted into the mechanisms of cell competition. This includes the development of machine learning algorithms for classifying variant cells in co-culture with wildtypes using RNA-seq data [53]. However, there are currently no automated tools for identifying genetically variant hPSCs from images. Outside of stem cell biology, the success of variant MDCK cells has been linked to their increased motility, allowing them to force wildtypes into areas of high density, resulting in apoptosis (death) of wildtype cells [54]. This information is very useful, but no such analyses have yet been applied to stem cells. Moreover, this feature identification relies on time-lapse data run over multiple days. Furthering this research, in 2021, a deep learning model with a variational auto-encoder output local density as the single most important factor for predicting MDCK survival in co-culture with competing cells [55]. Whilst this computational tool is of high interest, a custom microscope was built for the research, so it may not feasible to implement the model using data obtained across different imaging platforms. Additionally, a cell line tagged with fluorescent dye was developed, allowing for easy recognition of the distinct cell populations. Detecting genetic mosaicism from brightfield images is a much more challenging task, particularly as the exact mutation may not be known *a priori*. An algorithm that classifies genetically variant hPSCs from a minimal sample of images, obtained using standard machinery, would thus be a widely beneficial tool.

1.3.3 Thesis Objectives and Overview

Observed heterogeneous gene expression in (undifferentiated) pluripotent cell populations is not fully understood. In particular, single-cell heterogeneous expression of pluripotency markers has not been investigated for micro-patterned cell cultures, even though substantial focus has been applied to cell fate after differentiation. Therefore, the objective of Chapter 2 of this thesis is to develop a novel spatial analysis and spatial models for single-cell Nanog expression, using data from hESCs grown on disc micropatterns as inputs. This is achieved using several key model features relating to positional information and signalling across the cellular neighbourhood.

A major gap in understanding fate patterning of hPSCs that have undergone differentiation is a mathematical model that is (1) robust, (2) biologically informed, and (3) shown to recapitulate numerous fate patterns found in the literature [31, 45, 47]. One of the objectives of this work is to provide such a model, which is proposed in Chapter 3. For this, a novel theoretical paradigm is proposed for modelling node symmetries on a network, and a novel dynamical system for concentration levels of the differentiating morphogen at each node is given, where cell state equilibria preserve node symmetries. Time-invariant system equilibria are assumed to correspond to cell fate acquisition. The long term goal is to better understand hPSC signalling, which may inform optimal design of micro-patterned geometries to yield desired cell fates upon differentiation. If achieved, the associated costs of manufacture would be significantly lower than that of current differentiation protocols.

Chapter 4 of this thesis aims to add to knowledge about the mechanical strategies used in hPSC super-competition. Specifically, a local density quantification is applied to timelapse images of variant and wildtype hPSCs in co-culture, which is compared to their respective homeostatic culture densities. This novel application of local density analysis facilitated advancements in understanding how super-fit variant hPSCs out-compete their wildtype counterparts through an increased resilience to the mechanical stress of highly compact culture.

There currently lacks an automated tool for detecting genetically variant hPSCs at the single-cell level, using images of untagged cells as input. This kind of tool has a range of applications in stem cell super-competition, and also tissue therapies. Chapter 5 of this thesis aims to resolve this gap by proposing an ensemble classifier that accurately labels hPSCs as wildtype or variant from time-lapse images, taken over a short time frame of 3h. This methodology may be integrated into a laboratory workflow by using incubators equipped with a microscope for cell culture, which are already available for use. Live cell cultures should be imaged at regular intervals, particularly if the passage number of a flask is high and therefore more likely to produce genetic mutations. This time-lapse image data then serves as input for the model workflow. Cells labelled as variant by the classifier may be manually removed from culture to prevent proliferation. The long-term goal of this work is to improve the quality of stem cell research and therapies, by reducing the risk of genetic mosaicism in cell cultures used in medicine.

Chapter 6 concludes the work undertaken in this thesis, including a discussion of the results presented. Based on the implications and limitations of this work, suggestions for future directions of research are offered.

Chapter 2

Spatial Analysis and Modelling of Nanog Expression in Human Embryonic Stem Cells

2.1 Introduction

How neighbourhood cues orchestrate the onset of differentiation in hESCs remains unknown. In particular, Nanog - a key genetic regulator of pluripotency - exhibits heterogeneous expression in hESC cultures, resulting in poor understanding of how Nanog is down-regulated. In this chapter, Nanog expression levels of wildtype hESCs cultured on disc micro-patterns are characterised using a novel neighbourhood analysis, showing that a cell's Nanog intensity is inversely related to its neighbour count. These preliminary results form the basis for a novel empirical model that formulates the level of static single-cell Nanog intensity as a function of diffusive and juxtacrine signalling. This model, which predicts continuous Nanog signals, yields a reasonable distribution output despite low correlation between actual and predicted Nanog intensities. Interestingly, this model performs over twice as well for high-Nanog cells compared to low-Nanog cells, with low-Nanog cells also displaying a lesser relationship between Nanog intensity and neighbour count. These results suggest that as cells enter lineage specification, they become less susceptible to signalling cues from their environment.

Structure of this Chapter

The remainder of this chapter will be structured as follows. Section 2.2 provides background and a literature review within the field. Section 2.3 describes the micro-patterning experiments and data generated from these experiments, which are used in modelling. In Section 2.4, single-cell Nanog intensities are analysed with respect to the local cellular neighbourhood. Positional information of the cells is also considered, and it is demonstrated that the neighbourhood results are not confounded by global density patterns for each image. Section 2.5 develops data-informed regression models of single-cell Nanog expression levels in micro-patterned cell cultures, which incorporate the effects of neighbourhood features. Section 2.6 introduces a novel model, combining the effects of both diffusive and juxtacrine signalling, which is used to predict cellular Nanog expression levels. Finally, Section 2.7 provides a discussion of the results and conclusions.

2.2 Background and Literature Review

Stem cells within the human embryo are kept in a pluripotent state by a select number of genetic regulators. It is understood that the transcription factors (TFs) Sox2, Oct4, and Nanog (SON) comprise the essential network for regulating and maintaining pluripotency [56, 57, 58]. Of the target binding sites occupied by Oct4, 69.5% are also occupied by Nanog [59]. Similarly, Sox2 shares 66.1% of its binding sites with Nanog [59], suggesting that a large proportion of SON's target genes are regulated by SON in conjunction. Nanog is a downstream target of Sox2 and Oct4, and is the only TF expressed solely in pluripotent cells [60]. Therefore, down-regulation of Nanog in a cell indicates the onset of lineage specification.

By fusing ESCs with non-pluripotent stem cells, Nanog is sufficient to activate pluripotent genes in non-pluripotent stem cells [61]. Specifically, neural stem cells were fused with ESCs to form hybrid cells. These hybrids adopted a state of pluripotency as a result of Nanog over-expressed in the ESCs, which was found to be a uni-directional process. Interestingly, the activation of pluripotent genes by Nanog was able to reverse previous genetic and epigenetic programming of neural stem cells in hybrids.

Dynamic Genotypic Substates Impact Lineage Decisions

Importantly, not all stem cells are equally susceptible to different lineages when pluripotent. The notion of naïve and primed pluripotent states was first proposed in 2009, such that primed hPSCs have a differentiation propensity towards specific lineages, whilst all possible cell fates are equally likely for naïve hPCSs [62]. Embryonic stem cells may be either naïve or primed; derivation of hECSs from the blastocyst or epiblast yields naïve or primed cells, respectively [62].

Moreover, various lineages may be 'tried out' by hESCs prior to full commitment. A 2016 study provided evidence for this by showing that naïve or primed pluripotent states in hESCs are dynamic [63]. This means that the transition of hESCs from a naïve to primed state is not one-way, but rather a process in which cells can inter-convert between states. Naïve hECSs are less responsive to signalling pathways and thus less susceptible to signals that could interfere with their state of lineage-neutrality. As a result of this, the time between the signalling of morphogens that induce differentiation and hESCs acquiring fate specifications (priming) may be delayed. The loss of naïve pluripotency is therefore considered a gradual process. The authors conclude that lineage specification for hESCs relies on mechanisms of pre-patterning depending on the spatial organisation of the cells, although they do not state what they believe the impact of spatial organisation may be.

Similarly to these findings, NTERA2 hPSCs were found to inter-convert between positive/negative expression of the SSEA3 protein [64], which is a marker for undifferentiated hPSCs [65]. The authors developed a variational Bayesian expectation maximisation algorithm for hidden Markov trees, where each node in the Markov tree represented an SSEA3 expression state. This model was accurately fitted to time-lapse gene expression data to predict SSEA3 expression at different time points. This use of modelling alleviates the future need for a fluorescent reporter cell line which can be measured in real-time. Although the authors did not capture the probability of a cell moving between SSEA3 positive/negative states, the success of the Markov model suggests that such transitions are memoryless, i.e. dependent only on the previous state [66]. However, there is no evidence that this applies to genes other than SSEA3. In particular, no correlation was found between Nanog and SSEA1 (a pluripotency marker) [67], suggesting that multiple heterogeneous pluripotent states may exist simultaneously with potentially varied behaviour.

An important question arising from these discoveries is whether these primed cellular substates are able to be mapped to specific lineage decisions. This question was addressed in [33], firstly investigating whether these substates correspond to functional and self-renewing hESCs, then testing if these cells have a differentiation bias according to their substate. The authors designed a reporter cell line for GATA6, which is a marker for endodermal cell fate, by knocking a GFP reporter into the GATA6 locus. This means that the cells did not need to be fixed and stained in order to quantify GFP levels, corresponding to GATA6 expression, but rather GFP levels were able to be observed dynamically through time-lapse imaging of live cells. This method is highly useful for capturing the emergence of patterns in gene expression. Indeed, the cells were found to be capable of transitioning between states of expressing and not expressing GATA6, where cells in the GATA6 expressing substate had a higher differentiation propensity for endoderm and also mesoderm. This demonstrates that while hESCs remain in an undifferentiated state, they become primed for specific fates and are able to convert between these states freely. Lineage decisions may be inferred from these primed states, supporting the consensus that quantification of a cell's gene expression levels is an accurate indicator of cell fate. Notably, GATA6 is able to be directly repressed by Nanog, and low Nanog expression is able to be inferred from high GATA6 expression [67].

Dynamic Nanog Substates Regulate Pluripotency and Lineage Priming

Several studies have identified fluctuations in Nanog expression among individual ESCs as essential for regulating pluripotency and lineage priming. Considering mouse embryonic stem cells (mESCs), it was found in 2009 that cells' expression of Nanog falls into two distinct populations: low Nanog (LN) and high Nanog (HN) [68]. These populations were distinguished by two distinct peaks on the histogram displaying the number of Nanog molecules per cell among a Nanog-GFP mESC reporter line. Typically, only 5-25% of cells were in the LN state. Interestingly, the authors showed that taking a cell sample from either the LN or HN population resulted in the same qualitative Nanog distribution being generated over time. Hence, this behaviour is robust, and suggests to be intrinsic to the development of mESC populations.

The authors proposed that pluripotency may be modelled as a noise-driven excitable system. It is known that some level of stochastic transcriptional noise is key for producing genotypic differences in cell populations as a means of selection [69, 70], providing the rationale for the model. As Nanog and Oct4 are known to activate themselves and each other [71], with high levels of Oct4 leading to Nanog repression [59, 72], the authors propose the following ordinary differential equation (ODE) system:

$$\frac{dN}{dt} = \alpha_n + \frac{\beta_n N^n}{k_n^n + N^n} - \delta \frac{A^p}{k_x^p + A^p} N - \gamma_n N + \xi(t), \tag{2.1}$$

$$\frac{dA}{dt} = \alpha_a + \beta_a A N - \gamma_a A, \tag{2.2}$$

where N and A represent Nanog and Oct4 respectively, and the α values represent basal expression. Nanog self-activation is modelled by a Hill function with amplitude β_n, half-maximal activation constant k_n, and Hill coefficient n. The authors assume that self-activation of Oct4 is linear, with β_a as the strength of modulation by Nanog. Repression of Nanog by Oct4 is modelled as a Hill function with strength δ, half-maximal activation constant k_x, and Hill coefficient p. Both TFs have assumed linear degradation, with rates represented by the γ values. The term $\xi(t)$ represents Gaussian noise with 0 mean and correlation [68].

By analysing the system's equilibria, the authors determined that the LN state was unstable, i.e. sensitive to small perturbations about its equilibrium. This instability was backed up through experimental findings, whereby cells in the LN state were much more likely to differentiate than those in the HN state. An additional stochastic model provided further evidence that Nanog fluctuations were a result of transcriptional noise which regulates the LN and HN dynamics. Taken together, these results may indicate that fluctuations in Nanog expression are essential for heterogeneous lineage priming of pluripotent stem cells.

As an advancement of this work, the impact of LN and HN states on mESCs' lineage markers was explored by a different group of researchers in 2014 [73]. Interestingly, it was found that Nanog fluctuations can occur over a period of hours, rather than days as was previously thought [74]. Cells in the LN state had increased expression of lineage markers, but this priming was not biased towards any specific lineage. This suggests that Nanog fluctuations may be the mechanism which allows cells to explore lineages before full commitment. Moreover, each mESC was found to spend some time with Nanog transcriptionally inactive, measured through Nanog mRNA expression. Therefore, Nanog heterogeneity in stem cell populations is regulated at the transcriptional level, which further suggests that this is an intrinsic and key phenomenon. Another study in 2013 found that around 5% of mESCs do not display Nanog mRNA transcription [75], which also supports the small LN population size reported in [68].

Cell State is Dependent on the Cellular Neighbourhood

Cells in culture do not exist in isolation; signals transferred by cell-to-cell contact (juxtacrine signalling) and diffusion play a pivotal role in cell behaviour [76, 77]. To describe how cells react to the state of surrounding cells, the ‘neighbourhood watch’ model was introduced in 2022, which proposes that positional information drives stem cell fate in the chick embryo epiblast [78]. The authors are interested in how the primitive streak (PS) - a developmental stage in embryogenesis - forms with respect to the positional information of cells expressing an activator/inhibitor of PS formation. Cellular intensity of the Brachyury gene was used to infer the results, whereby increased Brachyury intensity provides increased evidence for PS formation.

A cell pellet expressing a PS activator is able to form PS for chick stem cells *in vitro* [79]. However, the authors of [78] found that by designing two neighbouring cell pellets both expressing a PS activator, the formation of PS was no stronger than for a single pellet. Interestingly, by taking a cell pellet expressing a PS inhibitor, then symmetrically introducing four neighbouring cell pellets (two either side) expressing a PS activator, much stronger PS formation was seen than when no PS inhibitor was present. The authors thus supposed that cell fate is decided with respect to the difference in signal between a cell and its neighbours.

Mathematically, the ‘neighbourhood watch’ model predicts that a PS will form next to cell i if

$$\frac{F_i - F_{\text{nbhd}}}{F_i} > \beta \tag{2.3}$$

where F represents a cell’s value of the SMAD protein complex, consisting of SMAD1/2/3/4/5/8. The reason for modelling the SMAD complex is that the amount of the PS activator/inhibitor received is able to be inferred through the levels of SMAD in a cell [80]. Here,

$$F_i = \frac{a_V V_i}{1 + a_V V_i + a_B B_i}, \tag{2.4}$$

where V_i and B_i represent the levels of SMAD complex inducer and inhibitor, respectively. The parameters a_V, a_B are scalings of the protein concentration. The neighbourhood signal

$$F_{\text{nbhd}} = \frac{\sum_j F_j}{2n}, \tag{2.5}$$

with $j \in [i-n, i+1] \backslash \{i\}$ where $(2n+1)$ is the full width of the neighbourhood. The threshold β, along with the other parameters, were determined through Markov-Chain Monte-Carlo (MCMC) Bayesian computation.

The authors found that this model - and notably not a model thresholding only the F_i value without neighbourhood considerations - was able to reproduce experimental data. Interestingly, the distance over which cells compared their relative signal to that of surrounding cells was concluded to be farther than the direct neighbourhood. The mechanism by which cells assess their relative signal levels is not explored within this work, and as such, poses an interesting question on the matter. As for some of the authors' experiments, the cell pellets for which dependent behaviour was observed were not in direct contact with each other, there is reason to suppose that signal transduction occurred through the diffusion of morphogens across the cell culture [81]. An exploration of these signalling mechanisms may present a useful direction for future work.

Stem cells' genetic state has been successfully modelled as directly regulated by neighbours' states [82]. The model does not describe a specific system; it is intended for some mutually exclusive transcription factors. Thus, this model is compatible with the finding that cell state is inversely related to neighbour state, as different genes exhibit different activatory/inhibitory behaviour. The authors note its possible relevance for the inner cell mass (ICM) of the mouse embryo, particularly with respect to the activation and inhibition of Nanog and Notch. Importantly, the signal being modelled is binary, i.e. each cell is treated only as positively or negatively expressing a particular gene. The neighbourhood signal s_i for the ith cell in the system is modelled as

$$s_i = \left(\sum_{j \neq i} u_j q^{d_{ij}-1} \right) \Bigg/ \left(\max_k \sum_{j \neq k} q^{d_{kj}-1} \right), \quad q \in [0,1], \tag{2.6}$$

where u_j is the signal of the jth cell, d_{ij} is the length of the shortest path between the ith and jth cells, and q is a parameter representing the strength of signal dispersion across the culture. This neighbourhood signal is incorporated into a larger ODE system which predicts time-dependent differentiation patterns, with cells computationally represented as either a tiling of regular hexagons or Voronoi objects. Neighbour-only signalling was found to result in a checker-board patterning of cell state. Conversely, signalling across the culture results in larger clusters of cell fate.

In 2023, this model was contextually applied to laboratory-generated data [83]. Data reporting Nanog and GATA6 expression within the ICM of mouse blastocysts was obtained via existing experimental studies, combined with novel data gathered by the authors. Interestingly, for early stage embryos, it was found that the percentage of Nanog positive cells across the whole tissue was representative of the percentage of Nanog positive neighbours a given cell was likely to have. Again, only a binary signal was considered. The best model fit was obtained by incorporating the impact of cell-to-cell signalling across the entire tissue, suggesting that cell-to-cell contact leads to a 'trickle down' juxtacrine signal of decreasing intensity as inter-cell distance increases. In agreement with the 'neighbourhood-watch' model [78], this model predicts that the signalling environment is not limited to the cell's direct neighbours.

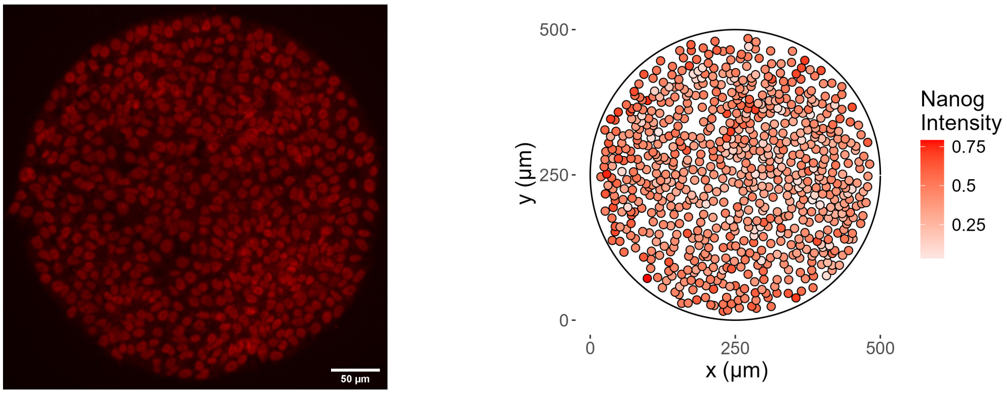
Notably, the model described by Equation 2.6 does not account for cells that have no direct neighbours, as the framework only accounts for contact-based signalling. However, as demonstrated by the 'neighbourhood-watch' model [78], cells are susceptible to signalling cues - likely from diffusion - even when not in contact. Moreover, cell pellets were used for these experiments, rather than single cells. For these reasons, determining single-cell Nanog expression for isolated cells, comparing this to cells with neighbours, and incorporating this into a combined model for juxtacrine and diffusive signalling would be a unique and promising direction of research.

To uncover any effects of positional information on gene expression, it is useful to have control over the organisation of the cellular environment. This is challenging in unconstrained cell cultures, but can be achieved using micro-patterned culture inserts that restrict cell growth to a particular geometry [84]. Edge-sensing mechanisms that have been reported in stem cell cultures may also be more easily investigated using micro-patterns to control the boundary of the culture [49].

2.3 Novel Data for Nanog Expression in hESCs on Disc Micro-Patterns

For data generation, wildtype H7 S14 hESCs were first dissociated to single cells (Appendix A.1.3), then seeded at a density of 1×10^6 cells/mm^2 on CYTOOChip Arena micro-patterns (Appendix A.1.4). Each micro-pattern was a 500μm diameter disc, 33 of which were used in total.

After 3 days from seeding, cells were fixed in place (Appendix A.2.1) and stained for Nanog and Hoechst (Appendix A.2.2). Hoechst fluorescent dye stains DNA, thereby acting as a nuclear counter-stain. This also allows for all cell nuclei to be identified. The DAPI/Cy5 channel was used for imaging Hoechst/Nanog, respectively. As Nanog is a nuclear expressed gene, cell membranes are not visible for the immunofluorescence images generated in this work. Each image was processed in ImageJ and subsequently



(a) Immunofluorescence microscopy image showing Nanog.

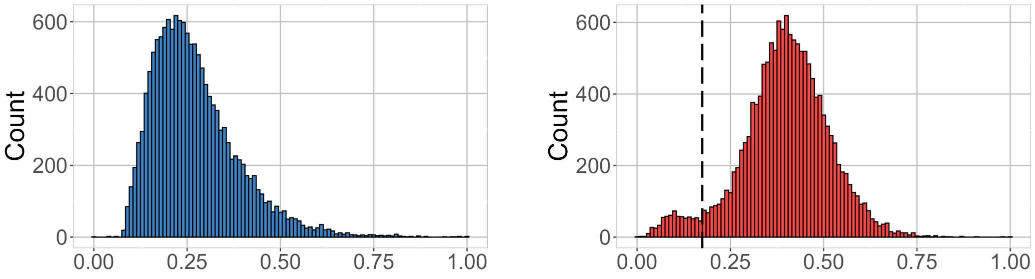
(b) Scatter plot of cells coloured by Nanog expression, quantified by pixel intensity.

Figure 2.1: Representative images of wildtype H7S14 hESC nuclei on a 500µm disc micro-pattern (CYTOOchip Arena). The same disc is displayed in both images.

CellProfiler, which performed automatic segmentation on the cell nuclei. Information for each cell was output by CellProfiler as a .csv file, containing 14765 total observations. The programming language R was implemented for all further analysis. Both Hoechst and Nanog pixel intensity values were standardised in the range $[0, 1]$. A representative immunofluorescence image showing nuclear Nanog expression, together with the corresponding scatter plot showing cell nucleus centres, are given in Figure 2.1.

Cells in 15 of the 33 micro-patterns displayed some significant overgrowth of the disc boundaries. For this reason, these 15 images were immediately removed from analysis, retaining 18 images of clearly defined disc cultures. Histograms were plotted to visualise the distribution of cells' Nanog and Hoechst intensities, shown in Figure 2.2. As reported in the literature [68], two distinct populations of Nanog intensities were observed. Cells with Nanog intensity ≤ 0.175 were classed as low Nanog (LN), and the remaining cells as high Nanog (HN). The LN population contained only 718 cells (4.85%), and the remaining 14048 cells were HN. Morphological analysis of the cell nuclei - performed automatically in CellProfiler in conjunction with object segmentation - revealed that the LN population exhibited significantly reduced nuclear area and perimeter than the HN population. This is demonstrated in Figure 2.3. A Kolmogorov-Smirnov test was used to quantify the statistical significance of both single-cell area and perimeter between LN and HN distributions [85]. While morphological nuclear differences are reported in the literature [86], this is the first time such analysis has been carried out in respect to levels of Nanog expression.

Hoechst intensity of an individual stem cell can be used to identify the cell cycle phase [87, 88], with two of the four phases known to promote pluripotency [89]. Specifically, the S and G2 phases positively regulate the pluripotent state, but only cells in the G1 phase are receptive to molecular signalling, while cells in the M phase are 'at rest'. It has



(a) Hoechst intensity distribution.

(b) Nanog intensity distribution.

Figure 2.2: Histograms showing the intensity distributions of Hoechst and Nanog for wildtype H7 S14 hESCs. Low Nanog (LN) and high Nanog (HN) populations were defined according to an intensity threshold of 0.175.

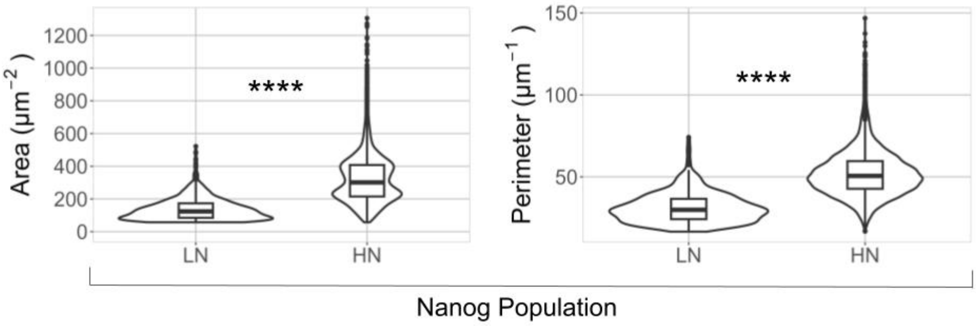


Figure 2.3: Morphological features of cell nuclei within either the LN or HN population.

been shown that the differentiation propensity of stem cells is coupled to phases of the cell cycle [90, 91, 92]. As Figure 2.2(a) does not correspond to a cell cycle profile (Appendix A.5), it is concluded that the cell cycle is unlikely to affect the results presented here. In case of observing a distinct cell cycle profile using Hoechst staining, only cells in the G1 phase - in which cells are receptive to their signalling environment - should be considered.

2.4 Spatial Analysis of Nanog Expression Levels

2.4.1 Defining the Local Cellular Neighbourhood

To quantify each cell's number of neighbours (i.e. how many surrounding cells share membrane contact), Delaunay triangulation was performed on each cell nucleus (Appendix B) [83]. This was combined with a radial distance method (RDM) [45]. The RDM computes the neighbour set N_i of cell x_i as

$$N_i = \{x_j | d(x_i, x_j) \leq T\} \tag{2.7}$$

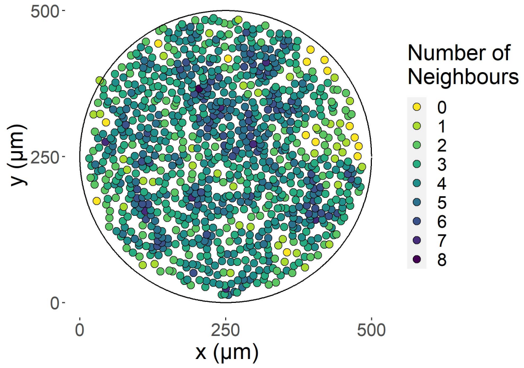


Figure 2.4: Representative scatter plot of cells coloured by number of neighbours, quantified using Delaunay triangulation with a radial distance threshold of 16.25μm.

for a distance threshold T and Euclidean distance metric $d()$.

The reason for combining these methods is that, for a cell of interest x_i and surrounding cells $x_j, x_k \in N_i$, the RDM alone cannot determine whether cell x_k is 'blocked' by cell x_j, i.e. cells x_i, x_k are not actually in contact with one another. Delaunay triangulation ensures that, in such cases, cells x_i, x_j will not share an edge. However, through Delaunay triangulation alone, a cell may be triangulated with other cell(s) outside its local neighbourhood if there are no immediate cells in a particular direction. Therefore, combining these methods is reasoned to give a novel and robust method for neighbourhood evaluation.

The average hESC diameter is 12-13μm [93], so the upper bound of 13μm was assumed in order to capture all neighbours. Then, assuming all cells are perfectly circular and have a central nucleus, cell nuclei at 13μm apart or less would be likely to be in contact with each other. However, hESC morphology in confluent cultures is rarely perfectly circular [94], so a 25% increase was assumed to account for cell ellipticity/eccentricity. By this calculation, $T = 16.25$μm. A representative spatial scatter plot of cell nuclei coloured by their resultant neighbour count is given by Figure 2.4.

To quantitatively verify that the chosen value of T was reasonable for each nucleus (node), the distance to another node forming part of the same Delaunay triangle was recorded. Considering the previously discussed possibility of non-neighbouring cells sharing an edge through Delaunay triangulation, only the distance between a node and its single closest connected node was recorded for verification of T. Indeed, the mean distance between a node and its nearest connected node was 12.46μm. This suggests that the threshold $T = 16.25$μm is reasonable for capturing all cells in contact.

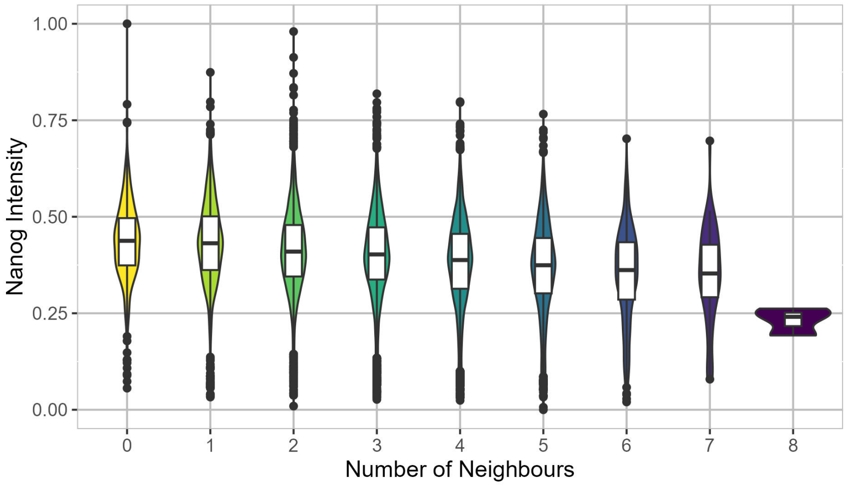


Figure 2.5: Violin plot showing single-cell Nanog intensity against the cell's number of neighbours. Results are across 18 images of disc micro-patterns.

2.4.2 Nanog Intensity is Inversely Related to Neighbour Count

Across all 18 images, each cell's Nanog intensity was investigated with respect to the cell's number of neighbours. The results are shown in Figure 2.5. Examining the median of each group, the plot shows an inverse relationship between a cell's number of neighbours and its Nanog expression. Computational results showed that this trend was present for slightly increased/decreased threshold values for the neighbourhood radius, which could account for varied cell ellipticity/eccentricity.

The behaviour of LN and HN cells with respect to number of neighbours was investigated, to understand if this differs between the two populations. On average, the number of neighbours for LN cells, 3.55, was reasonably similar to that for HN cells, 2.97. However, whilst the HN population displayed similar qualitative behaviour to that shown in Figure 2.5, the LN population displayed no clear trend between Nanog expression and neighbour count. As the LN population consisted of $< 5\%$ of the total cells, it was possible that the Nanog/neighbour relationship emerges only for suitably large cell populations. To test this, a random sample of the HN population was taken without replacement (719 cells) to match the size of the LN population (718 cells). Groups defined by neighbour count were uniformly sampled from. Interestingly, even for this small sample of the HN population, the Nanog/neighbour relationship was generally prevalent with some fluctuations; for cells with six or seven neighbours, the strictly decreasing trend in median Nanog intensity was not present. To visualise these results, Nanog expression for the LN and HN populations was plotted, as shown in Figure 2.6.

Considering the mechanisms of genetic regulation through the cellular neighbourhood, it is important to understand whether a cell's Nanog intensity is related to that of its

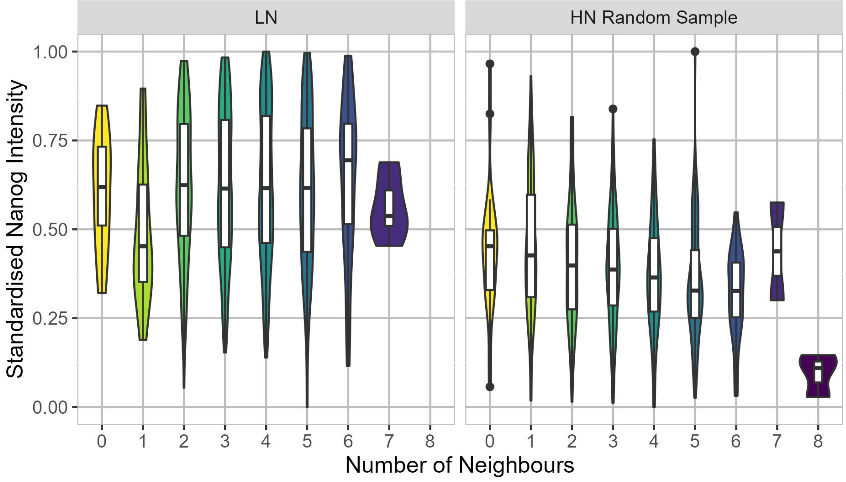


Figure 2.6: Violin plot showing single-cell Nanog intensity against the cell's number of neighbours. The HN population random sample (719 cells) was generated to approximate the size of the LN population (718 cells).

neighbours, or instead its neighbour count alone. For differentiating chick ESCs, cell fate is partially controlled by the difference between a cell's genetic signal and the average signal of surrounding cells [78]. To test whether this is the case for signalling in hESCs from direct neighbours only, the mean Nanog intensity across all neighbours of each cell was computed. Cells with no neighbours were excluded from this analysis. Nanog expression among the LN population displayed no relationship with the mean Nanog intensity of the cell's neighbours; the correlation coefficient between the two variables was -0.021. Taken with the behaviour seen in Figure 2.6, LN cells - i.e. cells that may have begun to acquire lineage - appear to behave independently of their local cellular neighbourhood. In contrast, as the mean Nanog expression of HN cells' neighbours increased, so did the cell's own Nanog expression, yielding a weak positive correlation of 0.31. These results are shown in Figure 2.7. The latter result is particularly interesting in conjunction with that of Figure 2.5; these findings suggest that hESCs express high Nanog in the presence of few neighbours, with those neighbours themselves expressing high levels of Nanog. However, as cells with no neighbours express the highest levels of Nanog on average, a further cell signalling distance - rather than just nearest neighbours - may control Nanog patterning [83].

2.4.3 Cells at Micro-Pattern Boundary Exhibit Increased Nanog Expression

An advantage of using micro-patterned cell cultures is the ease of control over the morphology of the culture boundary. As edge-sensing mechanisms have been previously

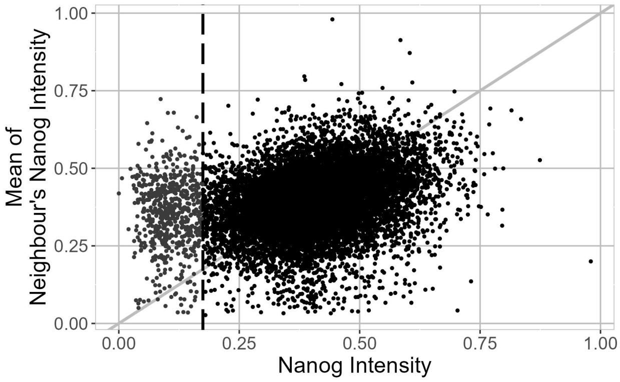
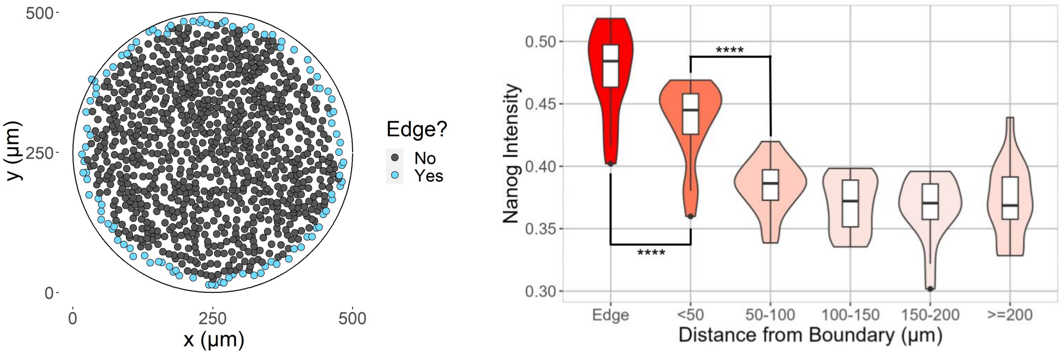


Figure 2.7: Scatter plot showing single-cell Nanog intensity against the mean Nanog intensity of the cell's neighbours. The vertical line thresholds LN and HN populations.



(a) Representative scatter plot of cells at the micro-pattern edge.

(b) Violin plot showing cells' Nanog intensity - averaged per image - vs. distance from the micro-pattern boundary.

Figure 2.8: Plots displaying the spatial location of edge cells in one image, and the variation of Nanog intensity averaged across images at the micro-pattern boundary. Edge cells were determined through computation of the concave hull, using cell nucleus centres as points.

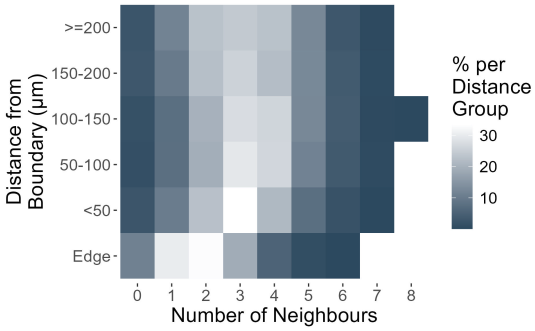


Figure 2.9: Heat map showing the percentage of cells with a given neighbour count for within each discrete group representing distance from micro-pattern boundary.

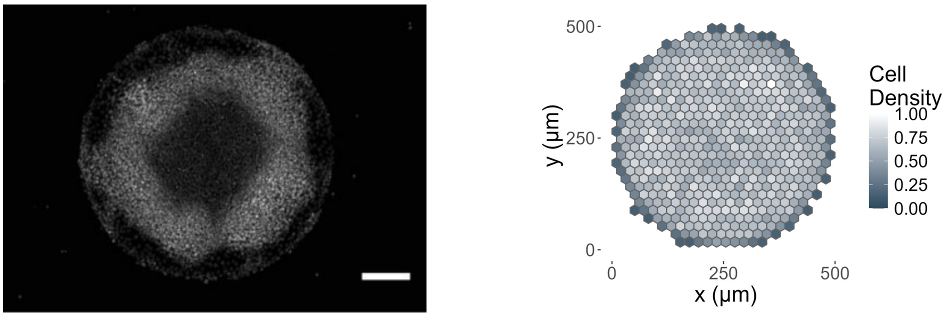
reported as a regulator of self-organisation in the human embryo [49], it was investigated whether such mechanisms appear to be present here.

The concave hull of the polygon enclosing all cell nuclei was computed for each image. From this, cells that were part of the concave hull were taken to be those at the edge of the micro-patterned culture. A representative example of this is given in Figure 2.8(a). For cells not at the micro-pattern edge, distances to the boundary were discretised into 5 spatial groups. It was found that cells at the edge of the micro-pattern displayed significantly higher Nanog intensity than non-edge cells. Interestingly, non-edge cells located 50µm or less from the boundary also showed increased Nanog intensity compared to cells further towards the centre of the disc. Cells further than 50µm from the boundary showed a reduced difference in distributions of Nanog intensities. These results are shown in Figure 2.8(b). A Kolmogorov-Smirnov test was used to quantify the statistical significance of Nanog intensities between distributions [85].

As cells at the edge of the culture are reasoned to generally have fewer neighbours than those more embedded in the culture, the result that edge cells show higher Nanog intensity is unsurprising in relation to the results of Section 2.4.2. However, the result that cells close to - but not at - the edge also display increased Nanog intensity is interesting. This may be further indicative of a cell signalling distance greater than direct neighbours, but there is no reason from these results alone to suspect a signalling distance inclusive of the entire culture.

To verify whether cells at the micro-pattern boundary had fewer neighbours than other cells, a heat map was generated to display the percentage of cells with a given neighbour count for within each discrete group representing distance from micro-pattern boundary. This is given by Figure 2.9. Indeed, the majority of edge cells had only one or two neighbours, whereas cells further from the boundary showed a wider range of neighbour counts.

Some studies have linked patterns of gene expression among micro-patterned pluripotent stem cell cultures with reproducible patterns of cell density [30, 45]. In particular,



(a) Figure S3(a) in Tewary *et al.* (2017): 'Representative immunofluorescent images of colonies stained for DAPI, SOX2, and NANOG of geometrically confined hPSC colonies cultured in BMP4 supplemented N2B27. Scale bars represent 200μm.'

(b) Standardised density map showing the distribution of hESCs grown on disc micro-patterns. Data across 18 images is aggregated, with maximum 25 points per bin.

Figure 2.10: Density distributions of hPSCs grown on disc micro-patterns, where (a) shows an immunofluorescence image in Tewary *et al.* (2017), and (b) shows an aggregated cell density map for the 18 images analysed for this work.

when using disc micro-patterns, a 3D stack of cells has been reported to form in a radially symmetric pattern [30]. An example of this is given in Figure 2.10(a), which shows the spatial distribution of every cell nucleus on a disc using immunofluorescence staining. To investigate whether the results for Nanog expression inversely relating to cellular neighbour count (presented in Section 2.4.2) were confounded by this phenomenon, a binned density map was generated for each of the 18 images individually. No patterns of cell density were observed. The aggregated density map across all 18 images is given in Figure 2.10(b).

2.5 Data-Informed Modelling of Single-Cell Nanog Expression Using Spatial Data Analysis

In this section, regression models are trained to predict the Nanog intensity vector for single cells. The response variable is continuous and scaled in the range $[0, 1]$. The data was split into training and testing sets with the aim of yielding unbiased models. Model architectures used were linear regression and Gaussian process regression (GPR). Linear regression has been successfully applied to stem cell gene expression data in [95], demonstrating that linear models are applicable for these problems. GPR models are non-parametric Bayesian regressors known to be powerful for fitting complex functions. They also are able to take into account stochastic noise that may be present in the data [96], which is extremely important in the case of single-cell gene expression [68, 73]. The desired train:test ratio was 8:2, which is well within acceptable ratios used [97].

However, it was important not to split portions of the data pertaining to the same image; doing so would lose spatial information with respect to surrounding cells. Therefore, the combination of images that gave the closest approximation to the desired ratio were selected. Thus, 15 images were used for training, and 3 used for testing. The actual train:test ratio obtained was 0.794:0.206, where the training set consisted of 11724 data observations, and 3041 for the testing set.

Based on the previously discussed results, a small set of spatial features were used for model training:

1. Distance from micro-pattern boundary (μm),
2. Number of neighbours ($\mathbb{Z} \geq 0$),
3. Mean Nanog intensity of neighbouring cells ($\mathbb{R} \geq 0$).

2.5.1 Regression Model Architectures and Performance Metrics

Consider a data set with N observations and n features. Then, the input and output data is

$$\{x_i, y_i | i = 1, 2, ..., N\}, \quad x_i \in \mathbb{R}^d, y_i \in \mathbb{R},$$

taken from an unknown distribution. The model architectures used in the remainder of this section may then be defined as follows.

Linear Regression

Linear regression models take the form

$$\hat{y}_i = \beta_0 + \beta_1 x_{i,1} + ... + \beta_n x_{i,n} + \epsilon_i, \quad (i = 1, ..., N), \tag{2.8}$$

where $\hat{y}$ represents the fitted values, β_0 is the y-intercept, the remaining β values are the coefficients fitted for each feature, and ϵ_i is the error [98]. Coefficients are estimated such that the error term is minimised.

Gaussian Process Regression

The GPR model can be defined as

$$h(x)^T \beta + f(x), \tag{2.9}$$

where $f(x)$ is from a Gaussian process (GP) with zero mean and covariance function $k(x, x')$, i.e. $f(x) \sim GP(0, k(x, x'))$. The set of basis function $h(x)$ transform the original feature vector $x \in \mathbb{R}^d$ into a new feature vector $h(x) \in \mathbb{R}^p$, and β is a $p \times 1$ vector of basis function coefficients. The actual response variable y can be modelled as the probability

$$P(y_i|f(x_i), x_i) \sim N(y_i|h(x_i)^T\beta + f(x_i), \sigma^2) \tag{2.10}$$

with error variance σ. The remaining terms and concepts are defined as follows.

A GP is a set of random variables, such that any finite number of them have a joint Gaussian distribution. If $\{f(x)|x \in \mathbb{R}^d\}$ is a GP, then given N observations $x_1, x_2, ..., x_N$, the joint distribution of the random variables $f(x_1), f(x_2), ..., f(x_N)$ is Gaussian. A GP is defined by its mean function $m(x)$ and covariance function $k(x, x')$. That is, if $\{f(x)|x \in \mathbb{R}^d\}$ is a GP, then

$$E(f(x)) = m(x) \text{ and } Cov([f(x), f(x')]) = E[\{f(x) - m(x)\}\{f(x') - m(x')\}] = k(x, x').$$

A GPR model explains the response by introducing latent variables, $f(x_i)$ $(i = 1, 2, ..., N)$ from a GP and explicit basis functions h. For every observation x_i, a latent variable $f(x_i)$ is introduced. The covariance function of the latent variables captures the smoothness of the response and basis functions project the inputs x into a p-dimensional feature space. The covariance function $k(x, x')$ is parameterised by θ, which is a set of kernel parameters or hyperparameters [99].

Model Performance Metrics

To evaluate model performance, several metrics were used. These were the normalised root mean squared error (NRMSE), mean percentage error (MPE), the correlation coefficient, and the mean absolute difference between the Moran's Index (MI) of predicted and actual response variables. The first three of these are standard model evaluation metrics for machine learning [100], and MI was successfully implemented to quantify spatial relationships between Nanog expressing hESCs in [83]. These metrics are now defined below.

The root mean squared error (RMSE) between the actual response y and predicted response $\hat{y}$ is

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}} \quad . \tag{2.11}$$

Then, the NRMSE can be defined as

$$NRMSE = \frac{RMSE}{(\max(\hat{y}) - \min(\hat{y}))}. \tag{2.12}$$

The MPE of $\hat{y}$ is

$$MPE = \frac{1}{N}\sum_{i=1}^{N}\left(\left| \frac{\hat{y}_i - y_i}{\hat{y}_i} \right| \times 100 \right). \tag{2.13}$$

The correlation coefficient $r \in [-1, 1]$ between y and $\hat{y}$ is

$$r = \frac{N\sum y\hat{y} - \left(\sum y\right)\left(\sum \hat{y}\right)}{\sqrt{[N\sum y^2 - \left(\sum y\right)^2][N\sum \hat{y}^2 - \left(\sum \hat{y}\right)^2}}, \tag{2.14}$$

where $r = 0$ represents no statistically significant correlation, $r = -1$ total negative correlation, and $r = 1$ total positive correlation.

For a spatial variable y, the MI of spatial autocorrelation is defined as

$$MI(y) = \frac{N}{W}\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \quad MI \in [-1, 1], \tag{2.15}$$

where y has N observations, $\bar{y}$ is the mean of y, w_{ij} are the elements of a matrix of spatial weights with zeroes on the diagonal, and W is the total sum of elements in w. So that cells with no neighbours could be measured by MI, the matrix w was defined such that the weight between a given pair of cells was the inverse of their Euclidean distance. A value of $MI = 0$ indicates no statistically significant autocorrelation, $MI = -1$ total negative autocorrelation, and $MI = 1$ total positive autocorrelation [101].

Then, the mean absolute difference between MI values for $y, \hat{y}$ is

$$MI(y, \hat{y}) = \frac{1}{N}\sum_{i=1}^{N}\left| MI(\hat{y}) - MI(y) \right|. \tag{2.16}$$

2.5.2 Results of Linear Regression Modelling

For actual Nanog intensity y and fitted intensity $\hat{y}$, the linear regression model fitted to the training set was as follows:

$$\hat{y} = 0.385 - 0.000382a_1 - 0.0158a_2 + 10.4a_3, \tag{2.17}$$

where a_1 = distance from boundary, a_2 = number of neighbours, and a_3 = mean Nanog intensity of neighbours.

Based on this limited spatial feature set, when applied to the test set, the model yielded an NRMSE of 0.405, an MPE of 20.7%, MI$(y, \hat{y})$ of 0.103, and correlation coefficient 0.449. The inverse relationship found between cellular Nanog intensity and neighbour count is represented by Figure 2.11, showing also a linear best fit for the data. This demonstrates that the means of each group may be very well approximated by a linear model.

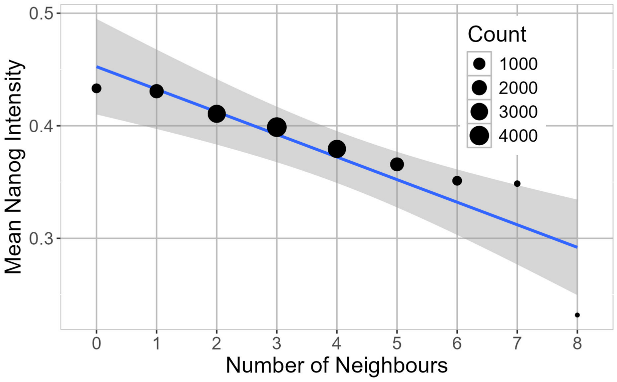


Figure 2.11: Plot showing mean Nanog intensity per neighbour count. A line of best fit with the formula $y \sim x$ is applied to the data, with confidence interval shown in grey.

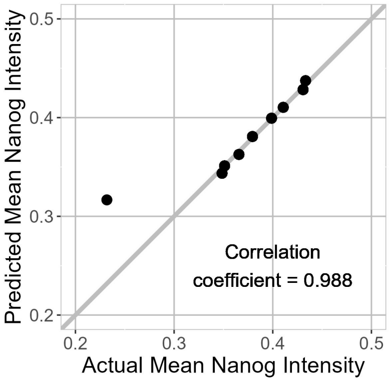


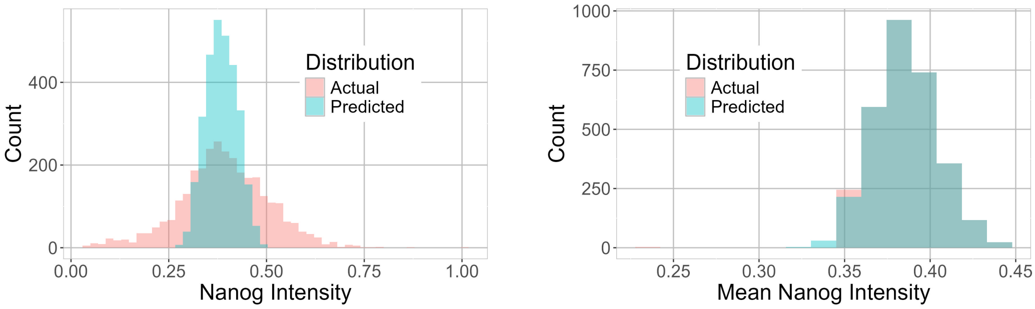
Figure 2.12: Scatter plot showing predicted vs. actual mean Nanog intensity (grouped by number of neighbours) for cells in the testing set. NRMSE = 0.0282, MPE = 0.309%, MI$(y, \hat{y})$ = 0.0117.

A linear regression model was fitted to the means of the variables a_1, a_2, a_3, denoted $\bar{a_1}$, $\bar{a_2}$, $\bar{a_3}$. The resultant model is given by

$$\hat{y} = 0.501 - 0.000956\bar{a_1} - 0.00929\bar{a_2} + 1.42\bar{a_3}. \tag{2.18}$$

This model yielded an NRMSE of 0.0282, an MPE of 0.309%, MI$(y, \hat{y})$ of 0.0117, and correlation coefficient 0.988. These results are - as expected - markedly improved as a consequence of predicting only the mean Nanog intensity per neighbour group. Results are shown in Figure 2.12.

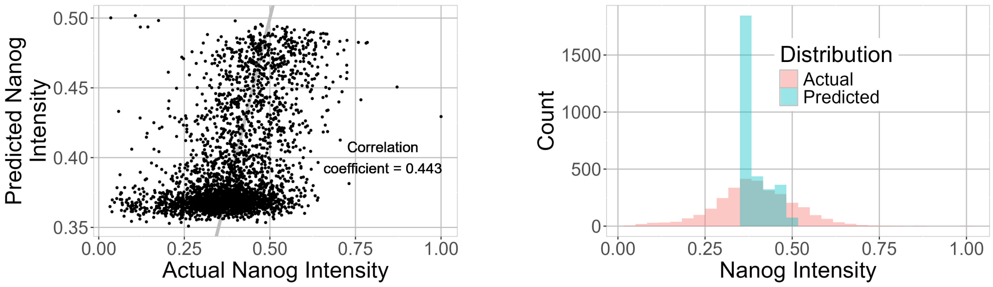
It was apparent that both linear regression models, given in Equations (2.17-2.18), predicted a reduced range of Nanog intensities compared to the actual data. To visualise this, overlaying histograms of predicted and actual response variables were plotted, which are shown in Figure 2.13. In terms of feature coefficients, distance from disc boundary and number of neighbours contributed negatively to the predicted value. This is in accordance with earlier results of this chapter shown in Figure 2.5. Conversely, the contribution from neighbouring Nanog intensities is positive, suggesting that cells exhibit



(a) Produced by Equation 2.17.

(b) Produced by Equation 2.18.

Figure 2.13: Histograms showing distributions of the predicted and actual response variable yielded by the models given in Equations (2.17-2.18), respectively.



(a) Scatter plot.

(b) Histogram.

Figure 2.14: Plots for visualising GPR model performance when used to predict Nanog intensity levels of cells in the testing set. NRMSE = 0.682, MPE = 20.0%, MI$(y, \hat{y})$ = 0.0194.

greater Nanog expression levels in the presence of increased average neighbourhood Nanog expression. This fits with the results shown in Figure 2.7. When Equation 2.18 was used to predict mean Nanog intensity per number of neighbours, all model performance metrics significantly improved from that yielded through Equation 2.17. This may suggest that there is an intrinsic relationship between Nanog expression and spatial features, which emerges for averaged populations, but may be confounded by stochastic Nanog fluctuations which introduce significant noise [68, 73].

2.5.3 Results of Gaussian Process Regression Modelling

The radial basis function kernel, i.e. the Gaussian kernel, was used for GPR model training. When used to predict the Nanog intensity vector of the testing set, the resultant GPR model had an NRMSE of 0.682, an MPE of 20.0%, a correlation coefficient of 0.443, and an MI$(y, \hat{y})$ of 0.0194. These results are shown in Figure 2.14(a). Similarly to the linear regression models, the GPR model produced a far reduced range of prediction values compared to actual Nanog intensity values. This is shown through overlayed intensity histograms for predicted and actual responses, given in Figure 2.14(b).

2.6 A Novel Spatial Model of Nanog Expression Combining Diffusive and Juxtacrine Signalling

2.6.1 Mathematical Formulation of Nanog Expression Model

The cell population is represented as a network, where nodes are centres of cell nuclei. There exists an edge between nodes if the corresponding cells are spatial neighbours (calculated according to Section 2.4.1), i.e. likely to be in physical contact. Nanog expression is then postulated as a linear combination of functions that relate to (1) positional information relative to the micro-pattern disc boundary, (2) cell-to-cell contact (juxtacrine signalling) in the neighbourhood, and (3) diffusive signals conducted through culture media (cells do not need to have neighbours to be subject to this).

The formulation of the cell network yields an undirected graph $G(V, E)$ with vertices V and edges E, where $|V| = N$ (the total number of cells) and $|E| \leq |V|$. Applying a shortest path algorithm to G, each cell pair i, j has a topological step distance

$$s_{ij} = \min(s(i, j)), \tag{2.19}$$

where $s()$ represents the number of steps involved in a possible path between i, j. As cells with no direct neighbours have no edges assigned to them in G, neighbour-less cells have no associated step distances. Cells are spatially non-uniform in culture, so the inter-cellular Euclidean distance measure $d()$ is also applied, such that

$$d_{ij} = (d(i, j)), \tag{2.20}$$

i.e. the linear distance. These may be regarded as edge weights.

The contact-based step distance given in Equation 2.19 may be used for the quantification of juxtacrine signalling. In conjunction with this, small molecules present in the culture media, which freely diffuse across the cell membrane, act as signals for the cells [102]. The diffusion profile often takes the form of a Gaussian distribution or, similarly, a first-order Bessel function [103]. Therefore, the Gaussian function

$$u_{ij} = \exp\left(\frac{-(d_{ij})^2}{4D}\right), \quad D \in [10^{-4.2}, 10^{-2}] \tag{2.21}$$

is used to describe the diffusive signalling strength between cells i, j, where D represents the diffusion constant [103, 104]. Diffusion does not rely on contact-based signalling, but is affected by Euclidean signalling distances. Examples of the diffusion profile with the maximum and minimum diffusion coefficient are given in Figure 2.15.

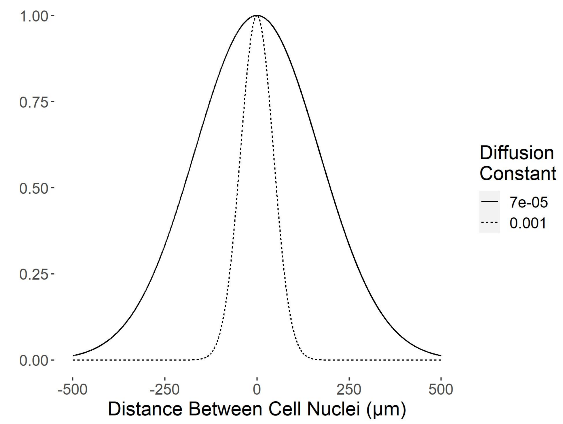


Figure 2.15: Diffusion profiles according to the Gaussian function $\exp(-(d_{ij})^2/4D)$, where d_{ij} is the Euclidean distance between the nuclei of cells i, j and D is the diffusion constant. Maximum and minimum values for D are plotted.

The model for predicting the Nanog expression of cell i is thereby given as

$$\begin{aligned} x_i &= \frac{c_1}{\Omega_i^{k_1}} + c_2 \sum_{j \neq i} x_j u_{ij} - \frac{c_3}{n_i^{k_2}} \sum_{j \neq i} x_j q^{(s_{ij}-1)} + \xi(i), \\ q, c_1, c_2, c_3 &\in [0,1], \quad k_1, k_2 \in [1,2], \end{aligned} \tag{2.22}$$

where Ω_i is the distance in µm between cell i and the disc boundary, n_i is the number of neighbours for cell i, and q represents the contact-based signalling strength. Notably, $q = 0$ results in signalling from only direct neighbours (those in contact), i.e. $s_{ij} = 1$, as then $0^{(s_{ij}-1)} = 1$ for all direct neighbours and 0 otherwise. Additionally, $q = 1$ assumes that culture-wide signals are maximally received, as $1^a = 1$ for all $a \in \mathbb{Z}$ [83]. As gene expression is subject to intrinsic noise that has significant impact on the overall cell state [105], the term $\xi(i)$ represents stochastic Gaussian noise with zero mean in the range $[-0.1, 0.1]$ applied to cell i. The y-intercept of the function is given by $c_1/\Omega_i^{k_1}$, with the remaining parameters c_2, c_3 representing the successfully received proportion of signalling. Parameters k_1, k_2 were permitted in the range $[1, 2]$ to investigate increasing non-linearity of the terms they are part of.

Model Optimisation

The goal of optimisation is to minimise a given cost function. For this work, the parameters $D, q, c_1, c_2, c_3, k_1, k_2$ were optimised to best fit the actual Nanog intensity vector. The basic minimisation problem is formulated as

$$\min_x f(x), \quad h(x) = 0, \quad g(x) \leq 0. \tag{2.23}$$

The solution to the optimisation problem was obtained using an interior-point algorithm, which solves a sequence of approximate minimisation problems. For each $\mu > 0$, the approximate problem is

$$\begin{aligned} &\min_{x,b} f_\mu(x,b) f(x) = \min_{x,b} f(x) - \mu \sum \ln(b_i), \\ &b \geq 0, \quad h(x) = 0, \quad g(x) + b = 0, \end{aligned} \tag{2.24}$$

where b_i are slack variables that must remain positive to ensure that iterations are kept in the interior of the feasible region. Notably,

$$\lim_{\mu \to 0} f_\mu = \min(f). \tag{2.25}$$

The logarithmic term in Equation 2.24 is a barrier function, which is a continuous function that tends to infinity as the interior point tends to the boundary of the feasible region [106].

As Equation 2.24 presents an equality problem, rather than the inequality problem posed by Equation 2.23, it is now easier to solve. The interior-point algorithm uses two step types at each iteration:

1. A direct step in (x,s), attempting to solve the Karush-Kuhn-Tucker (KKT) conditions using a linear approximation,

2. A conjugate gradient (CG) step, using trust conditions,

which are presented in order of choice. The auxiliary Lagrangian function is defined as

$$L(x, \lambda) = f(x) + \sum \lambda_{g,i} g_i(x) + \sum \lambda_{h,i} h_i(x), \tag{2.26}$$

where the vector $\lambda = [\lambda_{g,i}, \lambda_{h,i}]$ is the Lagrange multiplier, and g, h are the inequality and equality restraint vectors, respectively. Then the KKT conditions are:

$$\nabla_x L(x, \lambda) = 0, \quad \lambda_{g,i} g_i(x) = 0 \; \forall i. \tag{2.27}$$

Upon each iteration, the algorithm decreases a merit function:

$$f_\mu(x, b) + \nu ||(h(x), g(x) + b)||, \tag{2.28}$$

where the parameter ν may increase at each iteration to force the solution into the feasible region. If a step does not decrease the merit function, the step is rejected, with a new step implemented instead. Similarly, a new step is implemented if the function returns non-permitted values, e.g. a complex or non-numeric value.

The 'direct step' uses the following variables:

1. The Hessian of the Lagrangian of f_μ:
$$H = \nabla^2 f(x) + \sum_i \lambda_i \nabla^2 g_i(x) + \sum_j y_j \nabla^2 h_j(x).$$
2. The Jacobian of the constraint function g: J_g.
3. The Jacobian of the constraint function h: J_h.
4. $B = \text{diag}(b)$.
5. The Lagrange multiplier vector associated with constraints g: λ.
6. $\Lambda = \text{diag}\lambda$.
7. The Lagrange multiplier vector associated with h: y.
8. The vector of ones the same size as g: e.

Then, using a linearised Lagrangian, the direct step $(\Delta x, \Delta b)$ is as follows:

$$\begin{bmatrix} H & 0 & J_h^T & J_g^T \\ 0 & \Lambda & 0 & B \\ J_h & 0 & 0 & 0 \\ J_g & I & 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta b \\ \Delta y \\ \Delta \lambda \end{bmatrix} = - \begin{bmatrix} \nabla f(x) + J_h^T y + J_g^T \lambda \\ B\lambda - \mu e \\ h \\ g + b \end{bmatrix}. \tag{2.29}$$

For the 'conjugate gradient step', both x are b are altered, with $b > 0$. The approximate problem may be estimated by a quadratic expression in a trust region, subject to linearised constraints.
Introducing the parameter R, which is the radius of the trust region, the Lagrange multipliers
$$\nabla_x L = \nabla_x f(x) + \sum_i \lambda_i \nabla g_i(x) + \sum_j y_j \nabla h_j(x), \quad \lambda > 0, \tag{2.30}$$

are obtained by approximately solving the KKT equations using a least-squares method. Then a step $(\Delta x, \Delta b)$ is sufficient to approximately solve

$$\min_{\Delta x, \Delta b} \nabla f^T \Delta x + \frac{1}{2} \Delta x^T \nabla_{xx}^2 L \Delta x + \mu e^T B^{-1} \Delta b + \frac{1}{2} \Delta b^T B^{-1} \Lambda \Delta b, \tag{2.31}$$

subject to the linearised constraints

$$g(x) + J_g \Delta x + \Delta b = 0, \quad h(x) + J_h \Delta x = 0. \tag{2.32}$$

To solve Equation 2.32, the norm of the linearised constraints is computed and minimised by the algorithm within a region with radius scaled by R. By solving this equation to generate its residuals, Equation 2.31 is able to be resolved by matching its constraints with the residuals, remaining within R.

The barrier parameter μ, used in Equation 2.24, must be updated such that $\mu \rightarrow 0$ as iterations increase. To do so, the algorithm decreases μ by a factor of $1/100$ or $1/5$ when the approximate problem is solved with sufficient accuracy. The former fraction is used when the algorithm takes either 1 or 2 iterations to achieve sufficient accuracy, with the latter fraction used otherwise. The measure of accuracy is as follows:

$$\max\left(||\nabla f(x) + J_h^T y + J_g^T \lambda||, ||B\lambda - \mu e||, ||h||, ||g(x) + b||\right) < \mu. \tag{2.33}$$

This algorithm is taken from [107], which is implemented by MATLAB's 'fmincon' optimisation function.

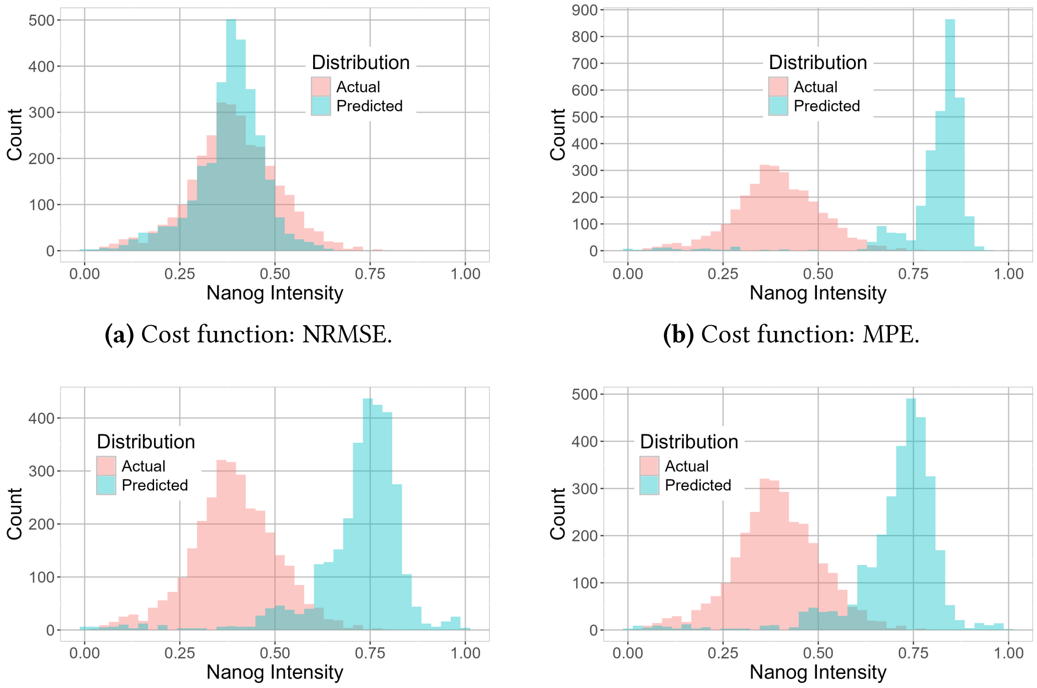
2.6.2 Modelling Results

To optimise parameters, four different cost functions (corresponding to the four model performance metrics used for regression modelling) were implemented:

1. NRMSE between predicted and actual Nanog intensity levels,
2. MPE between predicted and actual Nanog intensity levels,
3. Correlation coefficient between predicted and actual Nanog intensity levels,
4. Absolute value of the difference between the MI of predicted and actual Nanog intensity levels.

All optimisation was performed per image due to the construction of the optimisation function. Stopping criteria was defined such that optimisation outputs were within a 20% error tolerance for each cost function. For the first stage of parameter optimisation, cells that were assumed to respond only to diffusive signalling - i.e. those without neighbours - were selected from the image data. Then, the parameters D, c_1, c_2, k_1 were fitted on this subset. Next, using these results for the diffusive signalling portion of the model, the remaining parameters q, c_3, c_4, k_2 were fit on the cells that were assumed to respond to both diffusive and juxtacrine signalling, i.e. those with at least one neighbour.

When fitting parameters for cells assumed to respond only to diffusive signalling, results from using each of the cost functions were compared. The mean of each parameter value per image was computed and used as the fitted parameters. As there was little variation in each parameter across the images, this approach was deemed reasonable. Using these results, parameters were optimised for cells that have at least one neighbour, i.e. assumed to respond to both diffusive and juxtacrine signalling. After this was done for each image, the mean value of every parameter that had not yet been fitted was calculated, as again



(a) Cost function: NRMSE.

(b) Cost function: MPE.

(c) Cost function: correlation coefficient.

(d) Cost function: Moran's Index.

Figure 2.16: Histograms showing the distribution of actual and predicted Nanog intensity levels, generated by Equation 2.22, using the given cost function for model optimisation.

Table 2.1: Fitted parameter values for the Nanog expression model.

c_1	c_2	c_3	D	q	k_1	k_2
0.415	0.0469	4.55e-08	0.00552	0.213	1.25	1.37

there was little variation in these parameters across images.

Histograms for results generated using each of the four cost functions are shown in Figure 2.16. As this figure shows, the best fitting histogram is that which uses NRMSE as the cost function. Hence, this was taken to be the final model - fitted parameters for this model are given in Table 2.1. The model gave an NRMSE of 0.142, an MPE of 78.4%, a correlation coefficient of 0.0441, and an $\mathrm{MI}(y, \hat{y})$ of 0.0420.

2.7 Summary and Discussion

In this chapter, a novel spatial analysis and novel spatial models of Nanog expression have been presented. The goal was to progress understanding of how Nanog expression in pluripotent stem cells is affected by signalling cues from the cellular environment. A new method for quantifying a cell's number of neighbours was introduced, which combines Delaunay triangulation and a radial distance threshold, thereby removing the need for cell membrane staining during cell culture. Informed by results of neighbourhood analysis, a novel predictive model is developed for the static and continuous Nanog expression of single cells. The model is the first to combine the effects of both diffusive and juxtacrine signalling in pluripotent cell populations.

Investigation of the Nanog distribution across all cells revealed a large HN population, with a much smaller LN population ($\sim$5%) [68]. LN cells displayed significant differences in nuclear morphology, showing a reduction in nuclear area and perimeter compared to HN cells. From existing literature [86], this provides support that the LN cells were in a state of low pluripotency and more prone to differentiation.

Spatial analysis carried out using the new methodology for neighbour computation revealed that Nanog intensity was inversely related to neighbour count. Interestingly, while apparent for the total cell population, the LN cells alone did not display this relationship. Taking a random sample of HN cells that matched the size of the LN population generally preserved the relationship between Nanog and neighbour count, suggesting that LN cells are intrinsically less susceptible to juxtacrine signalling cues from their neighbours. As these cells are likely more prone to differentiation, these results may suggest that, while cells are pluripotent, the gene expression dynamics of pluripotency regulators in the local cellular neighbourhood plays an important role in regulating cell state. However, cells that have begun to lose pluripotency may become less dependent on the pluripotency state of their neighbours as a decision-making mechanism.

Further analysis was performed to investigate the relationship between a cell's Nanog intensity and the mean continuous intensity of its neighbours. This revealed that LN cells generally showed no relationship between these two variables, whereas the Nanog intensity of HN cells generally increased as the mean intensity of their neighbours increased.

Previous reports have been made regarding the edge-sensing mechanisms of stem cells [49]. As expected, cells at the edge of the disc micro-patterns exhibited fewer neighbours and higher Nanog intensities. Interestingly, cells that were close to - but not at - the boundary also had slightly increased Nanog intensities compared to cells more embedded in the cultures, despite having more neighbours. This may be due to diffusive signalling cues which travel across a wider signalling environment than neighbours only [78]. To determine whether the inverse relationship observed between Nanog expression and neighbour count was a result global cell density as has previously been reported [45, 30], a density map was produced across all images, showing a fairly uniform distribution of cells across each disc. Therefore, the spatial results were unlikely to be confounded by global density phenomena.

Linear regression models were trained on each image to predict continuous single-cell Nanog intensity using only three variables: (1) distance between the cell and disc boundary (µm), (2) the cell's number of neighbours ($\mathbb{Z} \geq 0$), and (3) the mean Nanog expression level of directly neighbouring cells ($\mathbb{R} \geq 0$). Across all images, the model yielded an NRMSE of 0.405, MPE of 20.7%, correlation coefficient of 0.449, and MI($\hat{y}, y$) of 0.103 between predicted and actual Nanog expression levels.

A data-informed spatial model of Nanog expression levels that accounts for (1) positional information relative to the micro-pattern disc boundary, (2) cell-to-cell contact (juxtacrine signalling) in the neighbourhood, and (3) diffusive signals conducted through culture media was developed. The model achieved an NRMSE of 0.142, MPE of 78.4%, correlation coefficient of 0.0441, and MI($\hat{y}, y$) of 0.0420. Unlike the linear regressor, this model also incorporates stochastic Gaussian noise to represent the intrinsic boisy fluctuations in cellular Nanog expression, which may negatively impact the predictive accuracy of the model.

Interestingly, for the HN population only, the model achieved a good NRMSE of 0.135, (less than that of the model overall), while the NRMSE for the LN population was 0.389, which is almost thrice that of the HN population. Thus, this chapter has shown that LN cells are less influenced by their cellular environment, which is supported through LN cell state less able to be accurately modelled by signalling cues from surrounding cells compared to the HN population.

The modelling presented in this chapter has limited application potential due to the low accuracies produced. Due to the extreme complexity of embryogenesis, a model that relies on a single genetic marker may not be sufficient. Moreover, stem cell state is constantly changing prior to fate acquisition [68, 108, 109]. It may be the case that predicting the static expression levels of a gene expressed during pluripotency - which captures only a snapshot of the noisy and dynamic cell state - based on spatial features alone may utilise too little information about the cells to give highly accurate results. However, it offers a much more cost and time effective approach to obtaining time-series gene expression data. Despite low correlation between actual and predicted intensities, the histograms of the two distributions matched reasonably well. Therefore, it may be concluded that, if seeking a population-wide estimate of a gene intensity distribution, the model proposed in this chapter could provide a useful tool - particularly to gain insight into neighbourhood signalling. Modelling the dynamic evolution of cell state during cell culture expansion requires dynamic models with temporal evolution, which may be more appropriate [82, 83].

Induced differentiation of PSCs towards one specific fate for the entire cell population minimises genetic heterogeneity of fate markers. Appendix C presents a detailed methodology for classifying single-cell binary Sox10 expression in H9TVD3 hESCs using information about the local neighbourhood; this model achieved an excellent accuracy of 92.7%.

Chapter 3

A Novel Symmetry-Based Dynamical Modelling Paradigm for Stem Cell Fate Patterning

3.1 Introduction

There are many factors known to influence lineage commitment of stem cells. A key factor uncovered over the past decade is the shape of the culture domain, which can be controlled through micro-patterned plate inserts which lead to selective cell adherence on the micro-pattern surface. At present, mechanisms by which culture geometry regulates hPSC fate patterning are not fully understood. In this chapter, this problem is addressed. Stem cell cultures are represented as a network, where nodes correspond to cells, and edges exist between cells in direct contact in culture. A novel method for defining symmetries of an undirected graph is then proposed. For this, an iterative procedure is developed to partition the graph's cells and edges into equivalence classes, where cell equivalence relations correspond to cell symmetries. This leverages existing methods by evaluating the graph in full, rather than each cell's local neighbourhood alone. Importantly, each cell has internal dynamics which take into account its spatial relation to all other cells. For this, novel dynamical model system is proposed, describing the spread of a differentiating morphogen across the cell culture. Thus, a dynamical vertex model of differentiating cells in culture is developed. Equilibria of the dynamical system are modelled as final cell fates, as morphogen concentration is temporally invariant at system equilibrium. Patterning of cell equilibria qualitatively recapitulates spatial fate patterning found in numerous published experiments. Moreover, the model is topologically flexible and can record single-cell gene expression evolution for growing cell populations. Thus, the novel model presented offers a promising and useful paradigm that may help to understand how cell fate patterning is impacted by culture geometry.

Structure of this Chapter

The remainder of this chapter is set out as follows. Section 3.2 provides background and a literature review within the field. In Section 3.3, the reaction-diffusion model given in [30] for morphogen concentration is re-created, and sensitivity of the results to sensible alterations to the initial Noggin profile is investigated. Section 3.4 models cell-to-cell interactions using the coupled map lattice architecture, and it is shown that this architecture lacks ability to accurately produce hPSC fate patterning for micro-patterned cultures. Section 3.5 formalises a novel methodology for assigning symmetry relations to nodes of an undirected graph. Section 3.6 introduces a novel dynamical system for morphogen concentration, which is modelled as internal dynamics for nodes of the graph. Simulated results are given in Section 3.7, representing a range of micro-patterned culture domains. Section 3.8 applies this model to a novel growing cell network for increased biological relevance, simulating the growth of a hPSC population from a single cell. Finally, Section 3.9 presents a conclusion of this chapter and a discussion of the results.

3.2 Background and Literature Review

One of the earliest lineage decisions that hPSCs make is differentiation into one of the three primary germ layers: ectoderm, mesoderm or endoderm. These germ layers are the basis of all somatic cell types in the body [25]. In unconstrained hPSC cultures, heterogeneous differentiation of cells into each of the germ layers can be observed, but there remains no clear link between cell fate and the spatial organisation of the cells [34].

Cell Fate Patterning is Dependent on Culture Geometry

To investigate whether the heterogeneous patterning of cell fate for unconstrained cultures is due to the colonies' irregular geometry, hPSCs can be grown on micro-patterns to control the size and shape of colonies. A groundbreaking study imposing radial symmetry on cultures of hESCs was conducted in 2014, where cells were grown on micro-patterned discs of 250, 500, and 1000 μm in diameter [31]. The micro-patterns used were CYTOOchips™, which have an activated adhesive surface that is able to be coated as normal in stem cell culture.

To induce differentiation of hESCs, bone morphogenetic protein 4 (BMP4) ligand was administered to cell cultures at a homogeneous concentration of 50ng/ml. BMP4 is crucial during early embryogenesis; performing a knockout of BMP4 leads to embryonic death, and BMP4 deficient mice do not develop mesodermal cells [110]. BMP4 has been found to induce differentiation into trophectoderm and also a combination of embryonic

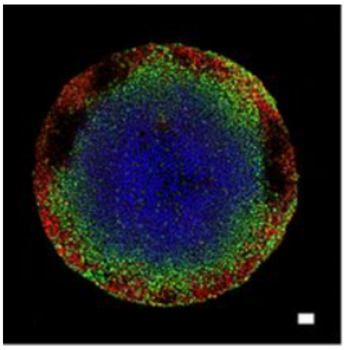


Figure 3.1: Immunofluorescence staining image from Tewary *et al.* (2017), showing rings of gene expression patterning in differentiated hESCs on a disc. Blue, green, and red staining shows Sox2, Brachyury, and CDX2 positive cells, respectively. Scale bar: 50μm.

and extra-embryonic mesoderm [111]. Trophectoderm refers to the outer layer of the blastocyst, also called trophoblast [112]. After administering BMP4, the cells were observed to have differentiated after 24 hours. After 42 hours, the cells were fixed with 4% paraformaldehyde, and subsequently stained for the genes Oct4, Nanog, Sox2, CDX2, Brachyury, Sox17, and Eomes. Pluripotent cells express the first three of these genes, but Sox2 is also maintained when cells differentiate to ectoderm and is down-regulated when cells differentiate to mesendoderm. Conversely, Oct4 and Nanog regulation behaves *vice-versa.* The remaining genes CDX2, Brachyury, Sox17, and Eomes are not associated with pluripotency, and are only expressed when cells begin to differentiate; CDX2 and Brachyury both mark differentiation to mesoderm, whilst Sox17 marks endoderm and Eomes marks mesendoderm.

On discs of $1000\,\mu\mathrm{m}$ diameter, it was found that gene expression among the cells followed a reproducible, radially symmetric pattern. Cell fate was subsequently inferred from the co-expression of genes. Starting at the centre of the colonies and moving outwards, rings of ectoderm, mesoderm, endoderm and trophectoderm were observed. The corresponding immunofluorescence microscopy image is given by Figure 3.1. Crucially, this result demonstrates how imposing a simple spatial constraint on hESC cultures can overcome the heterogeneous patterning that occurs in unconstrained cultures.

Furthermore, this phenomena was found to differ when the colony radius was decreased; the same gene expression was observed at the colony edges, but the Sox2 positive regions towards the colony centres were lost in favour of Nanog positive cells. In colonies of $250\,\mu\mathrm{m}$ diameter, Sox2 expression was lost entirely. This may indicate that patterning is controlled from the colony edge, possibly because a diffusible inhibitor with a high central concentration is lost at the colony edge. To test this, a gene knockdown experiment was conducted on the BMP inhibitors Chordin and Noggin, resulting in a higher expression of mesodermal markers in the centre of the colony, as was expected.

The protein Noggin, which is directly induced by BMP4 exclusively in hPSCs, has been found to be responsible for signalling asymmetries among cultures of hPSCs on disc micro-patterns [49]. Deleting Noggin was found to produce a spatially homogeneous profile of the gene pSMAD1, which was otherwise localised at the colony border after BMP4 treatment. Edge-sensing mechanisms were concluded to control fate patterning through the spatial organisation of TGF-β receptors on hPSCs. Specifically, at high cell density, hPSC TGF-β receptors are lateralised to face the centre of the cell colony, whereas cells at the colony edge - subject to reduced cell density - remain to have TGF-β receptors located at their apex. As BMP4 is part of the TGF-β family, this provides an explanation of the observed cell fate patterning.

The literature discussed above reflects the initial sole focus on disc micro-patterns as spatial control for hPSC fate. In more recent years, additional geometries have been exploited as micro-patterns for cell culture.

In 2020, the effects of differentiating hESCs with BMP4 with respect to mesodermal fate patterning was investigated using various micro-pattern geometries [47]. The purpose of using different geometries was to explore the link between mechanical tension in the cell culture and the resultant cell fate. As Brachyury is a mesodermal marker, immunofluorescence imaging was used to quantify Brachyury expression levels per cell. Particular attention was paid to an equilateral triangle micro-pattern, and a related trapezium which was obtained by removing the triangle's upper section. In the case of the equilateral triangle, Brachyury was found to localise at the three vertices 30 hours after BMP4 differentiation. At around 36 hours, hESCs began to express Brachyury around the entire perimeter of the micro-pattern in addition to the vertices. Upon measuring the traction stresses across the micro-pattern, the authors found that, at 30 hours, these corresponded to the triangle vertices. Therefore, high traction stress may induce mesodermal cell fate. As further investigation, a trapezium micro-pattern was created by removing the upper portion of the equilateral triangle. The result of this was decreased traction stress where the micro-pattern was cut, whilst retaining high traction stress at the lower two vertices. As the authors expected, Brachyury localisation was lost at the upper part of the trapezium, but remained localised at the lower vertices. It has since been found that the protein actomyosin regulates mechanical tension in the embryo of Drosophila [113], which may also apply to mammalian systems.

Mathematical Models for hPSC Lineage Decisions

Interpretations and inferences are often drawn from experimental results through quantitative analysis. In particular, how hPSCs arrive at a specific cell fate may be modelled mathematically, such as by considering the changing gene expression profile of each cell as the cell acquires lineage. An influential model of this kind was proposed in 2005 [114], inspired by Waddington's epigenetic landscape [115]. The concept is that every cell can be represented as a point in high dimensional state space, where the entire state space is akin to Waddington's landscape. Each axis corresponds to the cell's expression level of a particular gene; this is why the dimensionality of the system is high, as hPSCs may express some level of thousands of genes. Mathematically, for a system of N genes, this is represented by the following:

$$S(t) = [x_1(t), x_2(t), ..., x_N(t)],$$

where $S(t)$ is the state of the cell at time t, and $[x_1(t), x_2(t), ..., x_N(t)]$ is a state vector, with $x_i(t)$ (for $i = 1, ..., N$) having some effect on each other due to activatory/inhibitory gene responses. The authors use the term 'attractor state' to describe the steady states of the system, corresponding to cell fates. These attractor states may be thought of as sinks, as proposed by Waddington, in which cells that are close enough get drawn into. The goal of this research was to determine the most efficient way of pushing cells towards a specific lineage; if the state vector of a cell is known, together with the state vector of the desired fate (attractor state), the shortest path between these two vectors can be calculated. Moreover, a cell's trajectory need only enter the region of the state space that would pull it towards the desired attractor state, known as the basin of attraction. Knowing the minimal conditions required to achieve this could allow the design of efficient protocols for altering cell fate.

The research on this topic was continued in a paper published in 2007 [116]. To investigate the sort of behaviour that might be expected from the system, the authors considered a simple auto-regulatory network that takes the form of conventional gene regulatory network [117]. This models two transcription factors that are well understood: GATA1 and PU.1. These two proteins are known to govern the lineage of multipotent progenitor cells. Whilst progenitors are not hPSCs, they are closely related, and may offer useful insights into lineage acquisition using a far less complex system. Considering the cell state S, we have that $S(t) =$[GATA1, PU.1] at time t. Denoting GATA1 and PU.1 as x_1 and x_2 respectively, the authors derived the following system of ODEs:

$$\begin{aligned}\frac{dx_1}{dt} &= a_1\frac{x_1^n}{\theta_{a_1}^n + x_1^n} + b_1\frac{\theta_{b_1}^n}{\theta_{b_1}^n + x_2^n} - k_1x_1,\\ \frac{dx_2}{dt} &= a_2\frac{x_2^n}{\theta_{a_2}^n + x_2^n} + b_2\frac{\theta_{b_2}^n}{\theta_{b_2}^n + x_1^n} - k_2x_2,\end{aligned}$$

where a_1, a_2, b_1, b_2, k_1, k_2, and all of the θs are non-negative parameters. The first, second and third terms in each equation represent auto-stimulation, cross-inhibition and unregulated decay, respectively. The aim of this system was to analyse the qualitative behaviour exhibited by varying the parameters - rather than imposing restrictions by fitting the model to experimental data - to gain information about the general mechanisms surrounding cell fate.

Analysis of the system's time-invariant solutions revealed two stable and one unstable steady states, the latter taking the form of a saddle point around the basin of attraction of the two stable steady states. Biologically, the stable steady states correspond to specific cell fates (erythroid and myeloid respectively), whilst the third represents a bipotent progenitor state, able to acquire either fate. This simplistic model serves the purpose of showing that formalising cell state in this way indeed allows the differentiation potential of a cell to be quantified, and the stability of possible fates to be analysed mathematically.

For reasons explained by Huang in a solo-authored paper on the topic, published in 2010, a model of this type would need to be stochastic in order to be biologically accurate [118]. Upon introducing the state space model, the assumption was made that each cell may be regarded as a deterministic point. However, Huang concluded that the cell must be regarded as a cloud of points, as the inevitable gene expression noise of a cell will cause random fluctuations to the cell's state [73]. Consequently, any trajectory that the cell follows will also be stochastic, taking the form of a random walk. This further complicates the model beyond the already challenging high dimensional state space and difficulty obtaining the relevant single-cell data with which to validate the model, and no additional research has been published explicitly on the topic of modelling hPSC fate using these methods. This suggests that capturing the breadth of hPSC dynamics in such a model may be too difficult to achieve at present. Although inferring lineage trajectories from this type of model may not be currently achievable, it provided a mathematical framework for modelling hPSCs that many other researchers have built upon.

Alternative models may focus not on distances between gene expression states, but on how gene expression patterns change in response to varied differentiation signals. The formation of vertebrate embryos entails distinctions between the anterior-posterior and the dorsal-ventral axes [119]. Using a 3D model, the formation of human axes was investigated *in vitro* with hESCs [120]. BMP4 was administered to colonies of hESCs, showing that axial symmetry breaks despite the dose of BMP4 being homogeneously distributed. This followed a similar study conducted by the same authors in 2018 [121]. In accordance with previous studies [122], it was found that altering the concentration of BMP4 administered to the cultures was sufficient to produce differences in patterning. This may be a result of the increased probability of BMP4 at high concentrations coming into contact with the BMP4 binding site on hESCs.

Some useful mathematical tools are introduced in this paper to quantify symmetry breaking. The first is a vector, denoted by $\vec{\mu}$, which measures the spatial segregation between the Brachyury positive and Sox2 positive domains of the cultures. The magnitude of $\vec{\mu}$ increases/decreases as the degree of symmetry breaking increases/decreases. Using this metric, it was found that $\vec{\mu}$ is largely unaffected by the size of colonies. The second metric used is a cell mixing parameter σ, quantifying the compactness of the regions of Brachyury positive cells and Sox2 positive cells. A high σ corresponds to complete mixing, i.e. salt-and-pepper mixing. It was found that the regions were not well mixed irrespective of colony radius. However, regions of Brachyury positive cells had consistently higher density than regions of Sox2 positive cells. These results raise interesting questions about the link between culture morphology and signalling within the culture. In particular, the results may indicate a relationship between cell density and gene expression, though this is unexplored by the authors. Importantly, blocking specific signalling pathways led to symmetry breaking despite the cultures not morphologically resembling the embryo. This may suggest that the local cellular neighbourhood has a greater impact on cell behaviour than culture-wide features.

To explain the radially symmetric fate patterning yielded by differentiating hESCs on a disc micro-pattern [31], a novel reaction-diffusion system was proposed in 2017, consisting of two coupled partial differential equations (PDEs) [30]. The system describes the dynamics of BMP4/Noggin interaction in cultures of hESCs; recall that Noggin is an exogenous inhibitor of BMP4. This work followed from the experimental results, along with the simple reaction-diffusion model, earlier discussed. The goal was that by observing the patterns of BMP4/Noggin concentration predicted by the model, some degree of the heterogeneity in cell fate may be explained.

The non-dimensionalised reaction-diffusion system is as follows:

$$\frac{\partial bmp^*}{\partial t} = a_{BMP} bmp^* + b_{BMP} nog^* + c^*_{BMP} - d_{BMP} bmp^* + D_{BMP} \nabla^2 bmp^*, \tag{3.1}$$

$$\frac{\partial nog^*}{\partial t} = a_{NOG} bmp^* + b_{NOG} nog^* + c^*_{NOG} - d_{NOG} nog^* + D_{NOG} \nabla^2 nog^*, \tag{3.2}$$

with initial conditions

$$bmp^*(t=0) = BMPi, \tag{3.3}$$

$$nog^*(t=0) = \{0 \leq 1 - \left(\frac{x}{R}\right)^2 - \left(\frac{y}{R}\right)^2\}, \tag{3.4}$$

and boundary conditions

$$bmp^*(\Omega) = BMPi, \tag{3.5}$$

$$\frac{d(nog^*(\Omega))}{d\Omega} = 0, \tag{3.6}$$

where R represents the colony radius and Ω represents the boundary. Thus, Equation 3.6 is the Neumann boundary condition describing no-flux. The value $BMPi$ is the initial concentration of free bmp^* in the system. Parameter values are given in Table 3.1.

Table 3.1: Parameter values given by Tewary *et. al* (2017) for their reaction-diffusion model describing the concentrations of BMP4 and its inhibitor Noggin.

Parameter	Value
α	0.005
b_{BMP}	0.01
c^*_{BMP}	0.003
d_{BMP}	0.003
D_{BMP}	11
b_{NOG}	0
c^*_{NOG}	-0.015
d_{NOG}	0.009
D_{NOG}	55
γ_{BMP}	-
γ_{NOG}	0.0025
a_{BMP}	$\alpha(1 + BMPi \times \gamma_{BMP})$
a_{NOG}	$\alpha(1 + BMPi \times \gamma_{NOG})$

To simulate the results of the model computationally, the authors used finite element method to discretise the equations using the software COMSOL. As part of this process, a circular domain, representing the 2D colony, was partitioned by a triangular mesh. Videos were generated by running the simulation over an arbitrary number of discrete time steps. The model was able to produce results which closely resemble the authors' experimental data. The only values being varied each time were the initial concentration of BMP4 (0, 50, or 200ng/ml) and the colony diameter (1000 or $3000\,\mu\mathrm{m}$).

However, whilst this model does produce results that closely match what is observed biologically, deterministic reaction-diffusion models are often highly sensitive to their initial conditions [123]. This raises questions regarding how reliable the model is, as it is unlikely that precisely the same biological conditions are present during the development of each human embryo, particularly with respect to the initial profile of Noggin as a radially symmetric parabolic ellipsoid. Moreover, it has been shown that human embryos can be artificially split into its four blastomeres, with each blastomere going on to develop normally into an embryo [44]. This strongly implies that the mechanisms of early embryonic development are not contingent on the exact maintenance of specific conditions, but rather are able to overcome significant system perturbations. Thus, despite the accuracy of the model when simulated under chosen conditions, it may not be sufficient to capture the true robustness of molecular systems during early embryogenesis.

A contraction-reaction-diffusion model for hESC fate patterning was proposed in 2021 as a direct response to the issues discussed above [48]. The authors note that traditional reaction-diffusion systems are not capable of generating concentric ring patterns. In agreement with experimental results linking hPSC fate patterning to traction stress

[47, 113], this paper incorporates results from traditional reaction-diffusion with the effect of contraction leading to either flux-driven diffusion or the activation of signalling molecules. The paper is entirely theoretical, consisting of PDE systems and their corresponding computational simulations in COMSOL. The goal was to find conditions on parameters to best capture the rings of hPSC fate following BMP4 differentiation first demonstrated in 2014 [31].

This model is highly successful at producing robust ring patterning, as the initial condition can be randomised at the single-cell level whilst yielding the desired result. The wavenumber-related parameter is defined at the point of fastest growing instability within the system. The authors found that this parameter was crucial for controlling the activity of the two signalling molecules being modelled, such that a high wavenumber-related parameter corresponded to increased waves in the activity of the molecules. Importantly, this parameter dictates the symmetry of the resultant pattern.

The ring pattern is not an equilibrium of the contraction-RD model, as it is not a time-invariant state; simulations continue to show changing patterns of morphogen concentration if run further. This result opposes the popular reasoning originating from Waddington's epigenetic landscape [115]: a framework in which cell fates are points acting as 'attractor sites' which would take significant energy to push from their state.

Application of Vertex-Based Models to Stem Cell Biology

A modelling approach to self-organised hPSC fate patterning, focusing on how colonies form, was proposed in 2019 [46]. A gene knockdown of CDH1 and ROCK1 was performed on iPSCs, as these are known to affect stem cell colony organisation. Specifically, CDH1 modulation facilitates changes in how neighbouring cells adhere to one another. A multi-scale Cellular Potts model - a simple yet effective modelling paradigm where each cell occupies some number of squares which reflect their spatial organisation - was used in conjunction with parameter optimisation to produce desired fate patterns. The dynamics of a differentiating morphogen across the cell culture was able to be accurately modelled in this way. Interestingly, the authors point out that former reports of stem cell fate patterning, such as through the use of disc micro-patterns, loses key information by only investigating a static result. In other words, as cells are fixed in place before being antibody stained for immunofluorescence imaging, the process leading to the cells' final gene expression states are not seen.

Cell division was modelled as asynchronous, mimicking population growth kinetics. The direction in which cells moved was crucially influenced by the relative cell adhesion strength of neighbouring cells. After segmenting and tracking single cells *in vitro*, a basic grid search was used to fit the model parameters to the experimental cell velocity measurements. This model - in addition to an agent-based machine learning model that was able to grow from a small number of cells to reproduce iPSC gene expression patterning given a target image - were both accurate in recreating experimental fate patterning.

Potentially the most interesting features of this paper are (1) developing a growing model rather than static, and (2) not using micro-patterns to control culture boundary conditions. For point (1), as was later mentioned in other works [48], examination of a cell culture's static gene expression profile at an arbitrary time after differentiating the cells does not elucidate how cells arrive at that point. Rules for these patterning outcomes are still not understood, as no solid explanation is offered for how the model's results reflect biological processes. Interestingly, it was reported that observing cells 48 hours after inducing differentiation with BMP4 was not sufficient to accurately capture Brachyury patterning, which exhibited transient expression. This may have implications on static patterning studies, particularly with respect to the mathematical equilibria of gene expression states.

Embryogenesis is a highly complex process; combining various modelling architectures at different scales to recapitulate stem cell fate patterning may be necessary to capture this complexity. A combination of agent-based, RD, and gene regulatory models have been used for this purpose [50]. Interestingly, agents are not defined as cells, but rather genes: CDX2, Oct4, Sox2, Sox17, and TBXT. Gene activity was linked to three discrete states of morphogen concentration: high, low, or off. Depending on the combined levels of morphogens BMP4, Wnt3, Noggin, and DKK, together with activatory/inhibitory relations between genes, each gene expression state was also discretised into categories high, low, or off.

A state-flow style model architecture was used to simulate the system, whereby blocks of code would be run until the simulated micro-pattern was confluent. Actions that could be taken by agents were proliferation or differentiation. Initially, RD was set as inactive (as there is no morphogen signalling), and the agent-based gene model took a sparse pattern: Oct4 not expressed, Sox2 highly expressed, and the remaining genes randomly initialised. The subsequent onset of signalling from the gene regulatory network triggered the activation of RD, producing signalling gradients. In response to these gradients, the agent-based further drives gene regulation, leading to a feedback system.

This model does not incorporate an edge-sensing mechanism, yet experimental fate patterning results on a disc are reproduced. However, the model has only been simulated on a disc, so it is unknown whether it can be generalised across other cell culture geometries. The idea to model agents as genes, rather than cells, tackles the complication of cells' ability to express hundreds of genes. However, the relationship between the model's output and biological cells is not straightforward. There is a gap, therefore, between the theoretical research and its physical implementation.

Radial Asymmetry of Stem Cell Cultures Impacts Fate Patterning

When considering models where cell behaviour is assumed to rely solely upon the relationships between cells, without representing the cells on a rigid spatial structure, it is natural to think about how the topology and interactions of the network change as cells are added or removed.

The impact of culture geometry on cellular Brachyury expression among mouse embryonic stem cells (mESCs) was investigated in 2018 [45]. This research is rare in its approach to the spatial confinement of stem cell colonies: whilst much of the literature solely with disc micro-patterns, this paper addresses how breaking the radial symmetry gives rise to ordered asymmetric cell patterning. In doing this, the researchers show that such patterning is not merely due to diffusive signals and is heavily dependent on the local cellular neighbourhood, contrary to the assumptions of the reaction-diffusion model in other works [124], and in accordance with the theory of vertex models.

As their initial experiment, the authors plate mESCs onto both disc and ellipse micro-patterns, each with an area of $30\,000\ \mu\mathrm{m}^2$. Similarly to what was observed in [31], following from a differentiation procedure, Brachyury positive cells were located mainly towards the border of the discs. However, this pattern did not occur for ellipses; Brachyury positive cells were primarily located at the two tips of the ellipses, rather than across the whole border. Interestingly, the Brachyury positive cells localised at both tips of the ellipses only 35% of the time, and at just one tip 40% of the time. The authors postulated that the global density of cells during the pre-culturing stage, before plating onto the micro-patterns, may cause this variability. Cells were therefore subsequently grown at either high, medium or low densities in the pre-culture stage, and their Brachyury expression measured after 48 hours as before. Strikingly, although the percentage of Brachyury positive cells was found to be negatively correlated with global density, the number of Brachyury positive cells was unable to be normalised through controlling the shape and size of cultures. This suggests that while global density affects the number of Brachyury positive cells, it has no effect on their patterning. This led the authors to look instead at the impact of local cell density on Brachyury expression.

For each cell, the number of neighbouring cells was computed, using a manually specified radial distance threshold of $75\ \mu\mathrm{m}$. The authors found that Brachyury positive cells consistently localised in regions of lower local density than Brachyury negative cells on micro-patterns, although this phenomena was not seen for unconstrained cultures. Using a manual distance threshold to quantify each cell's number of neighbours provides flexibility; the threshold can be varied to examine the effect, if any, such variation has on the results. In this case, increasing the neighbourhood radius increased the difference in neighbour count between Brachyury positive and negative cells. This is unsurprising as the difference is presented as a raw difference in neighbour count, which may be misleading. Representing this information as a percentage difference is likely to be more a appropriate metric. Moreover, specifying a distance threshold in this way ultimately involves subjective judgment. There is no clear indicator that 75 µm is the most appropriate value to use for primary analysis, and so it may be useful to repeat the analysis using Delaunay triangulation to determine nearest-neighbour relationships.

The authors then considered whether a low concentration of Brachyury inhibitors at the ellipse tips was responsible for the localisation of Brachyury positive cells there. Micro-patterns in the shape of four-pointed flowers were developed, where each flower petal is an ellipse 25 µm apart from each other arranged in a cross shape. This distance between ellipses was as close as possible without the ellipses making contact. To test diffusive signalling between micro-patterns, each complete flower pattern was positioned 600 µm apart from the next. Plating cells on these micro-patterns led the authors to pivotal results regarding the decoupling of patterning with the effects of diffusive signalling. If cells were to secrete Brachyury inhibitors at a constant rate, it would be expected that there would be a higher concentration of inhibitors at the centre of the flower rather than the periphery, as the cells are closest to each other at the centre. Also, as a result of this, it would be expected that a low number of Brachyury positive cells would be observed at the centre. However, this was not the case. The patterning of Brachyury positive cells was unaffected compared to regular ellipse micro-patterns. Therefore, diffusive signalling does not occur between micro-patterns 600 µm apart, although it is possible that signalling occurs on a much smaller scale, or across all of the micro-patterns such that it is near homogeneous.

These results argue against what much other research assumes, which is a gradient of inhibitors emerging over time across the whole colony, which in turn is responsible for patterning. Rejecting culture-wide reaction-diffusion of chemical concentrations as the primary mechanism of stem cell patterning has vast implications, but does not mean that a cellular interaction model is incompatible with a reaction-diffusion model. Rather, some combination of the two may be necessary to mimic the events of early embryogenesis. For a vertex model where nodes each possess an internal dynamical system, such as proposed in other works [125], the assumption that these dynamics are able to be influenced by neighbouring cell dynamics means that the diffusive behaviour of chemical concentrations may be captured.

Notably, the authors do not present any mathematical model(s) in their paper to capture their results. In particular, there appears to be a strong focus on how local cell density affects patterning whilst straying from discussion of the geometry and symmetry of the cultures, despite the title and abstract of the paper stressing a focus on this. The development of a cell-to-cell interaction model for cells plated on the micro-patterns is likely to help elucidate the mechanisms of patterning beyond reaction-diffusion. Furthermore, as the researchers use mESCs, a repeat of their experiments using hESCs is necessary to test if the results differ. Although the same signalling pathways are present during gastrulation in mice and humans, the response to these pathways can differ vastly [126, 127]. Useful further work may be to determine whether cells arrange themselves prior to fixing and staining, such that gene expression patterning comes before spatial organisation, or whether the cells' position in the colony dictates their gene expression levels. Cells are known to move around during the early stages of lineage specification, so time-lapse imaging of the cells to observe the emergence of Brachyury expression is needed to determine a cause and effect relationship between patterning and position within the culture [128].

Mathematically Capturing the Symmetry of Cell Networks

The behaviour of a group of conceptual cells, where each cell represents a dynamical system, is explored with respect to the symmetry of the cell network [125]. Cells that are coupled share an edge, and so the authors define a symmetry of the network as "a permutation of the cells that preserves all internal dynamics and all couplings". The dynamical systems represented by each cell are ODEs, which take into consideration the effects upon a cell that might be induced by its coupled cell(s). The whole system is referred to as a 'coupled cell network'.

Patterns of synchrony may arise in a coupled cell network, where synchrony can be thought of as the action of a function. Let x_1 and x_2 be time-dependent variables with coupled dynamics. Then a synchronisation function f is such that, for the steady states $\bar{x_1}$ and $\bar{x_2}$, $f(\bar{x_1}) = \bar{x_2}$, f is continuous at $\bar{x_1}$, and f is consistent with the local dynamics of x_1 and x_2 [129]. This means that close to the steady state $(\bar{x_1}, \bar{x_2})$, f describes the predictability of the dynamics of x_2 from x_1. If such a function exists for x_1 and x_2 at time t, then x_1 and x_2 are said to be synchronous at t.

For a coupled cell network G, where C denotes the set of all cells in G, the authors define the symmetry groupoid of G as

$$\beta_G = \bigcup_{c,d \in C} B(c, d).$$

Through this definition, more information can be included than through the symmetry group; now, the input cell of each cell is captured, where the input set of a cell consists of itself and all connected cells. By analysing β_G, it can be determined whether two cells c and d exhibit patterns of synchrony. Solutions to synchronous cell dynamics may arise through Hopf bifurcations, indicating bistability [129].

The work discussed was not intended for a specific application, but it is easy to imagine how the concepts discussed by the authors may be applied to cultures of stem cells. Each cell may be represented as a node, and if two cells are neighbours, then the two cells share an edge. The time-dependent internal dynamics of each cell describe the cell's state, which may be presented as a gene regulatory network. However, unlike conventional gene regulatory networks, these equations also include a term to account for how a cell's dynamics are influenced by signalling from neighbouring cells. Cells with the same internal system of ODEs share a symmetry relation; note that for this to happen, it is required that the cells are neighbours.

A vertex mesh model has been applied to epithelial cell biology to accurately represent cell morphology [130]. Authors use the Linux software Chaste to subject cell vertices to forces, resulting in movement of the cells. This modelling architecture presents a flexible approach to cell culture which reflects true biophysical conditions. Graph theory has also been applied successfully to cell biology [131], though graph models for such applications are sparse in the literature.

Finite Group Theory

A group G is a set of objects together with a binary operation that associates any ordered pair of element g, h in G and also gh, the result of the binary operation. G is also subject to the following conditions:

1. *Identity.* There exists a unique identity element e in G such that $eg = ge = g$ for all g in G.

2. *Inverse.* For each g in G, there exists some inverse element g^{-1} in G such that $gg^{-1} = g^{-1}g = e$.

3. *Associativity.* For all g, h, k in G, $(gh)k = g(hk)$.

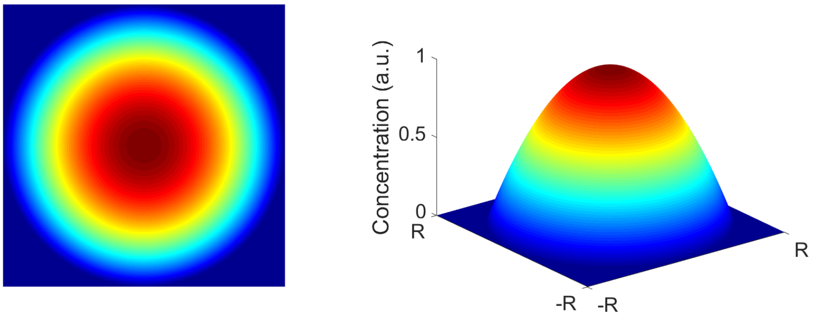
G is known as a finite group if the order of G is finite [132]. For a subset of G denoted by H, H is a subgroup of G if and only if the above three criteria are still satisfied.

In 1927, Brandt introduced 'groupoids'; a mathematical concept which has been heavily developed since [133, 134]. Unlike for groups, the identity element of a groupoid H need not be unique. It is only required that each element h in H has an identity element e_h such that $e_h h = he_h = h$.

The full symmetry group S_X of a set X is defined as the set of all of permutations σ of X such that $\sigma(x) = \sigma(y)$ if and only if $x = y$, and for every z in X, there exists an x in X such that $\sigma(x) = z$ [132]. It can be shown that S_X fulfills the group criteria, and represents the transformations for which the set X is invariant.

A partition of a set S is a collection of disjoint nonempty subsets of S that have S as their union. Letting I be an index set, the collection of subsets A_i for $i \in I$ forms a partition of S if and only if:

1. $A_i \neq \emptyset$ for $i \in I$.



(a) 2D.

(b) 3D.

Figure 3.2: Initial Noggin concentration profile $(t = 0)$, re-created from the reaction-diffusion model given by Tewary *et al.* (2017).

2. $A_i \cap A_j \neq \emptyset$ when $i \neq j$ for $i, j, \in I$.

3. $\bigcup\limits_{i \in I} A_i = S$.

A relation R on a set S is an *equivalence relation* if it is reflexive, symmetric, and transitive:

1. R is reflexive $\iff \forall \alpha \in S, \alpha R \alpha$.

2. R is symmetric $\iff \forall \alpha, \beta \in S$, if $\alpha R \beta$ then $\beta R \alpha$.

3. R is transitive $\iff \forall \alpha, \beta, \gamma \in S$, if $\alpha R \beta$ and $\beta R \gamma$ then $\alpha R \gamma$.

3.3 Recreating Tewary *et al.*'s Reaction-Diffusion Model of BMP4 and Noggin

For a complete understanding of the behaviour Tewary *et al.*'s reaction-diffusion model [30], the PDE system was simulated in MATLAB. In particular, as well as seeking to recreate the published results, this work aims to investigate the sensitivity to the model's initial conditions.

As shown in Table 3.1, the parameter γ_{BMP}, used to calculate a_{BMP} in the reaction-diffusion model given by Tewary *et al.* is not specified. For this work, the value $\gamma_{BMP} = \gamma_{NOG} = 0.0025$ was assumed. The initial concentration profile for Noggin is a parabolic ellipsoid, corresponding to Equation 3.4. The Noggin profile simulated for this work is shown in Figure 3.2.

The reaction-diffusion model was simulated in COMSOL, which used finite element analysis to solve the PDE system on a non-uniform 2D triangular mesh. However, a different approach may be taken to represent the system on a 2D square lattice. The advantage of this is that the results are easier to compare with those of well-documented lattice models that update the system state based on coupled cell-cell interactions. Such models will be explored in Section 3.4.

To discretise the continuous PDE system, finite difference approximations of first and second order differential equations were computed for this work. Finite difference approximations are numerical approximations of an equation, derived from the Taylor expansion [135]. For differential equations of greater than one order, approximations are computed either forward, centred, or backward. For this work, the centred finite difference was used, as this yields the most accurate approximation, i.e. that with the lowest error [136].

The finite difference approximation of the first order time derivative of a PDE is

$$\frac{\partial u}{\partial t} \approx \frac{1}{\Delta t}(u(t+\Delta t) - u(t)). \tag{3.7}$$

For a second order PDE on a 2D lattice with rows i and columns j, the centred finite difference approximation is

$$\frac{\partial^2 u}{\partial x \partial y} \approx \frac{1}{(\Delta x)^2}(u_{i-1,j} - 2u_{i,j} + u_{i+1,j}) + \frac{1}{(\Delta y)^2}(u_{i,j-1} - 2u_{i,j} + u_{i,j+1}) \tag{3.8}$$

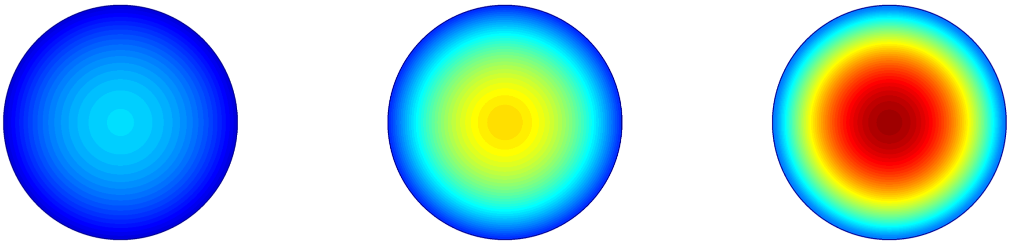
Letting $u = bmp^*$, $v = nog^*$, and (i,j) represent the spatial lattice sites corresponding to the reaction-diffusion model's (x,y) coordinates, then the centred finite difference approximation of Equations (3.1-3.2) are

$$\begin{aligned} u(t+\Delta t) = {} & u_{i,j}(t) + \Delta t((a_{BMP} - d_{BMP})u_{i,j}(t) + b_{BMP}v_{i,j}(t) + c^*_{BMP} \\ & + D_{BMP}\left(\frac{1}{(\Delta x)^2}(u_{i-1,j}(t) - 2u_{i,j}(t) + u_{i+1,j}(t))\right) \\ & + D_{BMP}\left(\frac{1}{(\Delta y)^2}(u_{i,j-1}(t) - 2u_{i,j}(t) + u_{i,j+1}(t))\right), \end{aligned} \tag{3.9}$$

$$\begin{aligned} v(t+\Delta t) = {} & v_{i,j}(t) + \Delta t((b_{NOG} - d_{NOG})v_{i,j}(t) + a_{NOG}u_{i,j}(t) + c^*_{NOG} \\ & + D_{NOG}\left(\frac{1}{(\Delta x)^2}(v_{i-1,j}(t) - 2v_{i,j}(t) + v_{i+1,j}(t))\right) \\ & + D_{NOG}\left(\frac{1}{(\Delta y)^2}(v_{i,j-1}(t) - 2v_{i,j}(t) + v_{i,j+1}(t))\right). \end{aligned} \tag{3.10}$$

A value of $\Delta t = 0.25$ and an initial concentration of $u = 0$ were used for simulations of this system with maximal time T, which are shown in Figure 3.3.

These results accurately reflect the published results from Tewary *et al.* when the BMP4 concentration is homogeneously initialised as 0.



(a) $t = T/3$. **(b)** $t = 2T/3$. **(c)** $t = T$.

Figure 3.3: Simulated concentrations of BMP4 at discrete time points t ($t = 0, ..., T$). The initial concentration of BMP4 = 0, and the disc diameter = 1000 μm. A centred finite difference scheme was implemented to approximate solutions to the reaction-diffusion system for BMP4 and Noggin. Parameters are given in Table 3.1, with $\gamma_{BMP} = 0.0025$.

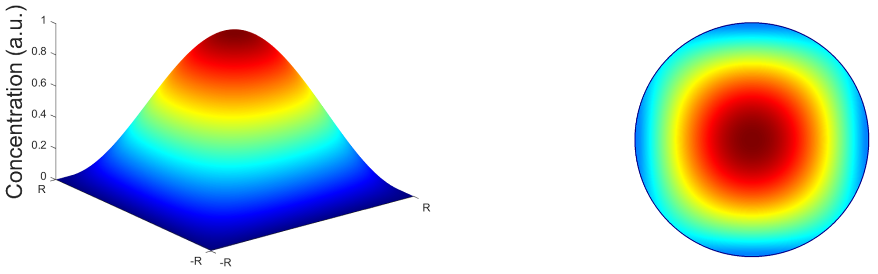
3.3.1 Altering the Initial Noggin Concentration Profile

Consider instead that the initial Noggin profile is diffusive, rather than a parabolic ellipsoid. The 2D parabolic diffusion equation is defined as

$$\frac{\partial u}{\partial t} = D\nabla^2 u, \tag{3.11}$$

which is used as the diffusive component in Equations(3.1-3.2). The parameter D_{NOG}, i.e. the diffusivity constant of Noggin, is given in Table 3.1. The parabolic diffusion equation was simulated using MATLAB code from Shankar [137]. A total of 20 discrete time steps were used when simulating 2D diffusion to yield the initial Noggin profile. This choice was arbitrary, as there is a lack of experimental data to indicate a suitable choice. As this serves to illustrate the difference in results when initialising Noggin with a diffusive profile compared to a parabolic ellipsoid, the exact diffusive profile used is not of high importance.

Figure 3.4 shows the related initial profile of Noggin, and the resultant final BMP4 concentration when the reaction-diffusion system was simulated with this initialisation. It is clear that changing the initial Noggin profile assumed by Tewary *et al.* - which lacks experimental evidence - to a diffusion profile which may also be biologically suitable affects the final BMP4 concentration profile. The symmetric concentric ring patterning is no longer seen, as the Noggin profile deviates from a ring as it reaches the disc boundary. demonstrates the sensitivity of the reaction-diffusion system to its initial conditions. Note that an entirely different initial Noggin profile may actually be present, which may disrupt the results to an even higher degree.



(a) 3D initial diffusive Noggin concentration profile ($t = 0$).

(b) 2D final BMP4 concentration profile ($t = T$).

Figure 3.4: Simulations of Noggin and BMP4 concentrations, where Noggin is initialised with a diffusive profile. Parameters for simulating the reaction-diffusion system leading to the final BMP4 profile are given in Table 3.1, with $\gamma_{BMP} = 0.0025$.

3.4 Development and Simulation of Coupled Map Lattice Models for Local Cell-Cell Interactions

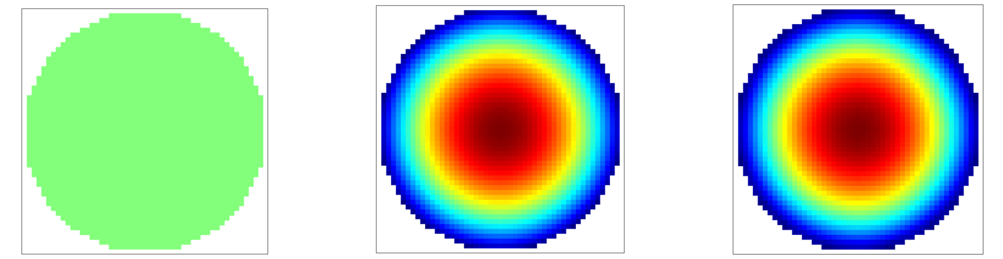
A model architecture that uses cell-cell interactions to drive system behaviour is the coupled map lattice (CML) model. The CML model was introduced by Kaneko [138] and others [139, 140] in the 1980s, but was most famously disseminated in a book written by Kaneko in 1991 [141]. Such models are designed to reflect non-linear system behaviour such as reaction-diffusion using discrete methods. Whilst traditional reaction-diffusion models operate in continuous time, the time component of CML models is discretised, and they are modelled on a discretised spatial domain (lattice) akin to cellular automata models [142]. They are particularly known for their ability to capture deterministic spatio-temporal chaos. To arrive at the CML model architecture, finite difference approximations of first and second order differential equations are computed. Currently, there are no published applications of CML models to explain hPSC fate patterning on micro-patterned cultures.

Letting $x_i(t)$ be a state variable for discrete time $t = 0, 1, 2, ...$ over a 1D lattice with sites $i = 1, 2, ..., N$, the CML model is given as:

$$x_i(t+1) = (1-\epsilon)f(x_i(t)) + \frac{\epsilon}{2}(f(x_{i+1}(t)) + f(x_{i-1}(t))), \tag{3.12}$$

which combines the discrete Laplacian operator for diffusion with a function $f(x)$ that is chosen to mimic the spatio-temporal dynamics of the system being modelled [141]. Considering instead a 2D lattice with sites $i, j = 1, 2, ..., N$, the CML model is given as

$$\begin{aligned} &x_{i,j}(t+1) = \\ &(1-\epsilon)f(x_{i,j}(t)) + \frac{\epsilon}{4}(f(x_{i+1,j}(t)) + f(x_{i-1,j}(t) + f(x_{i,j-1}(t) + f(x_{i,j+1}(t))). \end{aligned} \tag{3.13}$$



(a) Initial BMP4 concentration homogeneously defined as 0 ($t = 0$).

(b) Initial Noggin concentration as a parabolic ellipsoid ($t = 0$).

(c) Resultant BMP4 concentration after simulating the reaction-diffusion CML model ($t = 1$).

Figure 3.5: Initial and resultant concentration profiles when simulating the reaction-diffusion CML model with initial conditions given by Tewary *et al.* (2017). Parameters for reaction-diffusion are given in Table 3.1, additionally with $\epsilon = 0.1$.

How far each lattice site can 'see' surrounding sites, i.e. the distance at which any site is influenced by others, is controlled by the parameter ϵ. In particular, a value of $\epsilon = 0$ corresponds to a completely isolated site that is not impacted by any of its neighbours [143]. To investigate local dynamics, ϵ must remain sufficiently small, though what this means in a quantitative sense can vary per system.

3.4.1 Coupled Map Lattice Models for Reaction-Diffusion

Now, the reaction-diffusion system from Tewary *et al.* is revisited [30].

Letting $u = bmp^*$, $v = nog^*$, applying the finite difference discretisation and CML model architecture [143] to Equations (3.1-3.2) yields the coupled maps

$$f(u) = (1 + a_{BMP} - d_{BMP})u + b_{BMP}v + c^*_{BMP}, \tag{3.14}$$

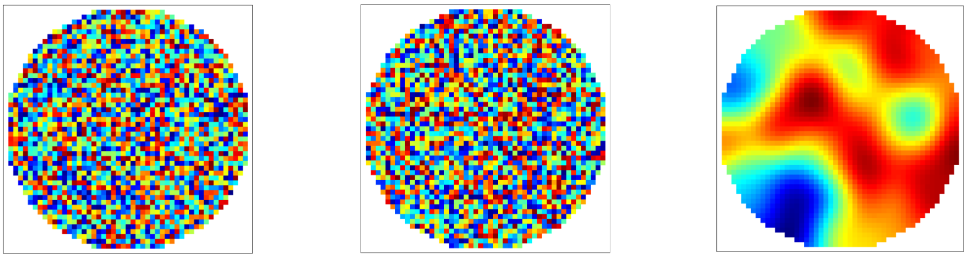
$$f(v) = a_{NOG}u + (1 + b_{NOG} - d_{NOG})v + c^*_{NOG}. \tag{3.15}$$

Taking these as the coupled maps corresponding to Equation 3.13, the reaction-diffusion system is simulated as a CML model.

For a homogeneous initial BMP4 concentration $u(0) = 0$, and the initial Noggin concentration $v(0)$ as the parabolic ellipsoid, ring patterning emerges for the BMP4 concentration after simulating the CML model for only one discrete time step. This is shown in Figure 3.5.

Clearly, this is not biologically representative, as gene expression profiles in hPSCs differentiated on a disc emerge transiently, moving through various states before reaching the final concentric ring patterning, rather than this being reached immediately.

To further a point made in Section 3.3, the initial Noggin profile may not display any



(a) Random initial BMP4 concentration ($t = 0$).

(b) Random initial Noggin concentration ($t = 0$).

(c) Resultant BMP4 concentration after simulating the reaction-diffusion CML model ($t = T$).

Figure 3.6: Initial and resultant concentration profiles when simulating the reaction-diffusion CML model with random initial conditions. Parameters for reaction-diffusion are given in Table 3.1, additionally with $\epsilon = 0.1$.

symmetries - previous simulations initialised Noggin as either a parabolic ellipsoid or a diffusion profile, which each possess symmetries. Considering a fully randomised distribution for BMP4 and Noggin, where random numbers are generated from a Gaussian distribution in the range $(0, 1)$, a highly robust model would be able to generate ring patterning. However, when such a system is simulated, a traditional reaction-diffusion profile is output, rather than a symmetric gradient. In other words, destroying symmetries of the model's initial conditions also destroys symmetries of the output. These results are shown in Figure 3.6.

As has been widely discussed, the reaction-diffusion model proposed by Tewary *et al.* presents initial conditions that are challenging to biologically justify. Patterns produced by CML models, like traditional reaction-diffusion, are highly sensitive to parameter values. Thus, despite robust cell-cell interaction-driven behaviour, CML models lack robustness in their parameter sensitivity. Ideally, a model should be able to generate the symmetric patterning seen *in vitro* without relying on these symmetries being present in the model's initial conditions.

3.5 A Novel Coupled Graph Formulation for Describing Cell Symmetry at Single-Cell Level

Whilst CML models show success in simulating reaction-diffusion processes, with 'sensing' distances for all lattice sites able to be controlled through model parameters, cells are necessarily confined to the rigid structure of the lattice. Vertex models allow more spatial flexibility than models relying on a defined grid, although they may also be simulated on a grid [130].

3.5.1 Defining the Coupled Graph

The following two definitions are borrowed from [125]. For an undirected graph, the coupled graph $G = (C, E, \backsim_C, \backsim_E)$ consists of:

1. A finite set $C = \{1, ..., N\}$ of nodes or cells.

2. A finite set of pairs $E \subseteq C \times C$ of edges. As G is undirected, if $\exists$ an edge (c, d) between cells c and d, then $(c, d) = (d, c)$.

3. An equivalence relation $\backsim_C$ on cells in C. The type of cell c is the $\backsim_C$-equivalence class $[c]_C$ of c.

4. An equivalence relation $\backsim_E$ on edges in E. The type of edge e is the $\backsim_E$-equivalence class $[e]_E$ of e.

Define also the input set for cell c as

$$I(c) = \{i \in C | (i, c) \in E\}. \tag{3.16}$$

Then $I(c)$ is the set of cells connected to c by an edge. It is assumed that there exists an internal edge for all cells $c \in C$ to themselves, such that $(c, c) \in E$. Then cell c is included in its own input set, i.e. $c \in I(c)$.

Now, these ideas from [125] are developed in this work, described as follows. The equivalence classes $[c]_C$ and $[e]_E$ are determined through an iterative procedure; transient equivalence relations on cells and edges will be assigned during this process. Transient cell/edge relations are defined by the following:

1. Let $\backsim_{C^*}$ be the transient equivalence relation on cells in C. The transient type of cell c is the $\backsim_{C^*}$-equivalence class $[c]_{C^*}$ of c.

2. Let $\backsim_{E^*}$ be the transient equivalence relation on edges in E. The transient type of edge e is the $\backsim_{E^*}$-equivalence class $[e]_{E^*}$ of e.

Determining Equivalence Relations Between Cells and Edges

The iterative procedure to find $[c]_C$ and $[e]_E$ is as follows:

Step 1: For each cell $c \in C$, assign transient cell type $[c]_{C^*} = |I(c)|$. These transient cell types form the $\backsim_{C^*}$-equivalence classes.

Step 2: For edges $(i, c), (j, d) \in E, (i, c) \backsim_{E^*} (j, d) \iff i, c$ can be ordered[1]such that $i \backsim_{C^*} j$ and $c \backsim_{C^*} d$. These transient edge types form the $\backsim_{E^*}$-equivalence classes.

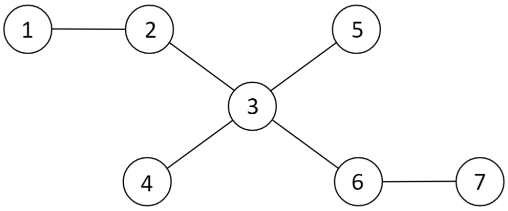
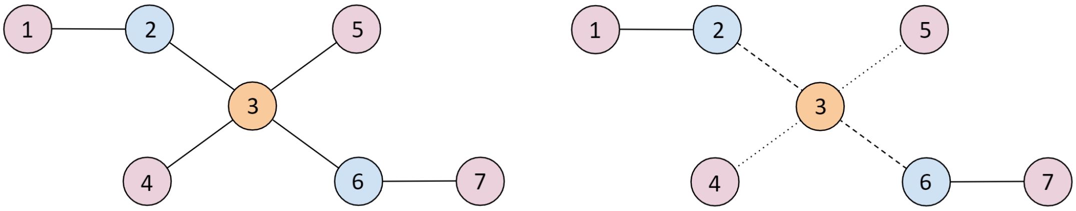


Figure 3.7: An example undirected 7-cell network.



(a) Cells are coloured by their transient cell type.

(b) Edges are marked by their transient edge type.

Figure 3.8: An example undirected 7-cell network. The transient cell type is given by the degree of each cell, and the transient edge type is given by the transient cell types joined by each edge.

Step 3: Cells $c, d \in C$ are assigned to the same $\backsim_C$-equivalence class $\iff \exists$ a base-cell preserving bijection $\beta : I(c) \to I(d)$ (meaning that $\beta(c) = d$) such that $\forall\ i \in I(c), (i, c) \backsim_{E^*} (\beta(i), d)$.

Step 4: If $\backsim_C = \backsim_{C^*}$ (up to reordering), then also assign $\backsim_E = \backsim_{E^*}$ and stop. If not, then assign $\backsim_{C^*} = \backsim_C$ and return to Step 2.

Example for Equivalence Class Computation

Consider the example undirected 7-cell network shown in Figure 3.7. A demonstration of how to use the iterative procedure described to determine the equivalence relations $\backsim_C$ and $\backsim_E$ is presented below.

Step 1: First, assign the transient cell equivalence $\backsim_{C^*}$ as the degree of each cell. Cells $1, 4, 5, 7$ have degree 1, cells $2, 6$ have degree 2, and cell 3 has degree 4. This means that the $\backsim_{C^*}$-equivalence classes are

$$\{1, 4, 5, 7\}, \{2, 6\}, \{3\}.$$

This partitioning of the cells into $\backsim_{C^*}$-equivalence classes is shown in Figure 3.8(a).

[1] Note that in Step 2 of the iterative process, choosing the order of i, c when determining transient edge equivalence arises from the fact that G is undirected, so the edge $(i, c) = (c, i)$. This means that i and c are interchangeable, but choosing the order of i, c may impact whether $i \backsim_{C^*} j$ and $c \backsim_{C^*} d$.

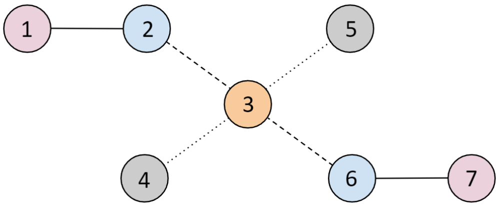


Figure 3.9: An example undirected 7-cell network. Cells are coloured according to their cell type and edges are marked according to their edge type.

Step 2: Assign the transient edge equivalence $\backsim_{E^*}$. Edges share transient equivalence if the edges connect the same transient type of cells. By examining Figure 3.8(a), we see that there are 3 transient edge types in the network: (*pink, blue*), (*blue, orange*), and (*pink, orange*). Recall that as the network is undirected, the ordering of cells - and therefore of (transient) cell types - within each edge pair is interchangeable. Then the $\backsim_{E^*}$-equivalence classes are

$$\{(1,2),(6,7)\},\{(2,3),(3,6)\},\{(3,4),(3,5)\}.$$

These $\backsim_{E^*}$-equivalence classes are shown in Figure 3.8(b).

Step 3: The next step is to assign $\backsim_C$-equivalence. For cells $c, d \in C$ to be $\backsim_C$-equivalent, there must exist a base-cell preserving bijection $\beta : I(c) \to I(d)$ (meaning that $\beta(c) = d$) such that $\forall\ i \in I(c), (i,c) \backsim_{E^*} (\beta(i), d)$. Consider cells $1, 5$ in Figure 3.8. Both cells are coloured *pink*, representing that $1 \backsim_{C^*} 5$. This means that $|I(c)| = |I(d)|$, so a β may be sought which maps the input set of cell 1 to the input set of cell 5, with $\beta(1) = 5$. Now, $I(1) = \{1,2\}$, and $I(5) = \{5,3\}$. Thus, the only base-cell preserving bijection existing between these two sets is such that $\beta(1) = 5, \beta(2) = 3$. However, it is required that $\forall\ i \in I(c), (i,c) \backsim_{E^*} (\beta(i), d)$. By looking at Figure 3.8(b), clearly the edge $(1,2)$ has transient edge type *solid*, i.e. the edge is represented by a solid line. The edge $(3,5)$ has transient edge type *dotted≠solid*, so $(1,2)$ is not $\backsim_{E^*}$-equivalent to $(3,5)$. Therefore, cell 1 is not $\backsim_C$-equivalent to cell 5. By repeating this process for all cells, the $\backsim_C$-equivalence classes are determined to be

$$\{1,7\},\{2,6\},\{3\},\{4,5\}.$$

This assignment of cell type is shown in Figure 3.9.

Step 4: Compare the cell equivalence relation $\backsim_C$ to the transient cell equivalence relation $\backsim_{C^*}$. It is found that $\backsim_C \neq \backsim_{C^*}$, as the cells within the network are partitioned differently as a result of Step 1 and Step 3 respectively. This means that $\backsim_{C^*} = \backsim_C$

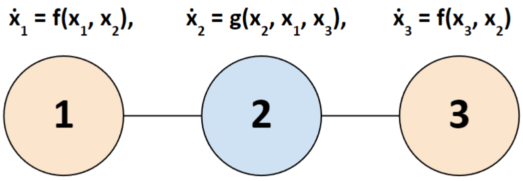


Figure 3.10: Example of symmetric dynamical functions for symmetric cells in a network. Here, the network is symmetric with respect to the permutation $\beta(1, 2, 3) = (3, 2, 1)$.

must be assigned and return to Step 2. In doing this, the edge equivalence relation $\backsim_{E^*}$ is seen to be unchanged from its former definition - this means that the edges within the network are partitioned into the same $\backsim_{E^*}$-equivalence classes in the second iteration of Step 2 as they were in the first iteration [2]. Moreover, after re-iterating Step 3, that the cell equivalence relation $\backsim_C = \backsim_{C^*}$ is found. Therefore, assign $\backsim_E = \backsim_{E^*}$ and stop the process. The equivalence classes $[c]_C$ and $[e]_E$ have now been found. Note that the diagram of the cell network continues to be represented by Figure 3.9.

3.6 A New Dynamical Model of Morphogen Concentration

For a stem cell population, each cell may be represented as a node, and an edge may be assigned between neighbouring cells, i.e. cells physically contacting in culture. Thus, for this work, hPSC cultures are modelled as a cell network $G(V, E)$. Notably, this approach is also taken in Section 2.6.

In this section, a mathematical model was sought to describe the dynamics of morphogen concentration at each node (cell) of G. Crucially, symmetric cells (as determined in Section 3.5) share symmetric internal dynamical functions. This means that, for symmetric cells i, j, the dynamical functions $\dot{x_i}, \dot{x_j}$ are symmetric with respect to a permutation β. Necessarily, β must map cell i to cell j, i.e. $\beta(i) = j$, and similarly all symmetry relations between other cells must be preserved. An example using a simple three-cell network is given in Figure 3.10.

The novel dynamical model is given by a set of ordinary differential equations (ODEs), with one ODE per node describing the morphogen concentration at that cell. This may be referred to as 'cell state'. Concentration at cell i is denoted x_i, with dynamics given

[2] This finding is not accidental, but rather a necessary part of concluding the iterative process - if the correct $\backsim_C$ and $\backsim_E$ are determined (but not confirmed) by completing iterative round $(n-1)$, then the procedure will stop after completion of round n, and $\backsim_{E^*}$ at $(n-1)$ will be equal to $\backsim_{E^*}$ at n. While this fact is useful to note, it is a superfluous condition to check when determining equivalence relations.

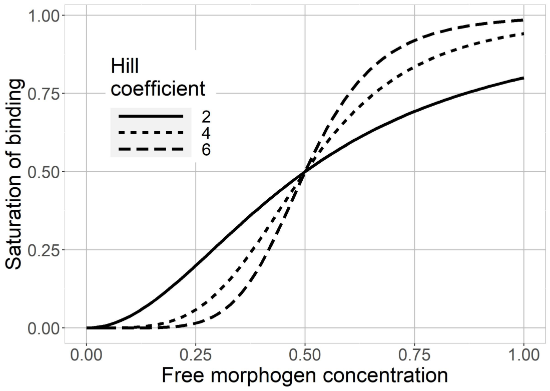


Figure 3.11: Example Hill functions with half-maximal saturation of binding constant $k = 0.5$ and Hill coefficient $n = 1, ..., 6$.

by

$$\frac{dx_i(t)}{dt} = \frac{\alpha x_i^n(t)}{k^n + x_i^n(t)} (1 - \epsilon(N_t)) - \frac{\alpha x_i(t)}{(k^n + 1)} + \alpha\epsilon(N_t) \sum_{j \neq i} \frac{x_j^n(t)}{k^n + x_j^n(t)},$$

$$\text{subject to } 1 \leq n \leq 6, \quad k \in [0, 1], \quad \alpha \in (0, 1), \tag{3.17}$$

where

$$\epsilon(N_t) = \begin{cases} 0 \text{ if } N_t = 1, \\ \frac{1}{N_t} \sum_{j \neq i} q^{(s_{i,j} - 1)}, \text{ with } q \in (0, 1), \text{ otherwise.} \end{cases} \tag{3.18}$$

Here, N_t is number of cells in the system at time t. The first term represents morphogen self-activation [144], which is assumed to take the form of a Hill function, which has Hill coefficient n, amplitude α, and half-maximal saturation constant k. Note that there is assumed zero basal concentration of the morphogen, as differentiating morphogens such as BMP4 must be externally administered to hPSC cultures. The coupling strength between cells i, j is $\epsilon(N_t)$, which incorporates the step distance $s_{i,j}$, i.e. the length of the shortest path between cells i, j. The parameter q represents signalling strength (as also defined for Equation 2.22) [83]. Coupling strengths are symmetric as $s_{i,j} = s_{j,i}$. Importantly, Equation 3.17 yields identical equilibria for cells within the same equivalence class.

Examples of Hill function profiles obtained by varying the Hill coefficient are given in Figure 3.11.

3.6.1 Characterising System Equilibria with Respect to Parameters

An equilibrium point of the ODE $dy/dt = f(y(t))$ is a solution $y^*(t)$ such that $dy^*/dt = 0$. Thus, at equilibrium, the system state is invariant with respect to time. An equilibrium point is locally stable if, when subject to small perturbations, the system returns to the equilibrium rather than diverges from it. The local stability of a non-linear dynamical system equilibria can be determined through analysis of the Jacobian matrix [145]. The Jacobian matrix is defined as

$$J = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

Thus, for the model given by Equations (3.17-3.18) with N cells, the Jacobian matrix becomes

$$J_{N_t} = \begin{bmatrix} \alpha\left(\frac{(1-\epsilon(N_t))nk^n x_1^{n-1}}{(k^n+x_1^n)^2} - \frac{1}{(k^n+1)}\right) & \cdots & \frac{\alpha\epsilon(N_t)nk^n x_{N_t}^{n-1}}{(k^n+x_{N_t}^n)^2} \\ \frac{\alpha\epsilon(N_t)nk^n x_1^{n-1}}{(k^n+x_1^n)^2} & \cdots & \frac{\alpha\epsilon(N_t)nk^n x_{N_t}^{n-1}}{(k^n+x_{N_t}^n)^2} \\ \vdots & \ddots & \vdots \\ \frac{\alpha\epsilon(N_t)nk^n x_1^{n-1}}{(k^n+x_1^n)^2} & \cdots & \alpha\left(\frac{(1-\epsilon(N_t))nk^n x_{N_t}^{n-1}}{(k^n+x_{N_t}^n)^2} - \frac{1}{(k^n+1)}\right) \end{bmatrix}.$$

Equilibria of the Single-Cell System

First considering a single-cell system (N_t=1), the model is given by the ODE

$$f_1 = \frac{dx_1}{dt} = \frac{\alpha x_1^n(t)}{k^n + x_1^n(t)} - \frac{\alpha x_1(t)}{(k^n+1)}. \tag{3.19}$$

The single-cell Jacobian matrix is then

$$J_1 = \left[\frac{\partial f_1}{\partial x_1}\right] = \left[\alpha\left(\frac{nk^n x_1^{n-1}}{(k^n+x_1^n)^2} - \frac{1}{(k^n+1)}\right)\right].$$

Let x_1^* denote an equilibrium of Equation 3.19. Then the equilibria of this model are found by solving $dx^*/dt = 0$, yielding the solutions:

1. $x_1^* = 0$ for $1 \leq n \leq 6$,

2. $x_1^* = 1$ for $1 \leq n \leq 6$,

3. $x_1^* = k^2$ for $n = 2, k > 0$ only,

4. $x_1^* = \frac{1}{2}\left(k^3 \pm \sqrt{(k^3(k^3+4))}\right)$ for $n = 3$ only.

By substituting $x_1^* = x_1$ into J_1, the Jacobian J_{1^*} is computed. Then, the equilibria of Equation 3.19 are asymptotically stable if $J_{1^*} < 0$, and unstable if $J_{1^*} > 0$ [146]. In case $J_{1^*} = 0$, then Equation 3.19 is stable if the function changes sign from positive to negative at the equilibrium [146]. Thus, for the four computed equilibria, it is found that

1. $x_1^* = 0$ is stable for $1 \leq n \leq 6$,
2. $x_1^* = 1$ is stable for $n = 1$, unstable for $2 \leq n \leq 6$,
3. $x_1^* = k^2$ is always stable for $n = 2, 0 < k < 1$, and stable for $n = 2, k = 1$ only if Equation 3.19 changes sign at $x_1^* = 1$.
4. $x_1^* = \frac{1}{2}\left(k^3 \pm \sqrt{(k^3(k^3+4))}\right)$ is stable for $n = 3$ only.

Equilibria of the Coupled Two-Cell System

Now considering a coupled two-cell system (N_t=2), the state model is

$$\frac{dx_1}{dt} = \frac{1}{2}\frac{\alpha x_1^n(t)}{k^n + x_1^n(t)} - \frac{\alpha x_1(t)}{(k^n+1)} + \frac{1}{2}\frac{\alpha x_2^n(t)}{k^n + x_2^n(t)}, \tag{3.20}$$

$$\frac{dx_2}{dt} = \frac{1}{2}\frac{\alpha x_2^n(t)}{k^n + x_2^n(t)} - \frac{\alpha x_2(t)}{(k^n+1)} + \frac{1}{2}\frac{\alpha x_1^n(t)}{k^n + x_1^n(t)}. \tag{3.21}$$

Note that as $s_{1,2} = s_{2,1} = 1$, then $\epsilon(2) = 1/2$ regardless of the value of q. The corresponding Jacobian matrix is therefore

$$J_2 = \begin{bmatrix} \alpha\left(\frac{nk^n x_1^{n-1}}{2(k^n+x_1^n)^2} - \frac{1}{(k^n+1)}\right) & \frac{\alpha nk^n x_2^{n-1}}{2(k^n+x_2^n)^2} \\ \frac{\alpha nk^n x_2^{n-1}}{2(k^n+x_2^n)^2} & \alpha\left(\frac{nk^n x_1^{n-1}}{2(k^n+x_1^n)^2} - \frac{1}{(k^n+1)}\right) \end{bmatrix}.$$

Let x_1^*, x_2^* denote the equilibrium values of x_1, x_2. As the ODE system preserves cell symmetry relations, symmetric cells share the same equilibrium value(s). Thus, $x_1^* = x_2^* = x^*$.

By solving $dx^*/dt = 0$, three equilibria are found:

1. $x^* = 0$ for $2 \leq n \leq 6$,
2. $x^* = 1$ for $1 \leq n \leq 6, n \neq 2$,
3. $x^* = k^2$ for $n = 2$ only.

At equilibrium, the Jacobian matrix for the two-cell system is

$$J_2^* = \begin{bmatrix} \alpha\left(\frac{nk^n x^{*(n-1)}}{2(k^n+x^{*n})^2} - \frac{1}{(k^n+1)}\right) & \frac{\alpha nk^n x^{*(n-1)}}{2(k^n+x^{*n})^2} \\ \frac{\alpha nk^n x^{*(n-1)}}{2(k^n+x^{*n})^2} & \alpha\left(\frac{nk^n x^{*(n-1)}}{2(k^n+x^{*n})^2} - \frac{1}{(k^n+1)}\right) \end{bmatrix}.$$

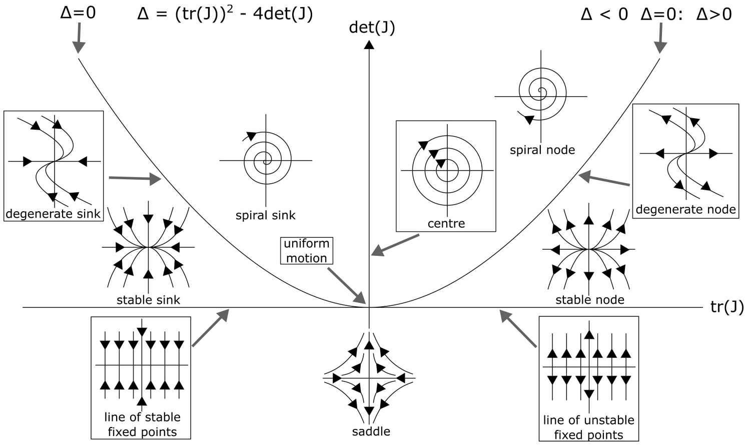


Figure 3.12: Poincarè diagram for classifying equilibria in the trace-determinant plane.

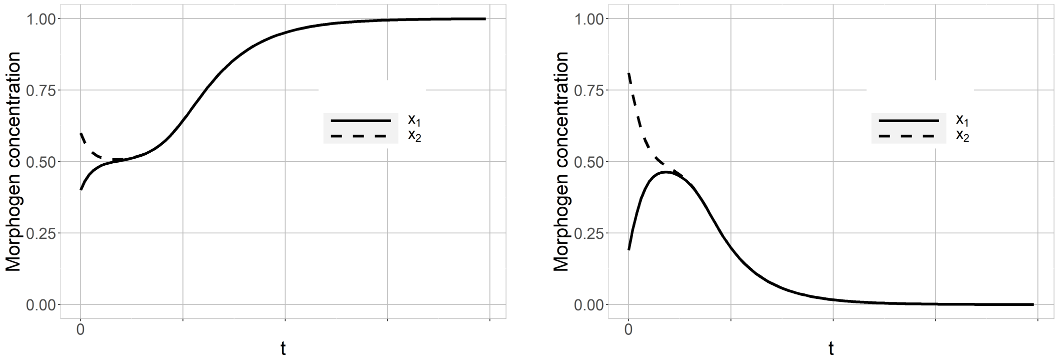
Analysis of the trace and determinant of the Jacobian matrix at equilibrium reveals the stability of the system equilibria [147, 148]. In particular, stability is achieved if and only if tr$(J) < 0$ and det$(J) > 0$. The parabolic equation

$$\Delta = (tr(J))^2 - 4det(J) \tag{3.22}$$

represents a separatrix in the trace-determinant plane such that equilibria can be characterised by where they are positioned with respect to the regions generated by Δ. The Poincarè diagram [149] shown in Figure 3.12 gives a visual representation of the trace-determinant plane, with the associated equilibrium classifications for values within different regions. In this case, using MATLAB's Symbolic Toolbox, the stability of equilibria was found to be as follows:

1. $x^* = 0$: stable degenerate node,
2. $x^* = 1$: stable node for $\frac{nk^n}{(k^n+1)} > 2$, saddle point otherwise. This condition is met for all values $k \in (0,1), 1 \leq n \leq 4$. The condition is also met for $n = 5$, $k \leq 0.9219$, and $n = 6$, $k \leq 0.8909$.
3. $x^* = k^2$: may take the form of a saddle point, stable (degenerate) node, or stable spiral depending on the choice of k for $1 \leq n \leq 6$, $n \neq 2$. For $n = 2$, this equilibrium is always a saddle point.

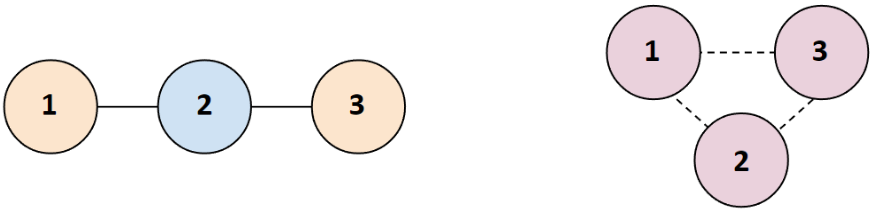
Note that saddle points are necessarily unstable. For this system, Figure 3.13 provides a demonstration of how two random initalisations of x_1, x_2 yield an equilibrium at either 1 or 0.



(a) Equilibrium at 1.

(b) Equilibrium at 0.

Figure 3.13: Time-dependent solutions of the ODE system for two cells. Random initialisation, constrained to sum to 1, results in the system settling at one of three equilibrium points (additional equilibrium at k^2). The equilibrium at 1 was found to be a stable node, and the equilibrium at 0 a stable degenerate node.



(a) $|E(G)| = 2$: $s_{2,1} = s_{2,3} = 1$, $s_{1,3} = 2$.

(b) $|E(G)| = 3$: $s_{1,2} = s_{1,3} = s_{2,3} = 1$.

Figure 3.14: Two possible topologies for G in a three-cell system. The number of edges in each case are not equal, demonstrating that the topology of G has more than one possibility for systems with greater than two cells.

Numerically Dependent Equilibria of the General System

For values $N_t > 2$, equilibria are challenging to compute for all possible parameter values. This is due to the potential differences in the topological structure of graph G for any fixed value of N_t.

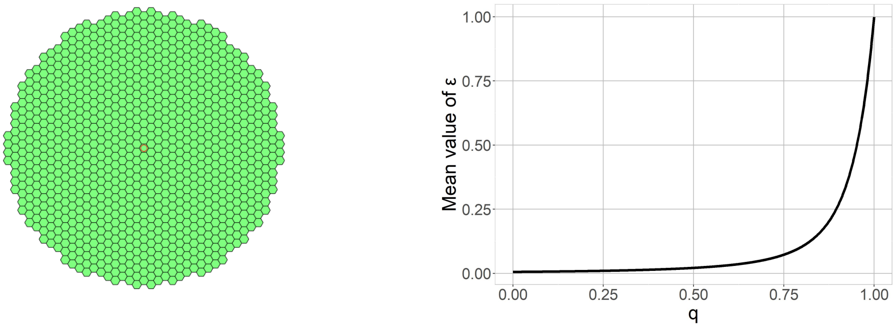
To illustrate this point with the simplest possible system, consider $N_t = 3$. Then $|V(G)| = 3$, but there are two possible values for $|E(G)|$: 2 or 3. This difference arises in the possibility that:

1. $(|E(G)| = 2)$: One cell has two neighbours - namely the other two cells in the system - which themselves each have only one neighbour.

2. $(|E(G)| = 3)$: All three cells are neighbours of every other cell in the system.

A visual representation of these two possible topologies for G when $N_t = 3$ is given by Figure 3.14.

In the first of these cases, step distances $s_{2,1} = s_{2,3} = 1$ and $s_{1,3} = 2$. However, in the



(a) $N_t = 1003$: simulated cells on a disc.

(b) Mean ϵ as q is varied for cells shown in (a).

Figure 3.15: (a) A demonstration of an arbitrary number of simulated hexagonal cells arranged to mimic a disc micro-pattern, (b) From Equation 3.18: plot showing the mean cell-to-cell coupling strength ϵ in the simulated culture given by (a) as the parameter controlling signalling strength $q \in (0, 1)$ is varied.

second of these cases, $s_{1,2} = s_{1,3} = s_{2,3} = 1$. Therefore, the cell-to-cell coupling strength ϵ, given in Equation 3.18, is heavily dependent on the topology of G.

Coupling between cells is defined by Equation 3.18, which is heavily regulated by the parameter $q \in (0, 1)$. To investigate how the inter-cellular coupling term $\epsilon(N_t)$ scales with the value of q, a system of size $N_t = 1003$ was simulated on a disc pattern. The topology of this simulated graph can be seen in Figure 3.15(a). Using the step distances between each cell pair, values of ϵ were computed while the parameter q was varied in the range $(0, 1)$. The mean value of ϵ for each q was then computed and plotted, as shown in Figure 3.15(b).

The figure reveals an 'elbow' in the curve relating ϵ and q. The elbow occurs at $q \approx 0.85 = q^*$, such that ϵ remains low (< 0.2) for $q \leq q^*$, then rapidly increases for $q > q^*$, maximising at $\epsilon = 1$.

Note that as every x_i represents a biological quantity, negative and/or complex solutions are not feasible.

3.7 Numerical Simulations of Cell Fate Selection on Micro-Patterns

Dependency of the dynamical state of cells on the topology of the underpinning contact network was evaluated by simulating micro-patterned cell cultures in line with experimental literature [30, 31, 45, 47]. Custom computational grids were created to simulate cells on triangular, square, and hexagonal micro-pattern geometries. All grids contained sites defined as regular shapes with equal area and perimeter. Each cell may be initialised with any value in the range $(0, 1)$ without altering the qualitative results. For our simulations, a homogeneous initial concentration of $1/N$ at each cell was applied. To compute equilibria, a custom MATLAB function was developed to automatically define the ODE system based on the structure of the cell network input, and integrated using the *ode45* routine.

It was assumed that half-maximal saturation of morphogen concentration occurs at the midpoint of minimum and maximum concentration. Hence, the half-maximal saturation constant $k = 0.5$ was used for simulations. When simulating large systems, the computational cost is increased, so the model was simulated with parameter values that would not change the qualitative output, but imposed sharp gradients on the associated functions to force fast convergence to the equilibrium. Therefore, a Hill coefficient of $n = 6$ and a signalling strength of $q = 0.9$ were used for simulations. The amplitude $\alpha = 1$ of the Hill function was selected for simplicity. It is worth bearing in mind that the final patterning results output by the model are highly qualitatively similar for sensible parameter choices; the case of $q = 1$ yields a homogeneous equilibrium.

3.7.1 Model Equilibria Recapitulate Published Experimental Results

The novel dynamical model was applied to simulated cells on a disc - generated using a hexagonal grid - to see whether the model was able to recapitulate experimental ring patterning. The equilibrium result is shown in Figure 3.16(a), with the corresponding immunofluorescence microscopy image from Tewary *et al.* (2017) in Figure 3.16(b). From this, it is seen that the model output matches very closely the experimental patterning result.

As the Brachyury gene in hPSCs has been found to localise at the tips of an equilateral triangle micro-pattern, an equilateral triangle was simulated, using a regular triangular lattice. Simulated cells were regarded as neighbours if any point of their respective perimeters were in contact - this includes cell contacts occurring only at the vertex of a triangle. The equilibrium result of this simulation is shown in Figure 3.16(c), with Figure 3.16(d) showing the immunofluorescence microscopy image from Muncie *et al.* (2020). Again, the model output is well-aligned with the associated microscopy image.

In addition to the equilateral triangle, Muncie *et al.* found that differentiating hPSCs on a trapezium micro-pattern - by removing the upper part of an equilateral triangle - de-localises Brachyury at the trapezium's apex. Figure 3.16(e) shows the simulated equilibria on a trapezium, which is in accordance with the behaviour of Brachyury described.

For differentiated mESCs, Brachyury has been shown to localise at micro-patterned ellipse tips. The model equilibria on a simulated ellipse, together with a hexagonally binned density map showing Brachyury localisation from Blin *et al.* (2018), are given in Figures 3.16(f)-(g), showing the model's success in recapitulating this result.

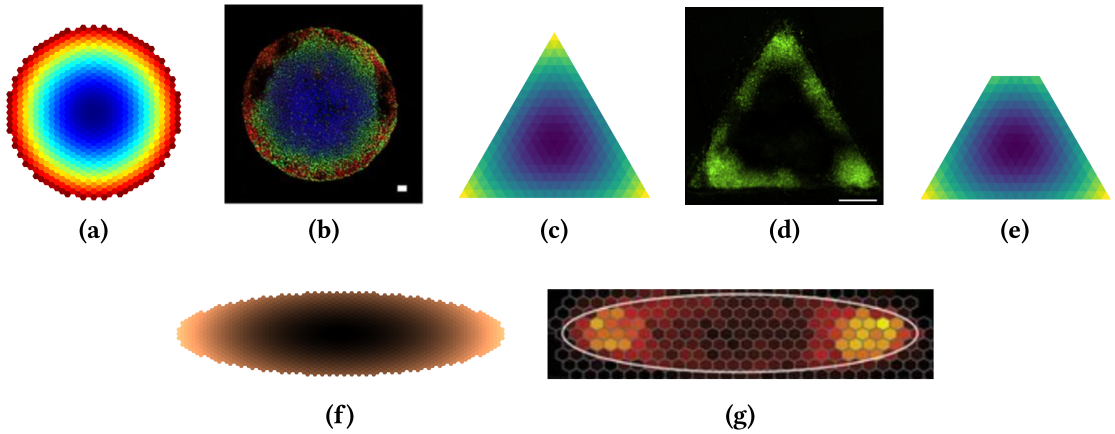


Figure 3.16: (a) Equilibrium simulation on a disc; **(b)** Figure 1(a) from Tewary *et al.* (2017): 'Representative immunofluorescence images of fate patterning of SOX2, BRA and CDX2 in BMP4-supplemented CM. Scale bars: 50μm'; **(c)** Equilibrium simulation on an equilateral triangle; **(d)** Figure 3(d) from Muncie *et al.* (2020): 'Representative brightfield [...] images of T-mNeonGreen expression for triangle hESC colonies on compliant gels following BMP4 addition. All scale bars, 250μm'; **(e)** Equilibrium simulation on a trapezium; **(f)** Equilibrium simulation on an ellipse; **(g)** Figure 2(a) from Blin *et al.* (2018): 'BDMs of the T- or T+ populations', aggregated across 4125 cells and 65 colonies. All simulations show equilibria of the model given by Equations (3.17-3.18) with parameters $k = 0.5$, $n = 6$, $\alpha = 1$, $q = 0.9$ on lattices.

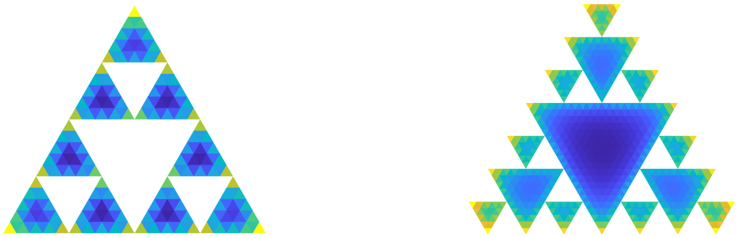
3.7.2 Could Fractal Micro-Patterns Yield Self-Similar Cell Fate Patterning?

A natural development of reasoning when dealing with symmetric systems is how to produce self-similar patterns. One way of recursively producing self-similar shapes is to develop fractal micro-patterned surfaces.

To investigate whether simulating cells on a fractal micro-pattern would yield self-similar cell fates, a Sierpinski triangle pattern was created. The Sierpinski triangle was mathematically formalised in 1915, having zero area and infinite boundary [150]. Here, cells are simulated on this geometry, again assigning neighbour relationships between triangles that come into contact at any point on their respective perimeters, including meetings occurring only at any of the three triangle vertices. The equilibrium results of these simulations are given in Figure 3.17.

3.7.3 Representing Cells in the Network as Voronoi Objects

The previous subsections modelled the dynamical system given by Equations (3.17-3.18) at node sites given as regular, homogeneous shapes. For the following model simulations, the Voronoi tessellation was used for representing each cell as an object with non-homogeneous morphology [151, 152]. This approach more closely mimics the irregularity of hPSC morphology in culture. The Voronoi tessellation of a plane, which is



(a) Sierpinski triangle ($N_t = 225$).

(b) Inverted Sierpinski triangle ($N_t = 925$).

Figure 3.17: Equilibrium simulations of the model given by Equations (3.17-3.18) on Sierpinski triangles, with model parameters: $k = 0.5, n = 6, \alpha = 1, q = 0.9$.

---

composed of individual Voronoi objects, is computed in the following manner [153]. Let a metric space $\boldsymbol{X}$ with distance function d. Let K be a set of indices and let $(P_k)_{k \in K}$ be a tuple - i.e. an indexed collection - of non-empty subsets in $\boldsymbol{X}$. Then the Voronoi object R_k associated with the site P_k is the set

$$\{P_i \in \boldsymbol{X} | d(P_i, P_k) \leq d(P_i, P_j) \text{ for any } i, j \neq k\}.$$

Then the Voronoi tessellation is the tuple of objects $(R_k)_{k \in K}$. For this work, $\boldsymbol{X}$ is the (x, y) plane, and d is the Euclidean distance.

Bounded Voronoi tessellations were generated using a computational algorithm by Sievers [154]. Boundaries for the tessellation were defined as either a disc or an equilateral triangle. Equilibrium simulations of these geometries are shown in Figures 3.18(a)-(b). From these figures, it is clear that the fate patterning demonstrated using lattices is preserved in the Voronoi model.

An interesting experimental finding was that ring fate patterning on discs was disrupted when disc diameter was reduced [30]. The most distinct patterning arose using discs of $1000\,\mu\text{m}$ diameter, but for discs of smaller diameter - and therefore smaller cell populations - some cell fates were not clearly emergent. This was particularly seen for discs of $500\,\mu\text{m}$ diameter or less, for which low expression levels of the pSMAD1 gene were not obtained.

To test whether the model given by Equations (3.17-3.18) reflects this phenomenon, the number of simulated cells was proportionally reduced such that the diameter of the disc was halved. As the population simulated in Figure 3.18(a) consisted of 1000 cells, the disc area was recorded as 1000 arbitrary square units. The corresponding disc radius r, and the number of cells N_i with which to initialise the reduced-sized disc simulation, were thus computed by the following:

$$1000 = \pi r^2 \implies r = \sqrt{\frac{1000}{\pi}} \implies N_i = \pi \left(\frac{r}{2}\right)^2 = \pi \left(\sqrt{\frac{1000}{\pi}} \bigg/ 2\right)^2 = 250. \quad (3.23)$$

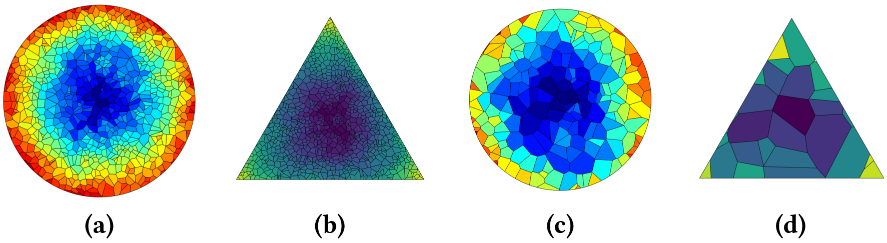


Figure 3.18: (a) Equilibrium simulation on a densely populated disc ($N = 1000$); **(b)** Equilibrium simulation on a densely populated triangle ($N = 930$); **(c)** Equilibrium simulation on a sparsely populated disc ($N = 207$); **(d)** Equilibrium simulation on a sparsely populated triangle ($N = 50$). All simulations show equilibria of the model given by Equations (3.17-3.18) with parameters $k = 0.5$, $n = 6$, $\alpha = 1$, $q = 0.9$, using Voronoi objects to represent cells.

To model this, 250 simulated points were initially distributed in a square domain. Applying the Voronoi tessellation algorithm - which involves trimming the square domain to the desired shape (here, a disc) - reduced the cell count to 207. The system equilibrium for simulated cells on this reduced-size disc is shown in Figure 3.18(c). As can be seen, the clear ring patterning demonstrated in Figure 3.18(a) is not emergent. Additionally, Figure 3.18(d) demonstrates a lack of distinct localisation of morphogen to the triangle tips for a simulation on a reduced-size triangular domain.

3.8 Applying the Dynamical Model to a Novel Growing Cell Network

Notably, fixed sized systems don't account for cell division and apoptosis. Such phenomena are beneficial to model for increased biophysical relevance. Moreover, any form of static model cannot capture the gene expression trajectory of single cells over time. Due to a combination of cell age and epigenetic memory, former and current stem cell states are highly related, which has been mathematically reflected [155]. In dynamical systems, bistability forms the basis for encoding memory into an agent (cell), thus driving pattern formation [156]. Therefore, the development of a model which combines the flexibility of a vertex model with the time-series nature of a growing model would be highly relevant for hPSC fate pattern formation.

This section aims to apply the model given by Equations (3.17-3.18) to describe emergent fate patterning for growing hPSC cultures. Previous mathematical models of hPSC fate link gene expression levels to the concentration of the differentiating morphogen at a spatiotemporal point [30]. In the growing model, therefore, cell state may be regarded as cellular expression of a particular genetic marker.

3.8.1 Novel Growing Model Formulation

The growing vertex model is initialised with one cell. At each discrete timestep $t \in [0, T]$, N_t represents the number of populated cells in the system. For cell i, cell state $x_i(t)$ is subject to stochastic fluctuations while remaining bounded in the interval $(0, 1)$, i.e.

$$x_i(t) = x_i(t-1) + \xi(i)_t, \quad t \geq 1, \quad x_i(t) \in (0, 1) \tag{3.24}$$

for stochastic noise $\xi(i)_t$ taken from a Gaussian distribution with zero mean.
A populated cell has a probability of division

$$p_{\text{div}} = \sigma\tau, \tag{3.25}$$

where σ is a rate constant governing how quickly cells divide when they are in the proliferation phase, and τ is a fraction representing the duration of time in which cells proliferate over the total cell cycle time. The fraction $\tau = 1/15.75$ was used, as hESCs spend 1h in the S phase of the cell cycle - during which they proliferate - and 14.75h as an average total in all other phases [157]. For cell i to divide, a randomly selected empty neighbour of cell i must be populated. Thus, if cell i has no empty neighbours, i.e. all cells in direct contact with cell i are populated, then cell i cannot divide. For biological cell cultures, neighbour-less cells are generally able to divide, as existing neighbours of the dividing cell are subject to mechanical forces and are displaced, allowing the daughter cell to form. However, it is a reasonable simplification to ignore this phenomena within the model, as the role of cell movement is incorporated through cellular Brownian motion.
If cell i divides at time t_k, then daughter cell j initially inherits the state value of its parent upon time of division, i.e.

$$x_j(t_k) = x_i(t_k). \tag{3.26}$$

Each populated cell also has an associated probability of apoptosis. For cell i, this is given by

$$p_{\text{apop}_i} = \left(\frac{1}{\gamma}\right)\left(1 - \frac{1}{|I(i)|}\right) \tag{3.27}$$

where $|I(i)|$ is the cardinality of the input set of cell i (as in Equation 3.16), and γ is a rate constant such that larger values of γ correspond to less frequent cell deaths. The reasoning for Equation 3.27 was that cells would be more likely to undergo apoptosis at high local density due to reduced availability of nutrients in the culture media. This is reflected in the equation by accounting for how many direct neighbours a cell has.
At each timestep t, all Voronoi cell vertices are subject to Brownian motion in the (x, y)

plane. Modelling the motility of hPSCs with Brownian motion has shown to be successful in capturing biological phenomena [158]. Brownian motion was modelled such that the updated $(\mathbf{x}, \mathbf{y})$ coordinate vector pair, denoted $(\mathbf{x}_{new}, \mathbf{y}_{new})$, are given by

$$(\mathbf{x}_{new}, \mathbf{y}_{new}) = \min(0, \max(0, (\mathbf{x}, \mathbf{y}) + \lambda_{\text{brown}}(\boldsymbol{\nu}_{x,y} - \sigma))), \tag{3.28}$$

where $\boldsymbol{\nu}$ is a random vector of length N_t with elements from the Gaussian distribution in the range $(0, 1)$, and λ_{brown} is the Brownian motility constant.

Values $\sigma = 10, \gamma = 4$, and $\lambda_{\text{brown}} = 0.2$ were applied. The selection of $\lambda_{\text{brown}} = 0.2$ was reasoned by measured diffusion coefficients for wildtype hESCs, which are around 0.2 [158], and small-scale Brownian motion of many particles yields diffusive processes [159]. Populations of hESCs are known to typically exhibit high rates of spontaneous proliferation and apoptosis [160], so the parameter values $\sigma = 10, \gamma = 4$ were selected to reflect this.

When the micro-pattern is fully populated, administration of the differentiating morphogen to the culture is simulated. This is modelled by applying the dynamical system given by Equations (3.17-3.18), using the random cell state vector as initial conditions. Cells at this stage are assumed to remain in Brownian motion without cell division or apoptosis to preserve the network topology. As before, the simulation ends when the system equilibrium is achieved.

3.9 Summary and Discussion

In this chapter, a novel vertex model with internal node dynamics was presented. This was applied to modelling hPSC fate at the single cell level, although it may be adapted for application to other systems. Accurate results were obtained when simulated on geometries mimicking micro-patterned hPSC cultures in the literature [45, 47, 31]. Neighbourhood interactions are accounted for through an ordinary differential equation (ODE) system describing nodular dynamics. In this case, nodular dynamics may represent the concentration of a differentiating morphogen such as BMP4 bound to cell surface receptors, or cellular expression level of a genetic marker. Akin to the reasoning behind Waddington's epigenetic landscape, the system state corresponding to fully acquired cell fate is a time-invariant equilibrium of the system [115].

Unlike many other models for hPSC fate patterning in the literature [50, 46, 30], the model presented here is simultaneously applicable to any 2D Euclidean space, able to reflect biological cell morphology, and incorporates established edge-sens-ing mechanisms of hPSC cultures [49]. The key tools implemented - such as defining the coupled graph, calculating inter-cellular distances, simulating time-depen-dent dynamics, generating Voronoi tessellations, and subjecting cells to Brownian motion - are all generalisable to 3D systems. Therefore, it is very possible that the model presented is able to be generalised to 3D cell models *in vitro*, but further work would need to be conducted to verify this.

Traditional models such as the coupled map lattice (CML) model and reaction-diffusion are unable to naturally generate ring patterning as is seen among differentiated hPSC cultures on disc micro-patterns. For these models, ring patterning is exclusively able to be output by defining the initial model condition as a radially symmetric gradient, i.e. the ring pattern is input rather than emergent. Conversely, the model given in this chapter is able to be initialised with any value in the range $(0, 1)$ whilst giving rise to not only ring patterning, but also patterning that matches experimental results involving micro-patterning of an equilateral triangle, trapezium, and ellipse [45, 47]. This is very promising; since the introduction of Wolpert's 'French flag' model in the 1960s, it has been well-known in mathematical biology that the spread of a morphogen is heavily linked with the positional information of cells that it acts upon [161]. However, existing models that accurately reproduce fate patterning do not do so with any kind of boundary control [46, 50]. The contraction-reaction-diffusion model for fate patterning in the literature is successful at generating target patterns, but is of high complexity due to the inclusion of mechanical tension [48]. The model presented in this chapter, on the other hand, is able to reproduce a range of experimentally observed patterns with a comparatively simple architecture.

Whilst the model offers a novel and useful paradigm for cell state modelling, it must be stressed that for the majority of cell populations simulated, the focus is on the qualitative - rather than quantitative - system equilibria. As further research, the growing vertex model would ideally be simulated with biologically informed gene expression data before initiating the simulation of morphogen administration. Then, using tracking constructs already present in the model code, the true single-cell evolution of gene expression could be recorded. In fact, this is exactly what has been initiated for the process described in Appendix A.1.5.

Continuing the above, as the model records changes in spatial organisation, including neighbour and parent/daughter relationships, together with cellular gene expression at each discrete time step, an information-rich output for performing analysis is provided. In particular, identifying the causal relationship between gene expression patterning and spatial position is of high interest. It has also been shown that rapidly altering morphogen concentration produces variations in hESC patterning [122]. This phenomenon could be investigated with reasonable ease through the proposed model.

Ultimately, the spatio-temporal distribution of a morphogen is highly likely to determined by a combination of (1) diffusion occurring across culture media, and (2) binding to cell surface TGF-β receptors [30, 49]. Although the model presented here does not explicitly incorporate diffusion, a great benefit of formulating the problem using a graph is that the diffusion equation may be easily applied [162]. Moreover, the application of spectral graph theory may facilitate quantification of structural similarity between different cell networks [163], which could be investigated with respect to differences in cell fate patterning. Finally, analysis of a cell network's adjacency matrix at each time step may reveal important diffusive behaviour across network edges [164].

It seems that overall, the model framework given in this chapter has proven to be accurate in qualitatively recapitulating several experimental results, thus providing a highly fruitful basis for time-dependent analysis of growing cell populations both prior to and during fate specification. There is a need to increase the biophysical relevance of the quantitative model inputs/outputs, for which further experimental and modelling work should be combined. Understanding how culture geometry guides hPSC fate patterning could lead to the optimal production of micro-patterns for yielding a particular fate, such as through self-similar designs. The long-term goal is to control a wider range of cell fates using micro-patterned cultures, optimising the results using mathematical modelling rather than experimental trial-and-error, as this would significantly reduce the associated labour and financial costs of laboratory protocols for controlling cell fate compared to the current state-of-the-art methods [165, 166]. Custom designs of micro-patterned chips for cell culture, intended to be manufactured for this purpose, are given in Appendix D. These designs were developed for this research, but due to the covid-19 pandemic followed by a lack of laboratory funds, the original experimental plan could not be completed.

Chapter 4

A New Quantification of Local Density in hPSC Cultures to Identify its Role in Mechanical Competition

4.1 Introduction

Unwanted genetic changes in hPSCs pose a threat to stem cell research. To advance regenerative medicine, a better understanding is needed for how genetically variant hPSCs take over wildtype cultures. In this chapter, local density analysis of wildtype and variant hPSCs is presented. Delaunay triangulation quantifies the spatial spread of cell nuclei in culture, which is translated into a local density measure. It was found that wildtype hPSCs grow at a significantly lower local density than variant cells in separate culture. Moreover, wildtype cells co-cultured with variants are forced into high density regions which they cannot survive due to a shift in localisation of yes-associated protein 1 (YAP) from nuclear to cytoplasmic. This culminates in wildtype cell apoptosis through mechanical competition.

Contributions of Work in This Chapter

The results of this chapter formed part of a journal article as detailed below. I.B., T.A.R., C.J.P., and D.S. conceived and designed the experiments. C.J.P., I.B., D.S., B.A.S., and J.L. performed the experiments. C.J.P., D.S., P.J.G., I.B., S.S., and B.A.S. analyzed the data. I.B., C.J.P., P.J.G., D.S., and T.A.R. wrote the paper. Contributions from authors are accredited in text where applicable.

Publications

Christopher J. Price, Dylan Stavish, Paul J. Gokhale, Ben A. Stevenson, **Samantha Sargeant**, Joanne Lacey, Tristan A. Rodriguez, Ivana Barbaric. Genetically variant human pluripotent stem cells selectively eliminate wild-type counterparts through YAP-mediated cell competition. *Developmental Cell*, 56(17):2455-2470, September 2021.

Structure of this Chapter

The remainder of this chapter is organised as follows. Section 4.2 provides background and a literature review within the field. Section 4.3 gives an overview of the hPSCs grown in separate/co-culture. Quantification and applications of local cell density analysis for wildtype and variant hPSCs within each culture condition are given in Section 4.4. Results of this analysis are presented in Section 4.5, with implications of these findings. Section 4.6 presents a discussion and conclusions for this chapter.

4.2 Background and Literature Review

Genetic Mutations in Pluripotent Stem Cells

Pluripotent stem cells in culture have the capacity to acquire genetic mutations over time, posing a significant risk to the quality of stem cell research [38, 167]. A major form of genetic abnormality in cells is changes to chromosomes. On a small scale, this may include a point mutation, i.e. the substitution of a single nucleotide for another. On a larger scale, more copy number changes may be present, which can include full aneuploidy, i.e. a full copy of the chromosome [168].

Whilst some level of genetic mosaicism is normal in mammalian cell populations for maintaining homeostasis [169], cells with particular mutations possess significant advantages, resulting in wildtype cell elimination. The most prevalent of such in hPSCs entail gains of chromosomes 1, 12, 17, 20, and X [170, 171, 172, 173]. Copy number variant (CNV) hPSCs with these genetic mutations can display growth advantages that display similarities to carcinogenesis [174, 175].

Concerningly, dopaminergic neurons derived from hESCs resulted in tumour formation when transplanted into the brains of primates with Parkinson's disease [176]. Although this has not been seen in humans for hPSC derived cells, human fetal neural stem cells that were used to treat a child with ataxia telangiectasia in 2009 resulted in a brain tumour [177]. Considering mutations in hPSCs, researchers at the RIKEN institute in Japan halted the first trial involving iPSCs in 2015 after finding both single nucleotide variants (SNVs) and CNVs in transplanted cells; fortunately, no adverse effects in patients were observed [178]. Advancements have since been made: earlier this year, dopaminergic neurons were efficiently obtained from iPSCs by inhibiting Notch signalling *in vitro* [179], demonstrating the potential of hPSCs in cell regeneration. However, the possibility of transplanting stem cells harbouring mutations remains topical [43].

Cellular Neighbourhood Interactions Mediate Super-Competition in Culture

The process by which advantageous cells ('winners') dominate cells less fit for the environmental conditions ('losers') through cell-cell interactions is referred to as cell competition [180]. Thus, 'winner' hPSCs out-compete 'loser' hPSCs by inducing apoptosis, senescence, or differentiation. As wildtype hPSCs are otherwise healthy - rather than unfit - yet lose in competition with variants, this is a subset of cell competition known as super-competition [181].

Various mechanisms have been found to contribute to cell competition. A technology for analysing individual cell behaviours promoting viability in co-culture was published in 2017 [54]. The cells used were wildtype MDCK (MDCK$^{\mathrm{WT}}$) cells and scribble$^{\mathrm{kd}}$ cells, the former of which competitively wins when co-cultured with the latter. Time-lapse microscopy images were used as inputs, where each cell was then tracked computationally using a Bayesian tracking algorithm with Brownian motion.

Whilst not applying to stem cells, the findings support that the immediate neighbourhood of a cell has a high impact on determining cell fate. The authors found that both local cell density and the state of a cell's neighbours can be used to identify which phase the cell is in (interphase, prophase/prometaphase, metaphase, anaphase/telophase, apoptosis). Delaunay triangulation was used to quantify cell spread, using cell nuclei as vertices for the triangulation. The local density for each cell was then taken as the inverse sum of the area of Delaunay triangles sharing a vertex with the cell of interest.

Moreover, the homeostatic density of scribble$^{\mathrm{kd}}$ cells was 1.4-fold higher than the density the cells reached after 80 hours in competition with MDCK$^{\mathrm{WT}}$ cells. Conversely, MDCK$^{\mathrm{WT}}$ cells had a similar density when cultured alone and when co-cultured with scribble$^{\mathrm{kd}}$ cells. This suggests that there are some mechanical features of MDCK$^{\mathrm{WT}}$ cells that allow them to out-compete scribble$^{\mathrm{kd}}$ cells by forcing them into high density regions. This could be due to a number of features such as increased motility, a higher diffusivity rate, and/or increased cell cycle time.

Furthering the role of the cellular neighbourhood, long non-coding RNA in cancer cells induces epigenetic changes to neighbours which yield fitness-decreasing mutations [182, 183]. As these epigenetic changes inhibit stem cell properties of cancer cells, their tendency to chemoresistance is decreased, resulting in better prognosis for the patient.

YAP as a Density-Dependent Regulator of Healthy Stem Cell Function

Yes-associated protein 1 (YAP) is key for coordinating stem cell self-renewal, proliferation, differentiation, and apoptosis [184], whilst also being positively linked to cytoplasmic rigidity [185, 186]. YAP and WW-domain-containing transcription regulator 1 (TAZ) are popular targets in cancer and regenerative medicine, as their activation promotes stem cell properties of cancer cells [187]. YAP/TAZ are part of the Hippo pathway,

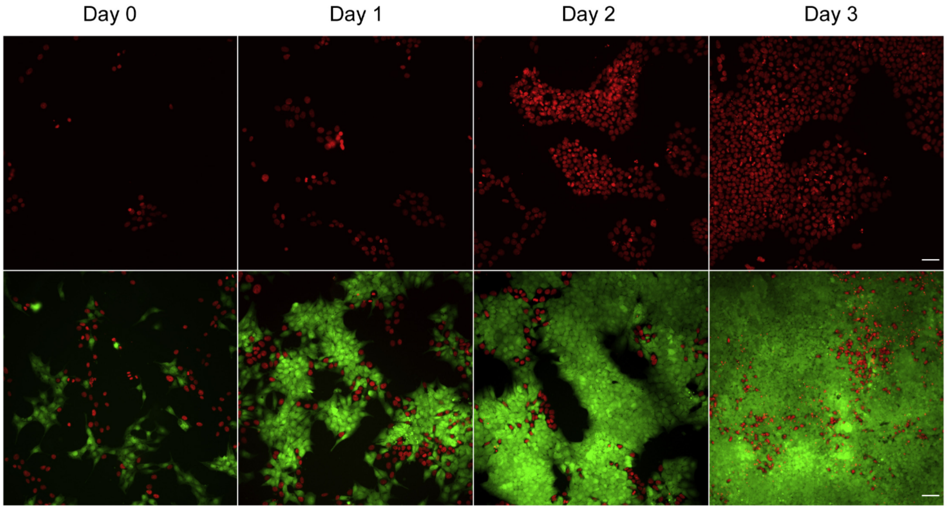


Figure 4.1: Figure (1C) from Price *et al.* (2021) showing H7 hPSCs in culture. The top row shows wildtype-RFP cells in separate culture. The bottom row shows wildtype-RFP and variant-GFP cells in co-culture. The scale bar represents 50 µm.

which is regulated by cell density *in vitro* [188]. In particular, high cell density was found to promote YAP phosphorylation in hPSCs, which also correlated with cytoplasmic localisation of YAP. These findings suggest that a forced increase in local cell density could result in adverse outcomes for hPSCs through YAP inactivation.

4.3 Data and Project Overview

The wildtype cells used for this work were diploid H7 hPSCs, which were either unlabelled or tagged with RFP (wildtype-RFP). These two conditions for wildtype cells had no impact on their behaviour, and acted merely as identification tools. The variant cells were aneuploid H7 CNV hPSCs tagged with GFP (variant-GFP). Variants had a gain on chromosomes 1, 12, 17q, and 20q. The wildtype and variant cells were grown in separate/co-culture, thereby yielding three distinct culture conditions. Time-lapse microscopy images of the cells were obtained using a BioStation, which is a non-invasive imaging platform, over a 96 hour period. Representative images at 24 hour intervals are shown in Figure 4.1 [189].

All time-lapse imaging data was generated and processed in CellProfiler prior to the work undertaken in this chapter. Resultantly, the data analysed was in the format of a .csv file, where each row corresponded to a single cell. The columns represented cellular features, including the (x,y) coordinates of cell nuclei centres. The programming language R was implemented for analysis.

4.4 Methods for the Quantification of Cell Compactness

In order to quantify the spatial distribution of cells in each image, Delaunay triangulation was performed according to the process given in Appendix B. Cell nuclei centres were used as vertices for the triangulation.

Next, using a method from the literature [54], the local cell density ρ for each cell was computed as

$$\rho = \sum_{i=1}^{N} \frac{1}{A(i)}, \tag{4.1}$$

where N is the number of Delaunay triangles sharing a common vertex with the cell, and $A(i)$ is the area of triangle i $(i = 1, ..., N)$. The smaller a cell's sum of Delaunay triangle areas, the more compact the cell is. Thus, taking the inverse of this sum results in a measure of the cell's local density.

4.5 Results

4.5.1 Local Density of Wildtype and Variant hPSCs

Equation 4.1 was applied to the wildtype and variant cells in separate culture, and then also in co-culture. Figure 4.2 shows the results of this analysis. Outliers were removed from the plot if they satisfied one of the following conditions:

$$x < Q1 - 1.5 \times IQR, \tag{4.2}$$

$$x > Q3 + 1.5 \times IQR, \tag{4.3}$$

where x is a data observation, $Q1$, $Q3$ are respectively the first and third quartiles of the data, and the inter-quartile range $IQR = Q3 - Q1$. The relevant quartiles for ordered observations are given by

$$Q1 = \begin{cases} (x_{n^*} + x_{n^*+1})/2 & \text{if } n^* \in \mathbb{Z}, \\ x_{\lceil (n^*) \rceil} & \text{otherwise,} \end{cases} \tag{4.4}$$

$$Q3 = \begin{cases} (x_{3n^*} + x_{3n^*+1})/2 & \text{if } 3n^* \in \mathbb{Z}, \\ x_{\lceil (3n^*) \rceil} & \text{otherwise,} \end{cases} \tag{4.5}$$

where n is the number of observations in the data, n^* is the index $n/4$, and $\lceil y \rceil$ is the ceiling function applied to y.

A statistically significant difference $(p < 0.01)$ was observed between the wildtype cells

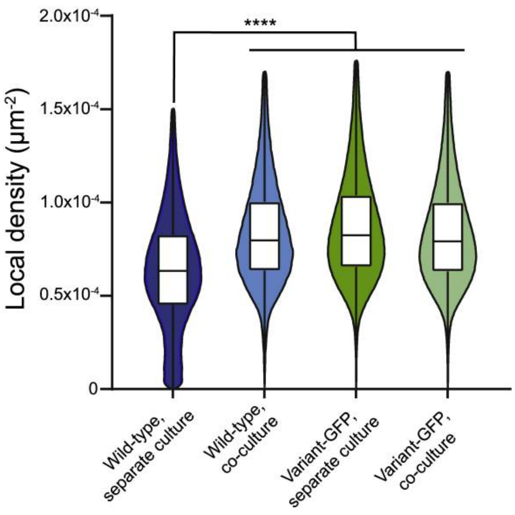


Figure 4.2: Violin plot showing local cell density against cell type. The local cell density for each cell is given by $\rho = \sum_{i=1}^{N} \frac{1}{A(i)}$, where n is the number of Delaunay triangles sharing a common vertex with the cell, and $A(i)$ is the area of triangle i $(i = 1, ..., N)$.

in separate culture and the remaining three culture conditions by performing a two-way student's t-test. This tests the hypothesis that a sample of size n, with mean and standard deviation $\bar{x}$ and s respectively, are randomly selected from a population with mean μ and unknown standard deviation. This information is combined in the t-statistic, which is given by

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}. \tag{4.6}$$

4.5.2 Implication for Mechanical Competition

Using immunofluorescence staining, the expression of YAP was inspected for wildtype and variant hPSCs in separate culture. It was observed that both cell types expressed YAP predominantly in their nuclei. However, when YAP expression was inspected for wildtype and variant hPSCs in co-culture, a different result was seen: variant cells retained their nuclear expression of YAP, whereas wildtype cells displayed a shift in YAP localisation from nuclear to cytoplasmic. This consequently inactivated YAP in wildtype cells.

For further investigation, YAP was artificially over-expressed in wildtype hPSCs. Doing so was found to increase the cells' growth rate and increase their density in separate culture. Moreover, wildtypes that over-expressed YAP had increased resistance to the compactness forced upon them in co-culture with variants. Finally, variants that were subjected to the knockdown of YAP were less successful at dominating wildtype cultures.

As a result of these findings, a protocol was conducted to promote the nuclear localisation of YAP in wildtype cells at high density. No cell competition between variant and wildtype cells was observed at low density, but as cells are forced into higher density regions, wildtype senescence and apoptosis begin. However, wildtypes that retained nuclear YAP expression were more resistant to mechanical competition, and higher proportions of these cells survived in co-culture. The combination of findings provides substantial support that YAP mediates the mechanical competition of hPSCs.

4.6 Summary and Discussion

The pivotal novel finding of this work was that genetically variant hPSCs employ high density growth as a super-competition tool against wildtypes in co-culture through the inactivation of YAP - a protein linked to hPSC proliferation and survival - in wildtypes. To achieve this result, a quantification of local cell density was applied in this chapter for the first time to hPSCs. This revealed that genetically variant H7 cells force their wildtype counterparts into significantly higher density regions than their homeostatic density, during which the wildtype cells are eliminated. Investigating this further, a migration from nuclear to cytoplasmic YAP localisation in wildtype cells co-cultured with variants was observed. This localisation shift consequently inactivates YAP, resulting in wildtype cells being unable to withstand the mechanical stress of highly compact culture.

The observation that hPSCs fare poorly in co-culture with those more genetically suited to the environment is in alignment with previous findings [190]. The results are also fitting with links between the nuclear localisation of YAP and cytoplasmic rigidity [185, 186]. A previously proposed quantification method for local cell density was implemented [54], which helps to fill a gap in existing knowledge linking YAP activation with local cell density [188].

Expectantly, the outcome of this work is to advance the use of the culture conditions proposed which increase the survival likelihood of wildtype hPSCs among variants. Specifically, growing cells at low density and/or promoting nuclear localisation of YAP in wildtype hPSCs could prevent cellular super-competition in culture.

In conjunction with this, it would be useful to implement an efficient method for identifying genetic variants in culture. This could lead to the removal of variants before wildtype cells are compromised, providing additional security. In developing such a method, cellular morphological and motility features may be useful for predicting how the cells fare in competition [191]. Moreover, many types of cancerous tumours are maintained by their own stem cells. These are similar to hPSCs in their self-renewal capabilities, but have advantageous mutations which allow them to out-compete normal cells [192]. Therefore, a technology that could extract the key features of 'winner' cells that differ from 'loser' cells could be of great medical benefit.

Chapter 5

A Novel Approach to Automatic Detection of Genetically Variant Stem Cells from Time-Lapse Microscopy Images Using Ensemble Learning

5.1 Introduction

Automatic detection of variant hPSCs is an important but challenging task to ensure the safety of stem cell replacement therapies. In this chapter, classification models are presented to distinguish genetically wildtype and variant hPSCs in culture, generated by a cell segmentation and tracking software from time-lapse microscopy images. Specifically, a collection of base machine learning models are trained using a reduced number of features determined by forward selection. Next, these base models are combined into a novel ensemble superlearner, where a weighted voting method is introduced for feature selection. For weighted voting, the vote from each model is weighted according to its relative accuracy, compared to other models, when previously implemented on a non-overlapping subset of the data. Boosting was used for the ensemble model, which improves predictive power and combats over-fitting. The resultant model developed - using training data derived from only 3 hours of time-lapse images - has an accuracy of 73.9% and an F1-score of 0.82.

Structure of this Chapter

The remainder of this chapter is organised as follows. Section 5.2 provides background and a literature review within the field. Section 5.3 gives an overview of the data analysed in this chapter, together with the individual classifiers used. Next, Section 5.4 describes how the data is prepared for analysis, including computational cell tracking and the calculation of cellular features. Full descriptions of individual classification architectures are given in Section 5.5 with feature selection in Section 5.6, then the analogous process for ensemble classifiers is given in Section 5.7. Section 5.8 provides the results from the previous two sections, including documentation for how the final ensemble model is produced. A summary and discussion of the work are presented in Section 5.9.

5.2 Background and Literature Review

Machine Learning Approaches are Needed for Cell Biology Applications

Traditional biological methods for detecting genetic heterogeneity across cell cultures - such as karyotyping, polymerase chain reaction (PCR), and quantitative PCR - are slow and have an inadequate level of sensitivity [170]. However, there is a need for fast detection, as many variant hPSCs grow faster than wildtypes, enabling them to rapidly dominate cultures [189]. Within the last four years, faster PCR and deep sequencing methods for genetically screening stem cell populations have been developed [39, 40, 41, 42, 43], but these fall into one or more undesirable categories: (1) they remain lengthy (over 24 hours), (2) they require disruptive manual processes to the cells, (3) they are costly.
Machine learning approaches for cell biology can drastically improve the speed and accuracy of high-throughput research [193]. For classification, supervised models rely on the class of the training data being known prior to model training, whereas unsupervised classifiers work on label-free data [194]. The advantage of unsupervised classification is this lack of need for *a priori* class knowledge. However, supervised classifiers generally yield substantially higher classification accuracies [195, 196]. Supervised models are hence are preferable if the class labels are known.

Appropriate Classification Models Within Cell Biology Yield Accurate Results

One of the most popular supervised classification architectures is the support vector machine (SVM) introduced in the early 1990s [197]. A novel algorithm that incorporates the linear SVM was developed in 2012, which was shown to improve binary classification in the case of having many more features than data samples [198]. For the first step of this algorithm, a student's t-test is performed per feature split by class, leading to the removal of features which have a statistically insignificant ($p > 0.05$) inter-class difference. This passes a reduced feature set to the SVM, reducing computational cost. Recursive feature elimination (RFE) is then used to select the optimal feature set. RFE is an iterative procedure where the feature with the lowest model contribution is removed per iteration, and is often used in conjunction with SVM models [199, 200]. The authors applied their algorithm to classify cancerous endometrial cells, yielding an accuracy of 88.5%.

As the study considered initially 1428 features, removing statistically insignificant features led to a significant reduction in computational effort [198]. However, as the student's t-test takes the form of a filter method, i.e. eliminating features without any knowledge of the classifier to be used, this may result in poorer model accuracy than if a wrapper method were used, i.e. feature selection that is embedded within an iterative classification algorithm [201]. Filter selection may be primarily advantageous when dealing with a sufficiently large feature set, due to the associated reduction in computational cost compared to wrapper selection methods. Thus, the trade-off between classification accuracy and computational cost should be considered.

Logistic regression (LR) is another important and popular classification tool, which unlike SVM models, must assume a linear boundary between data classes [202, 203]. For binary classes, LR algorithms have a similar form to linear regression, but the logistic function is used as the response in order to appropriately bound the output. As the output falls in the range $[0, 1]$, this is treated as the probability that the data belongs to one of the two classes. Thus, a binary class prediction is yielded by rounding the output to the nearest integer. Although developed for binary-class data, LR can be fairly easily extended to apply to multi-class problems [204]. In some cases, LR has been shown to out-perform SVM models [205].

Following also from linear regression, applying non-parametric smoothing to the regression function yields a Gaussian process model [206, 207]. Gaussian process models are flexible in their theoretical range of applications, but the requirement for the mean and covariance of the function to be specified *a priori* can be challenging to implement for real-life applications [99].

Another widely popular linear classifier is linear discriminant analysis (LDA), introduced in 1936 [208]. It has been demonstrated that both LDA and LR are appropriate for linear classification problems, with little difference in error between them [209]. However, LDA makes more assumptions about the data - in particular, that it is normally distributed - which LR does not [209, 210].

In further support of LR, it has been found to out-perform other classifiers when used in conjunction with iterative forward feature selection [211]. Forward selection (FS) algorithms use a chosen metric to select one feature, then iteratively build up the feature set. Often, a stopping criteria is imposed to halt the FS process at some event, such as when the classification accuracy no longer increases as features are introduced to the model. The authors use forward selection based on the likelihood ratio test, which is a FS method implemented prior to classification. By combining multi-class LR with FS, the authors most successfully predict radiation-induced esophagitis in lung cancer patients undergoing proton radiotherapy. Similarly to the student's t-test, however, the implementation of the likelihood ratio test prior to training classifiers may be beneficial for improving computational speed, but runs the risk of having a sub-optimal impact on the classification accuracy.

Logic-based algorithms are a particular form of classifier which differ from those discussed above. The most popular of these are decision trees, whereby the feature that best partitions the training data into its classes is taken as the node of a tree [212]. This procedure is iterated to produce sub-trees, stopping when all subsets of the training data are of the same size. However, decision trees are more appropriate for discrete and/or categorical features, rather than continuous [212]. Moreover, decision trees are often less stable than other classifiers, meaning that small changes in the input training data may yield very large changes in the model output [213]. Therefore, logic-based algorithms are not considered for this work.

Current Machine Learning Models Applied to Stem Cells and Their Limitations

Machine learning methods have been directly applied to stem cells to accurately detect certain properties from images, including mitosis [214], specific developmental stages [215], speed of replication [216], and whether the cells are dense, spread, differentiated, or debris [217]. Additionally, Gabor texture features, introduced in 1946 [218], have shown to be highly useful in hPSC classification. Gabor texture features can accurately distinguish between four behaviours of stem cells in culture when fed into a SVM classifier [219]. When combined with principal component analysis (PCA) for dimensionality reduction, Gabor texture features can also distinguish between images of hPSCs differentiating [220].

Single-cell entropy (SCENT) is an algorithm for classifying normal and variant cancer stem cells based on their signalling entropy [**?**]. The model was inspired by Waddington's epigenetic landscape, which is a conceptual construct whereby the state of a stem cell may be spatially represented on a smooth 3D surface [221]. The highest elevated states correspond to the highest differentiation potential and lowest stability, and are often associated with pluripotency. The low points, or sinks, of the landscape correspond to cell fates. Thus, lineage trajectories may be represented by a curve that tracks a cell from a point of high elevation on the landscape to a sink.

The main assumption of SCENT is that pluripotency can be characterised by a state of high uncertainty, or entropy, resulting from all fate choices being equally likely. Conversely, cells that have committed to a lineage to some degree lose this high entropy. This assumption seems reasonable, as the authors account for previous findings showing that some cells are primed for differentiation and hence have lower entropy than naïve pluripotent cells [63]. Another key assumption is that two genes are more likely to interact with each other if they are both highly expressed. Using a random walk to model probabilistic signalling between genes, the SCENT algorithm was developed to predict a cell's differentiation potential based on this assumption. Next, the global signalling entropy was calculated as the entropy rate of the probabilistic signalling. The entropy rate was given as a weighted average of the local signalling entropy across all genes and proteins within the network.

SCENT performed clustering from scRNA-Seq profiles to infer cell clusters that co-express genes. Cells were then assigned to (1) a co-expression cluster, and (2) a differentiation potency state. Lineage trajectory network diagrams were subsequently generated by correlating these features with those of a particular landmark - representing cell state - akin to points on the surface of Waddington's epigenetic landscape. The algorithm had high accuracy in estimation of true lineage trajectories, suggesting that the signalling entropy of a cell is a beneficial feature to measure for the classification of normal and mutant cells.

Despite this progress, the challenge remains to detect variations in stem cell genetics directly from images. A major challenge faced when developing a tool for classifying wild-type and variant hPSCs from time-lapse microscopy data is that genetic heterogeneity is visually subtle. The hPSC events that were successfully classified from images, however, were heavily characterised by morphological changes, and so posed less challenging to accurately detect than genetic mosaicism. For more sensitive tasks, achieving an acceptable classification accuracy is difficult. This is demonstrated when assessing the quality of hPSC colonies from images, resulting in a best accuracy of only 62.4% [222].

Tools for Cell Feature Quantification from Microscopy Images

StemCellQC is an automated toolkit which uses phase-contrast time-lapse microscopy imaging to output hPSCs features relating to motility, morphology, and pixel intensity [223]. The accuracy of this tool when classifying cells as healthy or dying was 96%, which indicates an excellent model. It is worth noting, however, that the onset of stem cell apoptosis is another easily visually apparent event, and the results relied on 48 hours of time-lapse data as input which may not always be available. Nevertheless, the results suggest that the features calculated by StemCellQC are pivotal for capturing useful properties of stem cells.

A more complete set of morphological, statistical, and pixel intensity features is output by the software *ilastik*, which enables a workflow for semi-automatic cell segmentation, tracking, and feature extraction [224]. The outputs are listed under the following categories: standard object features, convex hull features, and skeleton features. The former incorporates features relating to pixel intensity and basic measures of shape. The latter two incorporate more specialised morphological and statistical features. Although *ilastik* does not explicitly calculate motility features, its tracking output contains all required information for their easy computation.

Ensemble Models as a Balanced and Fast Approach to Stem Cell Classification

Classification tools must ideally be fast to deploy. A poignant question in hPSC research is when are variant cells - capable of super-competition - likely to be present in culture? A review of the predisposition of hESCs to acquire genetic changes showed that culturing hESCs with feeders prolonged their karyotypic stability up to 185 passages [225]. In feeder-free culture conditions, however, hESC lines can exhibit karyotypic abnormalities as early as 10 passages. For many applications in hPSC research, feeder-free conditions are preferred as the probability of transferring viruses to the cells is substantially reduced. Last year, the Bionano Genomics Saphyr optical mapping platform was used to detect copy number changes in iPSCs to subchromosomal level [226]. A gain of chromosome 12 - responsible for stem cell super-competition - was observed in the majority of cells after just 32 passages. This highlights the need for non-disruptive and accurate single-cell classification of hPSC variants from images.

A promising approach to improving predictions based on subtle data changes is ensemble classification. Ensemble classifiers can successfully comprise different base algorithms, or repeats of the same base algorithm trained on randomly sampled subsets of the training data, in order to improve the model's classification power [227]. Specifically, a linear combination of base algorithms are fitted to the training data, which is computed by re-weighting the original data to obtain different estimates, then taking the linear sum of these estimates [228].

Two popular methods for further improving the predictive power of ensemble models are 'bagging' (short for 'boostrap aggregating') [229], and 'boosting' [230, 231]. Bagging is advantageous for reducing the variance of the model output, with each sample fitting performed in parallel. Conversely, boosting algorithms fit models sequentially, with the feature weights at each iteration dependent on those at the previous iteration.

Studies have applied ensemble classifiers to cell biology to yield accurate classification outputs [232, 233]. For stem cell applications, in recent years, single-cell variations in hPSC gene expression were successfully identified through ensemble modelling [234]. Although the model uses RNA sequencing data rather than images of cells, this is a positive advancement, further supporting the applicability of ensemble classification in stem cell research.

5.3 Data and Project Overview

The timelapse imaging data for this work was generated using Phase Focus, which is a non-destructive imaging tool. Imaging of hPSCs took place over a period of 12 hours. Images were taken as 15 minute intervals, thus yielding 48 frames per video. A total of 24 videos were generated: 17 of variant hPSC lines, and 7 of wildtype hPSC lines. A total

Table 5.1: Summary of the time-lapse imaging experiments (24 video recordings).

Well	Class	Cell line	Cell count	% of Total
A1-A3, D5-D6	Wildtype	H7S14	12785	26.38
A4-A6	Variant	H7 v17q	6554	13.52
B1-B2	Variant	H7 v20q	4286	8.84
B3-B4	Variant	H7 v1q	1930	3.98
B5-B6	Variant	H7 v1,17q	4815	9.93
C1-C2	Variant	H7 v12,17q	3354	6.92
C3-C4	Variant	H7 S6-GFP v1,12,17,20q	4014	8.28
C5-C6	Wildtype	H9	2713	5.60
D1-D2	Variant	H9 v1q	4239	8.75
D3-D4	Variant	H9 v17q	3781	7.80

of 10 different cell lines were imaged. A breakdown of this data can be found in Table 5.1.

The model architectures used to classify variant and wildtype hPSCs from the images were:

1. Support vector machine with a linear kernel (SVM$_L$),
2. Support vector machine with a Gaussian kernel (SVM$_G$),
3. Linear discriminant analysis (LDA),
4. Logistic regression (LR).

Recursive feature elimination (RFE) and forward selection (FS) were respectively implemented for each model to determine the top feature set.

5.4 Data Preparation

5.4.1 Computational Cell Tracking

Prior to to this work, each image frame had been processed and normalised using the software *Fiji*, including automatic cell segmentation. As a result, there was minimal variation in colour properties - such as brightness and contrast - between the frames that were used in this work.

The open source software *ilastik* was used for semi-automated cell tracking. Initially, *ilastik*'s Pixel Classification app was trained on the first frame of the first video. Using several manual brushstrokes for training, the foreground and background were identified by *ilastik*. Then, 6 features were used to fine-tune the segmentation of cells:

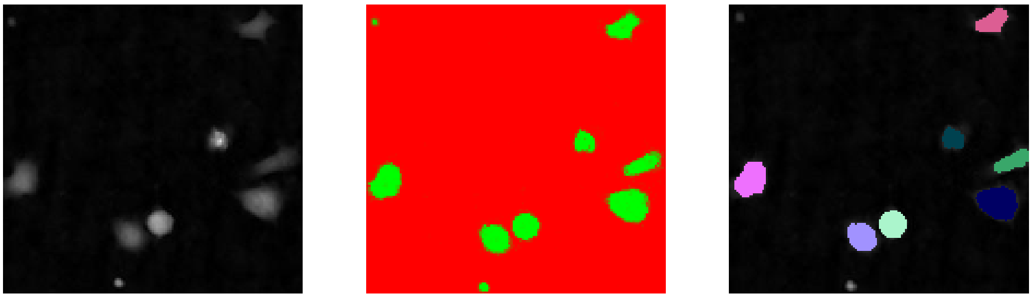
1. Gaussian smoothing for colour/intensity,

2. Laplacian of Gaussian for edges,
3. Gaussian gradient magnitude for edges,
4. Difference of Gaussians for edges,
5. Structure tensor eigenvalues for texture,
6. Hessian of Gaussian eigenvalues for texture,

using a standard deviation value of $\sigma = 1$. Detailed explanation of this process can be found in [224]. Results for all frames were batch processed according to training of this initial frame, and quality checked manually, which involved comparing by eye the pixel classification to the brightfield microscopy images.

Next, each video was individually imported into *ilastik*'s Tracking with Raw Data and Pixel Prediction Map app. Pixel intensity is in the range $[0, 1]$, where higher intensities were more likely to correspond to a cell, and lower intensities to the background. Using the pixel classification results, binary cell segmentation was performed on every frame according to a pixel intensity threshold of $0.6 \pm \epsilon$, where ϵ is a small parameter, kept constant across all frames per video, to compensate for marginally different overall brightness for some videos.

Manual training was performed on several cells to train *ilastik* to recognise cells that were dividing, and how many cells were contained within a segmented object (in the case of undersegmentation). For the former of these tasks, any cell in frame i that became two cells in frame $i + 1$ was marked at frame i, with the cell and its daughter then marked at frame $i + 1$. After sufficient manual training, visual inspection verified that *ilastik* could then automatically recognise dividing cells. For the latter of these tasks, a sample of undersegmented cell clusters were marked at frame i, together with inputting an integer value representing how many cells were within the cluster, determined by eye from microscopy images. Then, with specified weights for a cell dividing/transitioning, costs for cell appearance/disappearance, and a maximum transition neighbourhood, *ilastik* tracked the cells semi-automatically. It was found that a high penalty associated with assigning cell division/transitioning and appearance/disappearance, combined with a high distance permitted for neighbourhood transition, yielded by eye the most accurate cell tracks. This may be due to the fact that images were obtained across 15 minute intervals, so a given cell may have travelled far between consecutive frames. If these bounds were instead reversed, it was found that *ilastik* performed poorly at assigning the correct track to each cell, instead assigning many new cell IDs per frame rather than recognising a cell as existing at the previous frame. A representative image of cell sample in *ilastik* is given in Figure 5.1.



(a) Brightfield microscopy image of cells in culture.

(b) Binarised image in *ilastik*.

(c) Cells after semi-automatic tracking in *ilastik*.

Figure 5.1: Representative image of hPSCs imported to and processed by the tracking software *ilastik*. The imaged cells are H7 v20q variant. Sub-figure (b) shows two small non-cell objects detected by *ilastik* in the binarisation stage, which were later filtered from the tracking stage (c).

Retaining Single-Cell Data

Following from cell tracking, overall quality checks were manually carried out to ensure that cells had been well segmented and their trajectories accurately mapped. Although the former checks that had been carried out resulted in a good match between *ilastik*'s output and what was seen by eye, a loss of segmentation and tracking accuracy was observed over a prolonged timescale as cells began to group together. It was found that wildtype cells became less accurately identified after around 24 frames (6h), and variant around 12 frames (3h). This was unsurprising, given the result that wildtype cells grow at a significantly lower local density than variant cells in separate culture, as presented in Chapter 4. A representative image of this behaviour is shown in Figure 5.2.

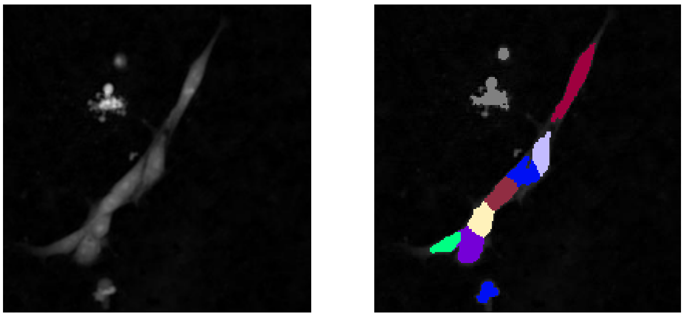
To overcome this issue, the 48 frames per video were trimmed to just the first 12 frames. Doing so ensured that the data analysed was of single cells, rather than clumps. Although videos of wildtype cells could retain up to 24 frames, reducing the number of hours for which cells must be imaged in order to apply the classification model being developed is beneficial.

5.4.2 Single Cell Features Output by *ilastik*

The output from cell tracking was 39 features for each cell per frame, quantifying standard object features, convex hull features, and skeleton features. These are as follows:

Standard Object Features

1. Bounding box maximum (x-axis),
2. Bounding box maximum (y-axis),
3. Bounding box minimum (x-axis),



(a) Brightfield microscopy image of cells in culture.

(b) Cells after semi-automatic tracking in *ilastik*.

Figure 5.2: H7 v17q variant hPSCs imaged after 10 hours. Cell boundaries are poorly segmented by *ilastik* when cells clump together.

4. Bounding box minimum (y-axis),
5. Size in pixels,
6. Kurtosis of intensity,
7. Maximum intensity,
8. Mean intensity,
9. Minimum intensity,
10. Skewness of intensity,
11. Total intensity,
12. Variance of intensity,
13. First principal component of the object,
14. Second principal component of the object,
15. Third principal component of the object,
16. Fourth principal component of the object,
17. Radii of the object (x-axis),
18. Radii of the object (y-axis).

Convex Hull Features

19. Convexity,
20. Number of defects,
21. Mean defect displacement,
22. Mean defect area,

23. Variance of defect area,
24. Convex hull centre (x-axis),
25. Convex hull centre (y-axis),
26. Object centre (x-axis),
27. Object centre (y-axis),
28. Object area.

Skeleton Features

29. Average branch length,
30. Number of branches,
31. Diameter,
32. Euclidean diameter,
33. Centre of the skeleton (x-axis),
34. Centre of the skeleton (y-axis),
35. Terminal 1 of the skeleton (x-axis),
36. Terminal 1 of the skeleton (y-axis),
37. Terminal 2 of the skeleton (x-axis),
38. Terminal 2 of the skeleton (y-axis),
39. Length of the skeleton.

For detailed descriptions of these features, please refer to the *ilastik* documentation [224].

Removing the Coordinate Dependency of Features

Features (1-4), (24-27), and (33-38) are all given as either an x or y coordinate value, representing their position in the image of hPSCs. However, the exact location of each cell is not of interest; for example, each image could be rotated - thus changing the cell coordinates - but the behaviour of cells would not differ. Therefore, the coordinate dependency of these 14 features was removed according to the following:

1. Convex hull/object centre distance (x-axis) = | Convex hull centre (x-axis) - Object centre (x-axis) |,
2. Convex hull/object centre distance (y-axis) = | Convex hull centre (y-axis) - Object centre (y-axis) |,

3. Skeleton/object centre distance (x-axis) = | Centre of the skeleton (x-axis) - Object centre (x-axis) |,
4. Skeleton/object centre distance (y-axis) = | Centre of the skeleton (y-axis) - Object centre (y-axis) |,
5. Bounding box length (x-axis) = | Bounding box maximum (x-axis) - Bounding box minimum (x-axis) |,
6. Bounding box length (y-axis) = | Bounding box maximum (y-axis) - Bounding box minimum (y-axis) |,
7. Terminal 1/object centre distance (x-axis) = | Terminal 1 of the skeleton (x-axis) - Object centre (x-axis) |,
8. Terminal 1/object centre distance (y-axis) = | Terminal 1 of the skeleton (y-axis) - Object centre (y-axis) |,
9. Terminal 2/object centre distance (x-axis) = | Terminal 2 of the skeleton (x-axis) - Object centre (x-axis) |,
10. Terminal 2/object centre distance (y-axis) = | Terminal 2 of the skeleton (y-axis) - Object centre (y-axis) |,

where $\mid x \mid$ denotes the absolute value of x. These 10 features were added to the feature set, while the 14 features used in their calculations were removed.

5.4.3 Motility, Local Cell Density, and Gabor Texture Features

Motility

The movement of wildtype and variant cells has been shown to have different characteristics *in vitro* - thus, cellular motility features were included as predictors for the classifier. To compliment the features output by *ilastik*, the following five motility features were computed on a cellular/cell population level.

At the single-cell level, across all video frames, the maximum distance travelled is given by

$$d_{max} = max_i(d(p_i, p_{i+1})). \tag{5.1}$$

The total distance travelled is given by

$$d_{tot} = \sum_{i=1}^{F-1} d(p_i, p_{i+1}). \tag{5.2}$$

The net distance travelled is given by

$$d_{net} = d(p_i, p_F), \tag{5.3}$$

where F is the total number of frames in the video of interest, p_i is the position of the centre of a cell in frame i, and $d()$ is the Euclidean distance.
To measure the linearity of cell motility, the confinement ratio (also known as the chemotactic/straightness index) is given by

$$r_{con} = \frac{d_{net}}{d_{tot}}, \tag{5.4}$$

which takes a value between 0 and 1 (0 corresponding to no net movement, 1 corresponding to an exactly linear trajectory).
The final motility feature computed was the mean squared displacement (MSD) [235], which is evaluated across a population of cells rather than at single-cell level. MSD was computed for each cell line, and is defined as

$$\text{MSD}(t) = \frac{1}{N}\sum_{n=1}^{N}(p_n(t) - p_n(0))^2, \tag{5.5}$$

where N is the total number of cells per cell line, and $p_n(t)$ is the position of the centre of a cell at time t. As the time is discretised by frame, t =0,1,...,F.

Local Cell Density

As detailed in Chapter 4, an important difference between wildtype and variant embryonic stem cells is the local cell density measure, as given by equation 4.1.
For cells close to the border of an image, a pixel threshold of 125 was set such that cells within this distance from the image edge had their local cellular density value doubled, i.e. assuming a mirrored distribution of cells.

Gabor Texture Features

The Gabor filter is a linear filter used for image processing, such as edge detection, texture analysis, and feature extraction. Gabor filters are regarded as the closest existing filter to approximating the mammalian visual cortex. A Gabor filter is given by the following equation:

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right)\cos\left(2\pi\frac{x'}{\lambda + \psi}\right), \tag{5.6}$$

where $x' = x\cos(\theta) + y\sin(\theta)$, $y' = -x\sin(\theta) + y\cos(\theta)$, λ is the sinusoidal wavelength, θ is the orientation of the normal to the stripes of the Gabor function, ψ is the sinusoid's phase offset, σ is the standard deviation of the Gaussian envelope, and γ is the ellipticity of the Gabor function (spatial aspect ratio). In particular, the parameter $\theta = \{0, 45, 90, 135\}$ is often selected to capture multiple stripe directions. The resultant set of Gabor texture features correspond to applying the set of Gabor filters to an image of an individual cell.

For this work, classification models trained on Gabor texture features produced poor results ($< 40\%$ accuracy). This may be due to the imaging platform used, as PhaseFocus creates highly pronounced contrasts across individual objects. The aim of this chapter was to create a model which can be accurately applied to data obtained across a range of imaging platforms; conversely, texture features are likely to vary significantly according to the imaging equipment used.

5.4.4 Time-Series Classification: Data Compression Methods

To remove the frame-dependency for each video, the following values were computed for each feature of a cell across all frames:

1. mean: $\bar{x} = \frac{1}{n}\left(\Sigma_{i=1}^{n} x_i\right)$ for a population with n observations,
2. variance: $\sigma^2 = \frac{1}{n}\left(\Sigma_{i=1}^{n}(x_i - \mu)^2\right)$ for a population with n observations where μ is the population mean,
3. kurtosis: $k = \frac{1}{\sigma^4}E(\underline{x} - \mu)^4$ for $\underline{x} = (x_1, x_2, ..., x_n)$ describing a population with n observations. Here, μ is the population mean, σ is the standard deviation of $\underline{x}$, and $E(t)$ represents the expected value of the quantity t,
4. skew: $s = \frac{1}{\sigma^3}E(\underline{x} - \mu)^3$, with the same variable and parameter definitions as above.

This increased the feature set by 4 times, whilst reducing the data observations.

5.4.5 Synthetic Minority Oversampling Technique for Handling Class Imbalance

For this data, the training set would contain significantly more variant than wildtype cells as approximately 77% of the cells were variant. Therefore, fitting a classification model to this data would likely not yield reliable results, as the model could achieve approximately 77% accuracy by always predicting that a cell is variant.

To overcome this problem, synthetic minority oversampling technique (SMOTE) was implemented [236]. SMOTE creates a synthetic sample of the under-represented class by perturbing an actual under-represented class sample by a small number which moves each point in the sample closer to its nearest neighbour. SMOTE is performed on a feature matrix $\mathbf{X}$ using the following algorithm:

1. Randomly select N data observations from the under-represented class in $\mathbf{X}$. The value of N was chosen such that the resultant number of observations belonging to each class was equal, i.e. $N = N_1 - N_0$, where N_1 is the number of observations in the over-represented class, and N_0 is the number of observations in the under-represented class.

2. From the above sample, select each data point's k-nearest neighbours within the sample. The nearest neighbour for data observation x_1 is the data observation x_2 for which the Euclidean distance vector $\mathbf{v} = d(x_1, x_2)$ is minimised in n-dimensional space, where n is the number of columns (features) in $\mathbf{X}$. For this work, $k = 4$.

3. Define $t \in (0, 1)$. Generate a synthetic data point $x_{new} = (1 - t)x_1 + tx_2$ for each nearest neighbour x_2 of x_1 found in step (2). For this work, $t = 0.3$.

4. Append the synthetic data to $\mathbf{X}$, and append the corresponding number of under-represented class values to the class vector.

5.4.6 *K*-Fold Cross-Validation of Feature Data

Cross-validation is an important tool for avoiding overfitting a model to data [237]. By this method, data is split into training and testing (with or without validation) sets, where the classifier is trained on the training set, yielding an associated accuracy when used to predict the class of the test set. A validation set is normally used if there is an iterative process of feature selection within each of the *K* folds [238]. If a validation set is used, models will be tested primarily on the validation set at each iteration of feature selection. This will inform which is the best model, which will finally be applied to the test set to generate an overall accuracy. This entire process is performed *K* times, where *K* is chosen based on the architecture of the data set. Importantly, data observations cannot be used in the test or validation sets multiple times.

For this project, 10-fold cross-validation was implemented. Thus, the data was randomly split into 10 training, validation, and testing sets. However, a constraint was implemented such that each of the 24 videos must be sampled from in equal proportions, each each of the sets should contain a representative sample of the total data [239]. For the individual classifiers, definitions of each generated set per fold are as follows:

- *T*: training set at each fold,

- *V1*: validation set at each fold,

- *T1*: testing set, used only once (after cross-validation is complete).

Then, informed by the results of this process, the top feature set is selected, where this process is given in Subsection 5.6.3. This reduced feature set is used to train a collection of ensemble learners. For cross-validation performed on the ensemble models, generated sets per fold are as follows:

- T: training set at each fold (invariant to that of individual models),
- *V2*: validation set at each fold,
- *T2*: testing set, used only once (after cross-validation is complete).

The T:*V1*:*T1*:*T2* splitting ratio was 7:1:1:1. To generate *V2*, 13.3% of data was randomly selected from T to reflect the size of *V1*.

Due to implementing cross-validation, SMOTE was performed once at each fold, i.e. for each training set after removing the test set from the model feature data. This is because the test set must be wholly separate from the training set; performing SMOTE on the full data would mean that relationships between features in the test set would be used to generate synthetic data for the training set, known as the malpractice of 'double-dipping'.

Before model fitting, the training data was standardised using the equation

$$\frac{x - \text{mean}(x)}{\text{standard deviation}(x)}, \tag{5.7}$$

with these mean and standard deviation values used to standardise the test data, again avoiding the problem of 'double-dipping'.

5.5 Machine Learning Classification Models

5.5.1 Classification Model Architectures

For all models described in this section, the actual class vector is denoted by y and the predicted class vector by y_{pred}. The total number of observations in the data is denoted by n.

Support Vector Machine

The SVM architecture is a non-parametric classification technique which involves drawing a decision boundary between two target classes with a $(N-1)$-dimensional hyperplane (N=number of features). The final hyperplane is chosen to maximise the margin between the two classes.

A cost function is used for margin maximisation which relies on the hinge loss. Considering hinge loss only, the cost function is defined as

$$c(x, y_{pred}, f(x)) = \max(0, 1 - f(x) \times y_{pred}), \tag{5.8}$$

where

$$f(x) = \begin{cases} 1 & \text{if } y_{pred_i} = y_i, \\ -1 & \text{otherwise.} \end{cases} \tag{5.9}$$

A regularisation parameter λ is also incorporated into the cost function to balance the hinge loss and the maximisation of the margin. Thus, the cost function becomes

$$c(x, y_{pred}, f(x)) = min_w \lambda ||w||^2 + \sum_{i=1}^{n} (1 - y_{pred_i} \langle x_i, w \rangle), \tag{5.10}$$

where w contains the weights associated with each feature.
The weights are updated according to two cases:

$$w = \begin{cases} w - \alpha(2\lambda w) & \text{if there is no misclassification,} \\ w + \alpha(y_{pred_i} \times x_i - 2\lambda w) & \text{otherwise.} \end{cases} \tag{5.11}$$

SVM models can incorporate linear or non-linear kernels to transform input data into the required form. This work used a linear kernel, which is defined as

$$k(x, y) = xy, \tag{5.12}$$

and a non-linear Gaussian kernel, which is defined as

$$k(x, y) = \exp\left(-\frac{||x - y||^2}{2\sigma^2}\right) \tag{5.13}$$

where $||x||$ is the $l1$-norm of x, and σ^2 = variance.

Linear Discriminant Analysis

Consider a feature matrix $\mathbf{X}$ such that each column vector $\mathbf{x}_i$ is the ith feature of $\mathbf{X}$ ($i = 1, ..., N$), and an associated binary target class y. Then the linear discriminant equation takes the form

$$\delta_k(\mathbf{x}) = \mathbf{x}^T \mathbf{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k^T \mathbf{\Sigma}^{-1} \mu_k + \log \pi_k, \tag{5.14}$$

where π_k is the proportion of rows in $\mathbf{X}$ belonging to the kth class, μ_k is the mean of the kth class, and $\mathbf{\Sigma}$ is the covariance matrix of $\mathbf{X}$.
LDA makes four main assumptions:

1. Multivariate normality: the conditional probability density functions $p(\mathbf{x}|y = 0)$ and $p(\mathbf{x}|y = 1)$ are both the normal distribution, with mean $(\boldsymbol{\mu}_0, \Sigma_0)$ and covariance $(\boldsymbol{\mu}_1, \Sigma_1)$,

2. Homoscedasticity: both target classes have identical covariance $(\Sigma_0 = \Sigma_1 = \Sigma)$,

3. Multicolinearity: increased correlations between features decreases prediction accuracy.

4. Independence: the value of a feature per observation is independent of other observations.

The high-dimensional feature vector **x** is projected onto a lower-dimensional vector **w**. The output class is determined by which side of a hyperplane, perpendicular to **w**, the data point (corresponding to **x**) is on.

Logistic Regression

Regression analysis models the probability of one event occurring from a choice of two event possibilities. Thus, the output y is bounded between 0 and 1. The logistic function is given by

$$p(x) = \frac{1}{1 + e^{-(x-\mu)/s}}, \tag{5.15}$$

where μ is a location parameter such that $p(\mu) = 1/2$, and s is a scale parameter.
Linear classification tasks can utilise a logistic learner to give binary output values. This is done by choosing a probability cut-off value (embedded in the learning algorithm), and labelling the outputs as either 0 or 1 depending on their closeness to the cut-off value.

5.6 Feature Selection

5.6.1 Recursive Feature Elimination

For feature selection, RFE was implemented, which removes the feature used in model training which is of the least importance for the model's performance.
For an SVM classifier, the linear score function is

$$f(x) = x'\beta + b, \tag{5.16}$$

where x is a data observation in the feature matrix $\mathbf{X}$, β contains model coefficients defining an orthogonal vector to the hyperplane (as discussed in Section 5.5.1), and b is the bias term for the model. The least important feature is that which corresponds to

$$\beta^* = min(\mid \beta^2 \mid), \tag{5.17}$$

which is then removed from the feature set prior to model re-training.
For an LDA classifier and a standardised feature matrix $\mathbf{X}$ with N features,

$$D = diag(\mathbf{X}^T * \mathbf{X}). \tag{5.18}$$

The regularised covariance matrix

$$\tilde{\Sigma} = (1-\gamma)\Sigma + \gamma D, \tag{5.19}$$

where γ is defined such that $\tilde{\Sigma}$ is non-singular whenever $\gamma \geq min(\gamma)$.
Let μ_k be the mean of the kth class in $\mathbf{X}$, and μ_0 be the global mean vector. Let also C be the correlation matrix of $\mathbf{X}$. Then the regularised correlation matrix

$$\tilde{C} = (1-\gamma)C + \gamma I, \tag{5.20}$$

where I is the identity matrix.
For the threshold δ as described in Section 5.5.1, there exists a $\delta_{pred}(i)$ $(i = 1, ..., N)$ such that if $\delta_{pred}(i) \leq \delta$, then

$$\delta \ \leq \mid \tilde{C}^{-1} D^{-1/2}(\mu_k - \mu_0) \mid . \tag{5.21}$$

If this condition is met, then the LDA classifier does not use feature i for predictions. Thus, the least important feature is that which corresponds to

$$\delta^*_{pred} = min(\mid \delta^2_{pred} \mid), \tag{5.22}$$

which is then removed from the feature set prior to model re-training.
For both the SVM and LDA model architectures, RFE was iterated such that the top 10 most important features were selected.

5.6.2 Forward Selection

To select features, the FS algorithm considers a feature matrix

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n] = \begin{bmatrix} x_1(1) & x_2(1) & ... & x_n(1) \\ x_1(2) & x_2(2) & ... & x_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(n) & x_2(n) & ... & x_n(n) \end{bmatrix}, \tag{5.23}$$

where all $\mathbf{x}_i$ are column vectors representing a feature. The output feature matrix begins empty and is 'built up', i.e. features are selected one at a time. This contrasts with RFE, where features are instead removed one at a time.

The output from FS is a feature matrix $\mathbf{F}$ containing the ordered selected features $\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_{n^*}$, where $n^* \leq n$ is the maximum number of features to be selected. For this work, $n^* = 10$.

The FS algorithm is as follows:

1. Train a classification model on each individual feature $\mathbf{x}_i$ in $\mathbf{X}$.
2. For the feature $\mathbf{x}_k$ achieving the maximum classification accuracy, let $\mathbf{f}_1 = \mathbf{x}_k$.
3. For each remaining feature $\mathbf{x}_i \neq \mathbf{f}_1$, consider $\mathbf{x}_i$ as the next candidate feature to enter matrix $\mathbf{F}$.
4. Train a classification model on $\mathbf{F}$.
5. Retain in $\mathbf{F}$ the feature $\mathbf{x}_k$ achieving the maximum classification accuracy in Step (4).
6. Repeat Steps (3)-(5) until the maximum number of features, n^*, are selected.

Orthogonal forward selection (OFS) was also considered as a feature selection technique. However, after preliminary results, OFS was not taken forward for two reasons: 1) poorer performance than FS, and 2) the columns of $\mathbf{X}$ were not linearly independent, although linear independence is an assumption of the Gram-Schmidt procedure which is used to generate an orthogonal decomposition of the feature matrix.

5.6.3 Selecting the Top Feature Set Using Novel Weighted Voting Technique

After performing K-fold cross-validation, the feature set obtaining the maximum classification accuracy may differ at each fold. However, a single feature set was sought to generate a final model.

Consider models trained on a data set such that each model is constrained to a maximum of N features (the total features in the data set may be larger than N). Then a voting system was used to select the top feature set [240]. Unlike majority voting methods, a weighted voting method is used, such that the vote of each model architecture is weighted by its relative accuracy when implemented on the individual models [241]. This was done according to the following procedure:

Size of the Top Feature Set

1. For each model i $(i = 1, ..., n)$ and cross-validation fold k $(k = 1, ..., K)$, determine the size of the feature set yielding the highest accuracy for model i, fold k. For a matrix ρ, assign $\rho_{i,k}$ as the resultant value.

2. For a vector a, assign $a_i = (1/K) \sum_{k=1}^{K} \rho_{i,k}$.

3. The size of the output top feature set is the mean of a_i, denoted n^*.

Top Features

4. Compute the mean accuracy m_i for each model i $(i = 1, ..., n)$ over all cross-validation folds k $(k = 1, ..., K)$.

5. Assign a probability p_i to each model i, representing the likelihood that model i is the most accurate. Thus $p_i = M_i / \sum_{j=1}^{n} m_j$.

6. For each model, assign a score to each feature $s_{i,j}$ $(j = 1, ..., N)$. This score is such that for RFE (or FS), the first feature to be eliminated (or the last feature to be chosen) receives a score of 1, and the final feature to be eliminated (or the first feature to be chosen) receives a score of N. Features that are not chosen are assigned a score of 0.

7. Remove features that have a score of 0 for any model. The resultant set of scores is denoted s_{i,j^*}.

8. Weight each score s_{i,j^*} by multiplying it to probability p_i. Then the final score for each feature is $f_{j^*} = \sum_{i=1}^{n} s_{i,j^*} p_i$.

9. Sort the features according to their weighted score (highest to lowest).

10. The first n^* features constitute the ranked top feature set.

The output was then used to train ensemble classifiers.

5.7 Ensemble Classification Models

Ensemble classification models combine individual learners to generate a single model output. The goals of this are to improve predictive power whilst also preventing overfitting the model to the data. Thus, a combination of several 'weak' learners - in this case, SVM$_L$, SVM$_G$, LDA, and LR - can form into one 'strong' learner. The ensemble models trained in this chapter use a MATLAB toolbox written by Victor Henrique Alves Ribeiro [242].

'Bagging' and 'boosting' are implemented with the aim of improving upon the basic ensemble model architecture. These methods rely on combining base models - confined to the model architectures given by the individual learners - to generate results. For this work, the number of base models trained by the bagged and boosted ensemble models ranged from 1 to 10. The performance of these models is discussed in Section 5.8.2.

For data pairs (X_i, Y_i), $(i = 1, ..., n)$ for d-dimensional feature vector $X_i \in \mathbb{R}^d$ and binary class response $Y_i \in \{0, 1\}$, the target function for classification is $\mathbb{P}[Y = j|X = x]$, $(j = 0, 1)$ [228]. The function estimator from a base algorithm is

$$\hat{g}(\cdot) = h_n((X_1, Y_1), ..., (X_n, Y_n))(\cdot) : \mathbb{R}^d \to \mathbb{R}, \tag{5.24}$$

where $h_n(\cdot)$ is a function that defines the estimator as a function of the data [228].

5.7.1 Bagging Method for Ensemble Models

Bagging - otherwise known as bootstrap aggregating - algorithms aim to decrease the variance in the prediction output by the ensemble model by taking a random sample of the original data with replacement, combining the observations within this sample [229]. The number of observations in the data set produced by bagging is thus equal to that of the original data set, i.e. by repeating observations within the sample. The ensemble classifier then uses the bagged data as its training input.

The bagging algorithm, as described in [228], is as follows:

1. Construct a bootstrap sample $(X_1^*, Y_1^*), ..., (X_n^*, Y_n^*)$ by randomly selecting n times with replacement from the data $(X_1, Y_1), ..., (X_n, Y_n)$.

2. Compute the bootstrapped estimator $\hat{g}^*(\cdot)$ by the equation

$$\hat{g}^*(\cdot) = h_n((X_1^*, Y_1^*), ..., (X_n^*, Y_n^*))(\cdot).$$

3. Repeat Step (1) and Step (2) M times, where M is user-specified (often 50 or 100),

thereby giving $\hat{g}^{*k}$, $(k = 1, ..., M)$. The bagged estimator is

$$\hat{g}_{bag}(\cdot) = M^{-1} \sum_{k=1}^{M} \hat{g}^{*k}.$$

4. Compute the mean of the bootstrapped probabilities

$$\hat{g}_j^{*k}(\cdot) = \hat{\mathbb{P}}^*[Y^{*k} = j | X^{*k} = \cdot], \quad (j = 0, 1),$$

giving an estimator for $\mathbb{P}[Y = j | X = x]$.

Instead of a fixed quantity n, the amount of data to be sampled can be given as a proportion p in the range $0 < p \leq 1$. For this work, a bagging proportion of $p = 0.5$ was selected. The models trained during the process of bagging are independent of each other, as the models are trained in parallel.

5.7.2 Boosting Method for Ensemble Models

Boosting algorithms iteratively update the weight of each feature c_k based on the previous iteration's fitted functions $\hat{g}_1, ..., \hat{g}_{k-1}$ [230, 231]. Thus, the models trained during boosting, unlike bagging, are dependent on one another.
Boosting as a functional gradient descent algorithm, as described in [228], aims to estimate a function $g : \mathbb{R}^d \rightarrow \mathbb{R}$, minimising expected loss

$$\mathbb{E}[\rho(Y, g(X))], \rho(\cdot, \cdot) : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}^+ \tag{5.25}$$

with loss function ρ.
The gradient descent boosting algorithm is as follows:

1. Given data $\{(X_i, Y_i) | i = 1, ..., n\}$, apply the base algorithm giving the function estimate

$$\hat{F}_1(\cdot) = \hat{g}(\cdot),$$

where $\hat{g} = \hat{g}_{X,Y} = h_n((X_1, Y_1), ..., (X_n, Y_n))$ is a function of the original data. Set $m = 1$.

2. Compute negative gradient vector

$$U_i = -\frac{\partial \rho(Y_i, g)}{\partial g}|_{g=\hat{F}_m(X_i)}, \quad (i = 1, ..., n),$$

evaluated at the current $\hat{F}_m(\cdot)$.

3. Apply the base algorithm to the gradient vector

$$\hat{g}_{m+1}(\cdot),$$

where $\hat{g}_{m+1} = \hat{g}_{X,U} = h_n((X_1, U_1), ..., (X_n, U_n))$ is a function of the original predictor variables and the current negative gradient vector as a pseudo-response.

4. Perform a one-dimensional numeric search for the best step size

$$\hat{s}_{m+1} = \text{argmin}_s \sum_{i=1}^{n} \rho(Y_i, \hat{F}_m(X_i) = s\hat{g}_{m+1}(X_i)).$$

Then, update

$$\hat{F}_{m+1}(\cdot) = \hat{F}_m(\cdot) + \hat{s}_{m+1}\hat{g}_{m+1}(\cdot).$$

5. Set $m = m+1$ and repeats Steps (2) and (3) until a stopping criterion M is achieved.

For this work, M represents the number of base models to be used for boosting, which will be varied to seek the best fit.

5.8 Results

5.8.1 Quantification of SMOTE Outputs at Each Fold for Handling Class Imbalance

As discussed in the Section 5.4, SMOTE was used to handle class imbalance by synthetically generating new samples for the minority class, which in this case was wildtype cells. SMOTE was applied 10 times to the data, i.e. when a new training set was defined upon each fold. A table showing the results of SMOTE on the training set per fold is given by Table 5.2.

5.8.2 Classification Performance of Individual and Ensemble Models

Performance of Individual Models

For each of the 10 cross-validation folds, each of the 161 features were either sequentially knocked out of (RFE), or added to (FS), the final feature set. The number of features used for model training was limited to a maximum of 10. As the RFE and FS algorithms used

Table 5.2: Results of using SMOTE on the training set at each of the 10 cross-validation folds. Values are given as before/after performing SMOTE on the data.
Key: WT = 'wildtype cells', V = 'variant cells', # = 'number', % = 'percentage'.

Test fold	WT before (#)	V before (#)	WT after (#)	V after (#)	WT before (%)	V before (%)	WT after (%)	V after (%)
1	11	33	33	33	25.0	75.0	50	50
2	14	30	30	30	31.8	68.2	50	50
3	17	27	27	27	38.6	61.4	50	50
4	9	35	35	35	20.5	79.5	50	50
5	17	27	27	27	38.6	61.4	50	50
6	12	32	32	32	27.3	72.7	50	50
7	18	26	26	26	40.9	59.1	50	50
8	12	32	32	32	27.3	72.7	50	50
9	19	25	25	25	43.2	56.8	50	50
10	11	33	33	33	25.0	75.0	50	50
Mean	13.6	29.8	29.8	29.8	30.9	67.8	50	50

for this work were wrapper methods, classification models were trained in order to determine which feature should be removed from/added to the final feature set each time. Therefore, several classification accuracies were generated at each fold, as each model was tested on the validation set *V1* and compared to *V1*'s target class. To output a single accuracy per model architecture, the model which generated the maximum accuracy per fold was selected. This model was then used to predict the class of the test set *T1*. The resultant accuracies for RFE and FS are presented in Table 5.3. If more than one model achieved the maximum accuracy, that which used the fewest features was selected.
As a consequence of these results, RFE was discontinued due to poor performance in comparison to FS. Moreover, the RFE algorithm does not accommodate non-linear classification models, so the results yielded by this method cannot incorporate SVM$_G$ or LR.

Feature Set Selection for Ensemble Modelling

Prior to identifying the top features, the best number of features at each of the 10 cross-validation folds was examined. These were taken according to the following process:

1. The model must yield a higher accuracy when predicted on the training set than when predicted on the validation set.

2. The model(s) which achieve maximal accuracy whilst adhering to (1) are selected.

Table 5.3: Classification model accuracy results for individual learners at each of the 10 cross-validation folds. Either RFE or FS was used for feature selection, and the best model per fold was tested on the set *T1*. Accuracy results are given as a percentage of correct predictions. The number of features used to train the corresponding model is given in brackets.

Test fold	SVM$_L$	LDA	SVM$_L$	LDA	SVM$_G$	LR
	(RFE)	(RFE)	(FS)	(FS)	(FS)	(FS)
1	50.0 (10)	60.0 (10)	63.6 (7)	54.5 (6)	72.7 (3)	45.5 (7)
2	50.0 (10)	50.0 (10)	54.5 (5)	63.6 (7)	90.9 (4)	81.8 (6)
3	30.0 (10)	40.0 (7)	81.8 (6)	18.2 (5)	90.9 (2)	54.5 (4)
4	80.0 (10)	70.0 (10)	90.0 (10)	80.0 (10)	80.0 (3)	80.0 (8)
5	50.0 (10)	70.0 (10)	70.0 (7)	70.0 (1)	60.0 (4)	60.0 (1)
6	40.0 (10)	40.0 (10)	70.0 (1)	70.0 (1)	90.0 (5)	70.0 (9)
7	44.4 (10)	77.8 (10)	70.0 (5)	90.0 (5)	90.0 (4)	70.0 (5)
8	33.3 (10)	55.6 (10)	80.0 (3)	70.0 (1)	90.0 (1)	40.0 (9)
9	88.9 (10)	22.2 (10)	60.0 (8)	80.0 (8)	60.0 (6)	60.0 (6)
10	77.8 (10)	55.6 (10)	60.0 (6)	80.0 (2)	90.0 (3)	60.0 (10)
Mean	54.4 (10)	54.1 (10)	69.2 (6)	63.1 (5)	80.5 (4)	60.8 (7)

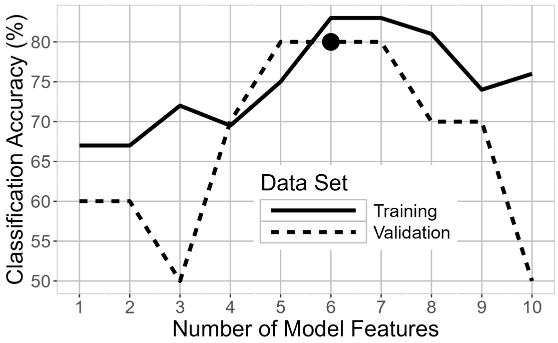


Figure 5.3: Representative plot showing the accuracy of an SVM$_G$ classification model with respect to the number of features used for model training. The best result is marked by a dot, and corresponds to the lowest number of features which maximise validation accuracy, whilst retaining a higher training accuracy. FS was used for feature selection.

3. The model from (2) which uses the least number of features is the best model, with an associated best number of features.

A representative plot displaying this process is presented in Figure 5.3.

When the best number of features at each cross-validation fold were found, the top feature set was generated using the algorithm given in Section 5.6.3 (using FS to inform the process). The best number of model features output by the algorithm was 6, which were as follows:

1. Variance of the first principal component of the object,
2. Kurtosis of the skewness of intensity,
3. Variance of the second principal component of the object,

Table 5.4: Classification accuracy results for individual and ensemble models using the top feature set. The number of base models used for training the boosted and bagged models was 3, which reflects the highest accuracy results. Each table column corresponding to a model architecture shows the accuracy of that model, given as a percentage of correct predictions. The final model is selected based on the highest validation accuracy such that the training accuracy is higher, as shown in bold.

Set	SVM_L	LDA	SVM_G	LR	Classic Ens.	Boosted Ens.	Bagged Ens.
Training	61.0	60.3	78.2	61.3	67.1	**83.3**	63.6
Validation	-	-	-	-	64.9	**66.0**	61.9
Testing	56.5	58.7	65.2	56.5	64.1	**65.2**	62.0

4. Mean of the maximum intensity,
5. Mean of the distance of an object between two consecutive frames,
6. Mean of the mean intensity.

Notably, the total number of features output by the algorithm was 15, which gave poorer performed when used together than the 6 selected features above.

Performance of Ensemble Models

Using the above feature set, ensemble classification models were trained which combine SVM_L, LDA, SVM_G, and LR with/without bagging or boosting methods to increase predictive power. The validation set used was *V2*, the test set *T2*, with the training set being the full data without *V2*, *T2*.

For ensemble models that used either boosting or bagging methods, a number of base models ranging from 1 to 10 were used for training. The number of base models which maximised the overall classification accuracies across all of the three ensemble models was 5. These classification accuracies are presented in Table 5.4.

5.8.3 Implementing K-Fold Cross-Validation for Ensemble Classifiers

Considering Table 5.4, the boosted ensemble model was selected due to it obtaining the best accuracy when predicted on the validation set. However, the accuracy of this model when predicted on the validation and test sets was not as high as sought (<70%). Moreover, some of the features have poor interpretability. To investigate whether these results could be improved upon, K-fold cross-validation was applied directly to the ensemble models. FS was used as the feature selection technique. For the boosted and bagged

Table 5.5: Classification model accuracy results for ensemble learners for each of the 10 base models. FS was used for feature selection, and the best model per fold was tested on the set *T1*. Accuracy results are given as a percentage of correct predictions. The number of features used to train the corresponding model is given in brackets. The row which maximises accuracy is shown in bold.

Base models	Classic Ens.	Boosted Ens.	Bagged Ens.
1	54.3 (5)	46.9 (5)	45.3 (4)
2	-	61.3 (8)	45.7 (2)
3	**-**	**63.4 (5)**	**57.6 (10)**
4	-	56.4 (8)	54.1 (4)
5	-	54.1 (7)	49.7 (4)
6	-	58.3 (7)	50.8 (5)
7	-	58.4 (8)	59.9 (3)
8	-	54.6 (10)	53.6 (5)
9	-	51.2 (9)	52.1 (5)
10	-	52.0 (4)	56.6 (2)
Mean	54.3 (5)	55.5 (7)	51.8 (4)

ensemble models, the set B containing the number of base models used for training was unchanged. The results are shown in Table 5.5.

From these results, using 3 base models was seen to maximise the total model accuracy. Therefore, for the boosted and bagged ensembles, only the K-fold cross-validation results using 3 base models were further considered. Results from the classic ensemble with its maximum of 1 base model were also considered for increased feature insight. As before, the algorithm given in Section 5.6.3 was implemented for selecting the top feature set, with resultant size = 6. The feature ranked in 4th place was the mean of the variance of intensity. However, with the aim of improving the model's generalisability across different imaging platforms, this feature was disregarded. The resultant top feature set was:

1. Mean of the Euclidean diameter,
2. Skewness of the size in pixels,
3. Kurtosis of the distance between the terminal 2/object centre (y-axis),
4. Mean squared displacement (MSD),
5. Mean of the distance between the convex hull/object centre (y-axis),
6. Kurtosis of the distance between the convex hull/object centre (x-axis).

5.8.4 The Final Boosted Ensemble Classification Model

As before, SMOTE was used to handle class imbalance. A final boosted ensemble model was trained using 3 base models and the above 6 features. The accuracy was 73.9% when the model was used for class prediction on the final test set *T2*. Additional prediction accuracies were 94.4% for the training set, and 70.6% for the validation set.

The performance of a classification model can be further evaluated through its precision and recall, where

$$prec = \frac{TP}{TP + FP} \quad (0 \leq prec \leq 1), \tag{5.26}$$

and

$$recall = \frac{TP}{TP + FN} \quad (0 \leq recall \leq 1). \tag{5.27}$$

Combining these results givess the F1-score, which is often referred to as the 'harmonic mean' of the precision and recall:

$$F1 = 2 \times \frac{prec \times recall}{prec + recall}. \tag{5.28}$$

The acronyms TP = true positive, TN = true negative, FP = false positive, and FN = false negative. These report the number of times that the model is correct/incorrect for predicting each class label.

These definitions are rooted in diagnostic models to predict whether a patient is unwell with a given illness. Therefore, the recall can be interpreted as the probability of a positive diagnosis if the patient is unwell, and the precision as the probability that when a positive diagnosis is made, it is done so correctly.

For this model, positive/negative classes were assigned as variant/wildtype, respectively. As the precision and recall represent the probability of the model making correct predictions, a value of 1 corresponds to the ideal for that parameter. The ensemble model precision = 0.85, and the recall = 0.78. The F1-score = 0.82. The associated confusion matrix is given in Figure 5.4.

5.9 Summary and Discussion

This chapter has presented a novel methodology for feature selection based on weighted voting of each ensemble model. This leverages majority voting methods, as previously described [240], as weighted voting can yield a conclusive output for an even split of votes, and accounts for the relative accuracy of each model that is voting [241]. Using the feature set generated by weighted voting, a boosted ensemble classification model is constructed for the accurate detection of wildtype and variant hPSCs in culture. Four classification architectures were individually trained before being combined into ensem-

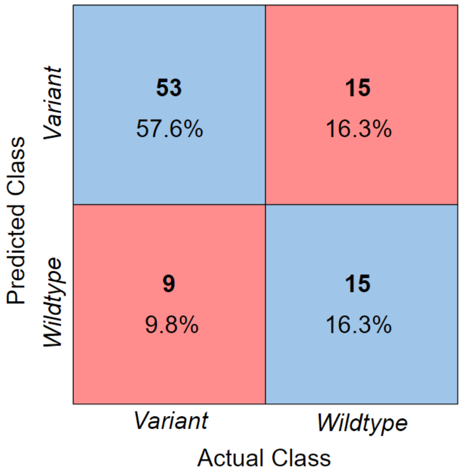


Figure 5.4: Confusion matrix for the final ensemble classification model. The number of members and percentage of overall data are given for each category. Accuracy = 73.9%, precision = 0.85, recall = 0.78, F1-score = 0.82.

ble learners. These were: (1&2) linear and Gaussian kernel support vector machines, (3) linear discriminant analysis, and (4) logistic regression. Individual learners were trained with 10-fold cross-validation, during which both recursive feature elimination (RFE) and forward selection (FS) were respectively implemented to determine the best feature set, favouring FS. These 6 features were used to train the ensemble classifier. The resultant accuracy achieved was 73.9%, with a precision of 0.85, a recall of 0.78, and F1-score 0.82. This result is promising, as it is markedly better than a model which assigns a random output class (expected accuracy 50%), whilst seemingly avoiding the problem of overfitting (expected accuracy $>$ 90%). Notably, the model's recall represents the probability that a variant cell is identified as such by the classifier. This is arguably of the highest importance when considering the context; undetected variant cells may proliferate and initiate cell competition. The model's precision corresponds to the probability that, given that a cell is classified as variant, this classification is correct. Thus, if precision is low, time may be wasted erroneously extracting wildtype cells from the population.

It was shown that FS out-performs RFE. A reason for this may be that, per iteration, FS evaluates model performance for every candidate feature added to the training set, then the feature with the best result is selected. However, RFE uses a single result of the previous iteration's model performance to eliminate a feature at the current iteration. This means that fewer possible outcomes are evaluated by RFE than FS. Evaluating the results from 10-fold cross-validation using FS, LDA giving the poorest mean accuracy may be a consequence of its modelling assumptions. Specifically, LDA models assume that, given any feature, the value at each observation is unaffected by the value at any other observation. However, it is known that neighbouring cells do affect each other; interactions between neighbouring cells impact cellular genetics and lineage [243, 244].

Although no direct links between cellular neighbourhood and the selected features have been made (while cells are still single), it would not be unreasonable to suppose that interactions between cells are present. It is hence good that the SVM models take precedent in training the ensemble learner.

Notably, no local density measures (mean, variance, skewness, kurtosis) were selected as top features, as there was no significant difference in these features between wildtype and variant cells. However, this is not in opposition with results in Chapter 4; rather, it is an expected consequence of trimming the video frames to those showing single cells (retaining frames over 3 hours). This meant that, for this work, the cells were not given the opportunity to clump together and thereby display preferential differences in local density. The model given in this chapter is not intended to address mechanisms of cell competition, but instead to identify variant cells as quickly as possible. Cell clumping would therefore be undesirable to observe for this purpose.

To verify that density findings in this chapter are consistent with those in Chapter 4, local density analysis was performed on the full 48 frames per video (12 hours of imaging). On average, the local density of variant cells between consecutive frame increased by $0.11\text{x}10^{-6}\mu\text{m}^{-2}$, whereas that of wildtype cells increased by only $0.052\text{x}10^{-6}\mu\text{m}^{-2}$. This shows that, on average, variant cells increased their local density by over twice as much as wildtype cells over a sufficient time period.

Fewer than 4% of the total feature set was used to train the ensemble model, which is beneficial for two main reasons: 1) the computational cost of the model is reduced, and 2) few measurements need to be taken for the model to operate, allowing greater generalisability. Furthering the latter point, the selected features are predominantly statistical/morphological, with one motility feature. These are likely to be standard across different imaging platforms; hence, the model may be generalised to images obtained via platforms other than Phase Focus (on which the data was obtained). Such an ease of use was a goal of the work, so it would be useful to verify this hypothesis in the future. For maximal impact, the model should also be tested on videos of wildtype and variant cells in co-culture, as this is representative of how variants would appear in laboratory applications.

The model's hypothesised generalisability could offer great use to researchers working with hPSCs. Cell culture - particularly on a large scale - may be performed in conjunction with imaging, e.g. using a microscope within the incubator. The pipeline developed in this chapter for 1) cell tracking using *ilastik*, 2) pre-processing the data, and 3) applying the ensemble model may be integrated with reasonable ease into such a laboratory set-up. Currently, there are drawbacks in the requirement of both manual training of *ilastik*, and the associated offline pre-processing. The monetary and temporal impact of these requirements are low; the time taken to train some instances of cell behaviour in *ilastik* is estimated at 1h, when using the pipeline developed in this chapter. This is less time spent than that for a qPCR experiment, which also has additional steps that can introduce error, such as fixation of the sample, transportation, and storage [245]. Ideally however, the cell segmentation and tracking would be incorporated within the MATLAB framework to improve the workflow, which may be useful future work.

The ensemble model results show that just 3 hours of imaging is sufficient to classify the cells, which improves significantly upon existing automated tools for stem cell classification, which can require 48h of data [223]. An ensemble model developed for multi-class labelling of HEp-2 cells achieved an accuracy of 80.25%, which is slightly higher than the accuracy of the classifier presented in this chapter. This is a good benchmark for which to evaluate the model, though a direct comparison of these results is not recommended, as the differences in morphology between HEp-2 cells and hPSCs are significant. In particular, HEp-2 cells generally show high ellipticity, leading to simpler computational segmentation of the cells from microscopy images. Moreover, the result in the literature uses immunofluorescence microscopy images, which greatly aids computational segmentation and tracking.

To conclude, the ensemble model presented here is the first model to classify wildtype and variant hPSCs without a need to screen the genotype of cells, yielding a very good accuracy of 73.9%. Using subtle cell features such as morphology for classification implies that the model may be implemented using data obtained across different imaging platforms, as variations such as pixel intensities are unlikely to affect the model's output. The workflow developed requires low associated labour, though improvements may be made by developing/incorporating a pipeline for cell segmentation and tracking directly within MATLAB. For a laboratory to use this workflow, it is recommended to use incubators fitted with time-lapse cameras for cell culture (3h); the data stream may then be directly input to *ilastik* (1h), the output of which then feeds to MATLAB (30m), giving a total estimated run time of 4.5h. The potential impact of this model is to reduce confounding results in hPSC research that are attributable to cells acquiring unwanted mutations in culture.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

Within the field of mathematical approaches to developmental biology, there is currently an important opportunity to take advantage of modern experimental research. In particular, there is an increasing number of studies into the differentiation of hPSCs cultured on micro-patterned geometries. The data generated by these studies may be used for modelling cell fate patterning. These models provide a basis for understanding stem cell fate decision-making, with the objective to better understand hPSC signalling behaviour. Additionally, if regenerative medicine is to be safely progressed, a detection and intervention methodology must be established to eliminate unwanted genetic variants in hPSC cultures. The work presented in this thesis aims to develop new modelling and analysis approaches that exploit the wealth of data available at singe-cell level. In this section, it is demonstrated how these objectives were met.

To this end, Chapter 2 introduced a novel spatial analysis and data-driven modelling of single-cell expression levels of Nanog (a key pluripotency regulator) for hESCs on disc micro-patterns. Nanog patterning was seen to be heterogeneous at population level, but single-cell expression was impacted by local spatial features. Specifically, a negative statistical trend was found between a cell's number of direct neighbours (cells in physical contact) and Nanog expression level. This was particularly prevalent for cells in the high Nanog (HN) population, whereas low Nanog (LN) cells displayed no significant relationship between neighbour count and Nanog expression. This suggests that hESCs are less responsive to signalling cues from their local environment when in an LN state.

This conclusion is further backed up by results of the novel model for single-cell Nanog expression levels, which models effects of both diffusive and juxtacrine (contact-based) signalling, informed by results of the spatial analysis prior conducted. This model yielded an NRMSE of 0.135 for HN cells, and 0.389 for LN cells. In other words, modelling single-cell levels of Nanog expression for LN cells using spatial features output almost thrice as much error on average than for HN cells, showing that LN cells have a reduced response to environmental signalling compared to cells in the HN population.

Chapter 3 proposed a novel dynamical network model for predicting fate patterning of hPSC populations differentiated on micro-patterns. Successful recapitulation of qualitative fate patterning reported in previous studies on various micro-patterned geometries was achieved. Notably, cell fate selection was modelled as a function of juxtacrine signalling only. This was computed based on the shortest topological path between physically contacting cells in culture. Therefore, the results may suggest that hPSCs primarily localise their cell-surface receptors which bind to the differentiating morphogen in response to signals passed through cell-to-cell contact, which is of greater importance than diffusive signals travelling through the culture medium.

Within existing literature, there is a wide range of experiments dealing with differentiating hPSCs on micro-patterns. The approach taken for this thesis was to model the spatio-temporal dynamics of a differentiating morphogen across the cell culture, which aligns with the findings of previous studies. The most well-known of these studies concerns disc micro-patterns, reporting cellular expression of several genes to determine fate. Additional studies into ellipse, triangular, and trapezoid micro-patterns only report patterning of Brachyury (a mesodermal fate marker) after differentiation.

A novel mathematical vertex model was developed for defining node symmetries on an undirected graph, where the graph is given by $G(V, E)$ with nodes V and edges E. Topological positional information was then used to assign equivalence classes (symmetries) to nodes. Specifically, equivalence was assigned to nodes i, j if the topological structure of the remainder of the network connected to i, j was identical. In this way, stem cell cultures were represented as cell networks, where nodes correspond to cells. Edges exist between neighbouring cells, i.e. physically contacting cells in culture.

Every cell was regarded as having internal dynamics, representing concentration levels of the differentiating morphogen at the cell. This was modelled using an ordinary differential equation system. Importantly, this dynamical system relates to the graph symmetry definition, as symmetric cells were modelled as having symmetric internal dynamical functions (variant only with respect to some cell permutations). Thus, time-invariant dynamical system equilibria were identical for symmetric cells. Equilibria of the system were assumed to correspond to final cell fates, with the reasoning that, at system equilibrium, morphogen concentration is unchanging and hence cells have selected fate.

In Chapter 4, a novel application of local density analysis for hPSC populations was implemented. This aims to improve understanding of super-competition mec-hanisms used by genetically variant hPSCs to out-compete wildtypes in culture. An automated tool for quantifying local cell density was applied to time-lapse images of separate cultures of wildtype and variant hPSCs, then the same method was applied to time-lapse images of wildtypes and variants in salt-and-pepper mixed co-culture. Comparisons between resultant distributions of single-cell local densities were made across each culture condition.

It was revealed that, whilst variant cells showed no significant difference in local density between separate and co-cultures, wildtypes grew at a significantly lower local density in separate culture compared to in co-culture with variants. This suggests that forcing wildtypes to grow far more compactly than their homeostatic density is a key mechanism in mechanical hPSC competition. Facilitated by these results, it was found that, in the presence of variants, wildtype hPSCs undergo re-localisation of YAP protein from the nucleus to cell membrane. As YAP regulates cytoskeletal rigidity, wildtypes are resultantly unable to withstand the mechanical stress of high density culture, leading to apoptosis. This is a key result in stem cell research.

Chapter 5 develops a novel tool for identifying genetically variant hPSCs from time-lapse images, which forms an automated framework that may be integrated into a laboratory workflow. The methodology comprises a novel ensemble classification model exploiting the 'boosting' method for improving predictive power, weighting constituent models by their performance. A very good classification accuracy of 73.9% and F1-score of 0.82 was achieved.

No other such models of this kind exist at present, which is likely due to the highly challenging nature of classifying variant hPSCs based solely on visual characteristics. This is because mutations at the chromosomal level incur extremely subtle phenotypic variations to the cell. Thus, the ensemble model substantially leverages existing methods for detecting genetically variant stem cells, as other methods tend to be costly and/or disruptive to cells, requiring the manual use of specialised machinery for population screening. These novel results are therefore very promising.

Data for several wildtype and genetic hPSC lines was used for model training, and the ensemble classifier was trained on only 3 hours of time-lapse image data. Four base model architectures were combined within the ensemble model: a support vector machine (SVM) with a linear kernel, linear discriminant analysis, binary logistic regression, and an SVM with a Gaussian kernel. It was found that the SVM using a Gaussian kernel was able to yield the most accurate results of the four models, which may be indicative of non-linear relationships between input features. Initially, 110 features describing cell morphology and motility were used as inputs to the four model architectures. Then, the feature set was reduced using forward feature selection on the individual and ensemble models, the latter of which employed a weighted voting algorithm for determining class labels. The final feature set consisted of only six features: (1) mean Euclidean diameter of the cell, (2) skewness of cell size in pixels, (3) kurtosis of distance between the terminal 2/object centre of the cell on the y-axis, (4) mean squared displacement (MSD) of the cell line, (5) mean distance between the convex hull/object centre of the cell on the y-axis, and (6) kurtosis of distance between the convex hull/object centre of the cell on the x-axis. Thus, only one motility feature at population level - the MSD - and five single-cell morphological features were selected.

A range of applications may be associated with this work; particularly, within the field of cell-based therapies for regenerative medicine, and for preventing stem cell competition, both of which require interventions to eliminate genetically variant hPSCs in culture. Ideally, hPSC populations to be used for cell therapies must not reach the stage of super-competition if genetic mutations occur in culture. In case of this, a great loss of cells would be incurred - this would not only result in a large financial loss, but also a lack of quality surrounding stem cell research. The framework presented in this thesis may be integrated into a laboratory workflow with substantial ease by fitting incubators with microscopes capable of imaging cells; these machines are currently available.

This thesis is therefore shown to positively impact (1) understanding hPSC signalling regulating gene expression, and (2) detection of genetically variant hPSCs. This is achieved by providing automated tools that deal with heterogeneity in hPSC populations through single-cell mathematical modelling.

6.2 Recommendations for Future Research

Future work is identified in two areas: (1) experimental, and (2) data-driven modelling and analysis. These are discussed below, as per the contribution chapter the work relates to.

Chapter 3

Experimental work:

1. Differentiating hPSCs on novel micro-patterned culture chips (such as those presented in Appendix D). Investigation of self-similar cell fate patterning may be facilitated through conducting these experiments, which could result from hPSC differentiation on fractal geometries. From correspondence, companies such as CYTOO™ are able to produce custom micro-patterns using computer-aided design (CAD). These designs were created as part of an experimental framework after restrictions arising from the covid-19 pandemic were alleviated. Hence, it was expected that this lab work may proceed. However, a lack of laboratory funds suddenly prevented the continuation of this work.

2. Continuing the experimental research described in Appendix A.1.5 and Appendix A.3. This entails transfecting H9TVD3 hPSCs with nuclear RFP, such that cell nuclei exhibit constant fluorescence. As the H9TVD3 cell line expresses GFP in correlation with Brachyury, the temporal evolution of single-cell Brachyury expression

may be tracked using time-lapse imaging, which would provide a far more complete picture of how cells acquire lineage compared to a static image. For this, cell nuclei could be identified and computationally segmented using RFP signals, then each cell's Brachyury expression quantified through the GFP intensity of pixels overlapping the segmented nuclei. This method is effective as Brachyury is a transcription factor, meaning that it is also localised within the cell nucleus. A novel cell line was successfully established by using electroporation to transfect the H9TVD3 cell line with RFP using the pCAG-H2B-RFP-IRES-PURO plasmid. This line was carefully grown and monitored, and cell samples were sent for karyotyping prior to their use in experiments. Unfortunately, the onset of the covid-19 pandemic and a lack of laboratory funds thereafter prevented the continuation of this work.

Modelling and analysis:

1. Using experimental data from point (1) above, predictions made by the novel dynamical model proposed in Chapter 3 could be validated. Moreover, this would provide quantitative - rather than solely qualitative - single-cell data, which is crucial for refining the model to reflect biophysical processes.

2. Time-lapse data generated by point (2) above may be used as input for the growing cell network model. Currently within the literature, gene expression patterning is reported at a static time point with a static image, representing final cell fates. Epigenetic memory - known to be an influential factor for final cell fate - is thus lost in such data. The growing model is handily set up to produce tracking constructs for time-series single-cell data as the model iterates. The intention behind this was to calibrate the growing model - which is initialised with a single cell that undergoes stochastic division and apoptosis - using time-lapse data of transfected H9TVD3 cells plated as single cells (protocol in Appendix A.1.3). Model outputs would include (1) the adjacency matrix of cells, (2) parent/daughter cell relations, and (3) gene expression levels per cell, each given for every discrete time point. Spectral analysis may be used to mathematically understand how cellular memory impacts fate. Moreover, algorithms for simulating diffusion on a graph may be implemented to incorporate the effect of diffusive signals within the model (which currently considers juxtacrine signalling only).

Chapter 5

Experimental work:

1. Generating time-lapse data of wildtype and variant hPSCs in salt-and-pepper mixed co-culture. PhaseFocus should be used to image cells at 15 minute intervals for a

total duration of 3 hours, as these were the imaging conditions used in Chapter 5. Additionally, it would be ideal to generate the same data as just described, but imaged using a BioStation rather than a PhaseFocus microscope.

Modelling and analysis:

1. Time-lapse data imaged using PhaseFocus may be used to validate the ensemble model presented in Chapter 5. The reason for this is that the model was trained on a data set containing time-lapse images of 7 wildtype and 17 variant cell lines in separate culture. For real-life applications of the classifier, variants would be present among majority wildtype cells, so model accuracy in such scenarios should be determined. However, there is no reason to suspect that the model would perform any better or worse in this case, as all classification was performed at single-cell level with the same imaging settings used for each video. Additionally, if similar time-lapse data was obtained using a BioStation, it may be useful to apply the model to this data. This is because the BioStation is a more standard imaging platform compared to PhaseFocus. If the classifier remains accurate in this task, there is a high likelihood that it may benefit a large number of stem cell research laboratories. A possible obstacle may be that the computational framework is unable to accurately segment cells when contrasts are altered. In this case, parameters may be altered in *ilastik* (software used for cell segmentation and tracking) to improve results.

Appendix C

Experimental work:

1. Equivalent to point (2) made for the future work for Chapter 3; continuing the experimental research described in Appendix A.1.5 and Appendix A.3, yielding time-lapse data.

Modelling and analysis:

1. A major question in stem cell biology may be addressed using this data: which comes first, single-cell gene expression patterning or positional information? Although such analysis was attempted for H9TVD3 cells in Appendix C, the lack of constant nuclear fluorescence made object segmentation extremely challenging.

Bibliography

[1] Wojciech Zakrzewski, Maciej Dobrzyński, Maria Szymonowicz, and Zbigniew Rybak. Stem cells: Past, present, and future. *Stem Cell Research & Therapy*, 10(1):68, 2019.

[2] Antonio P Beltrami, Laura Barlucchi, Daniele Torella, Mathue Baker, Federica Limana, Stefano Chimenti, Hideko Kasahara, Marcello Rota, Ezio Musso, Konrad Urbanek, et al. Adult cardiac stem cells are multipotent and support myocardial regeneration. *Cell*, 114(6):763–776, 2003.

[3] Mirko Corselli, Chien-Wen Chen, Mihaela Crisan, Lorenza Lazzari, and Bruno Péault. Perivascular ancestors of adult multipotent stem cells. *Arteriosclerosis, thrombosis, and vascular biology*, 30(6):1104–1109, 2010.

[4] M Hamadani, M Craig, FT Awan, and SM Devine. How we approach patient evaluation for hematopoietic stem cell transplantation. *Bone marrow transplantation*, 45(8):1259–1268, 2010.

[5] Zengrong Zhu and Danwei Huangfu. Human pluripotent stem cells: an emerging model in developmental biology. *Development*, 140(4):705–717, 2013.

[6] Michal Amit, Melissa K Carpenter, Margaret S Inokuma, Choy-Pik Chiu, Charles P Harris, Michelle A Waknitz, Joseph Itskovitz-Eldor, and James A Thomson. Clonally derived human embryonic stem cell lines maintain pluripotency and proliferative potential for prolonged periods of culture. *Developmental biology*, 227(2):271–278, 2000.

[7] James A Thomson, Joseph Itskovitz-Eldor, Sander S Shapiro, Michelle A Waknitz, Jennifer J Swiergiel, Vivienne S Marshall, and Jeffrey M Jones. Embryonic stem cell lines derived from human blastocysts. *science*, 282(5391):1145–1147, 1998.

[8] Kazutoshi Takahashi and Shinya Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–676, 2006.

[9] Junying Yu, Maxim A Vodyanik, Kim Smuga-Otto, Jessica Antosiewicz-Bourget, Jennifer L Frane, Shulan Tian, Jeff Nie, Gudrun A Jonsdottir, Victor Ruotti, Ron Stewart, et al. Induced pluripotent stem cell lines derived from human somatic cells. *Science*, 318(5858):1917–1920, 2007.

[10] Zhumur Ghosh, Kitchener D Wilson, Yi Wu, Shijun Hu, Thomas Quertermous, and Joseph C Wu. Persistent donor cell gene expression among human induced pluripotent stem cells contributes to differences with human embryonic stem cells. *PloS one*, 5(2):e8975, 2010.

[11] Kazim H Narsinh, Ning Sun, Veronica Sanchez-Freire, Andrew S Lee, Patricia Almeida, Shijun Hu, Taha Jan, Kitchener D Wilson, Denise Leong, Jarrett Rosenberg, et al. Single cell transcriptional profiling reveals heterogeneity of human induced pluripotent stem cells. *The Journal of clinical investigation*, 121(3):1217–1221, 2011.

[12] Xuetao Sun, Jun Wu, Beiping Qiang, Rocco Romagnuolo, Mark Gagliardi, Gordon Keller, Michael A Laflamme, Ren-ke Li, and Sara S Nunes. Transplanted microvessels improve pluripotent stem cell–derived cardiomyocyte engraftment and cardiac function after infarction in rats. *Science translational medicine*, 12(562):eaax2992, 2020.

[13] Chris Mason and Peter Dunnill. A brief definition of regenerative medicine. *Regenerative Medicine*, 3(1):1–5, 2008.

[14] Wenlin Li, Ke Li, Wanguo Wei, and Sheng Ding. Chemical approaches to stem cell biology and therapeutics. *Cell Stem Cell*, 13(3):270–283, 2013.

[15] Anthony E Lang and Andres M Lozano. Parkinson's disease. *New England Journal of Medicine*, 339(16):1130–1143, 1998.

[16] Marios Politis and Olle Lindvall. Clinical application of stem cell therapy in parkinson's disease. *BMC medicine*, 10(1):1–7, 2012.

[17] Frank Soldner, Dirk Hockemeyer, Caroline Beard, Qing Gao, George W Bell, Elizabeth G Cook, Gunnar Hargus, Alexandra Blak, Oliver Cooper, Maisam Mitalipova, et al. Parkinson's disease patient-derived induced pluripotent stem cells free of viral reprogramming factors. *Cell*, 136(5):964–977, 2009.

[18] Sinisa Hrvatin, Charles W O'Donnell, Francis Deng, Jeffrey R Millman, Felicia Walton Pagliuca, Philip DiIorio, Alireza Rezania, David K Gifford, and Douglas A Melton. Differentiated human stem cells resemble fetal, not adult, β cells. *Proceedings of the National Academy of Sciences*, 111(8):3038–3043, 2014.

[19] Cale N Street, Simonetta Sipione, Lisa Helms, Tanya Binette, Ray V Rajotte, R Chris Bleackley, and Gregory S Korbutt. Stem cell-based approaches to solving the problem of tissue supply for islet transplantation in type 1 diabetes. *The international journal of biochemistry & cell biology*, 36(4):667–683, 2004.

[20] Yang D Teng, Erin B Lavik, Xianlu Qu, Kook I Park, Jitka Ourednik, David Zurakowski, Robert Langer, and Evan Y Snyder. Functional recovery following traumatic spinal cord injury mediated by a unique polymer scaffold seeded with neural stem cells. *Proceedings of the National Academy of Sciences*, 99(5):3024–3029, 2002.

[21] Edward A Copelan. Hematopoietic stem-cell transplantation. *New England Journal of Medicine*, 354(17):1813–1826, 2006.

[22] Martin Körbling and Zeev Estrov. Adult stem cells for tissue repair—a new therapeutic concept? *New England Journal of Medicine*, 349(6):570–582, 2003.

[23] Han Qin, Miroslav Hejna, Yanxia Liu, Michelle Percharde, Mark Wossidlo, Laure Blouin, Jens Durruthy-Durruthy, Priscilla Wong, Zhongxia Qi, Jingwei Yu, Lei S. Qi, Vittorio Sebastiano, Jun S. Song, and Miguel Ramalho-Santos. YAP Induces Human Naive Pluripotency. *Cell Reports*, 14(10):2301–2312, 2016.

[24] Jianfeng Zhou, Jindian Hu, Yixuan Wang, and Shaorong Gao. Induction and application of human naive pluripotency. *Cell Reports*, 42(4):112379, 2023.

[25] Gail R. Martin. Teratocarcinomas and Mammalian Embryogenesis. *Science*, 209(4458):768–776, 1980.

[26] Michal Amit and Joseph Itskovitz-Eldor. Feeder-free culture of human embryonic stem cells. *Methods in Enzymology*, 420, 2006.

[27] Wenxiu Zhao, Xiang Ji, Fangfang Zhang, Liang Li, and Lan Ma. Embryonic Stem Cell Markers. *Molecules*, 17(6):6196–6246, 2021.

[28] Scott J. Rodig. Cell Staining. *Cold Spring Harbor Protocols*, 2022(6):Pdb.top099606, 2022.

[29] Kyuseok Im, Sergey Mareninov, M. Fernando Palma Diaz, and William H. Yong. An introduction to Performing Immunofluorescence Staining. *Methods in molecular biology (Clifton, N.J.)*, 1897:299–311, 2019.

[30] Mukul Tewary, Joel Ostblom, Laura Prochazka, Teresa Zulueta-Coarasa, Nika Shakiba, Rodrigo Fernandez-Gonzalez, and Peter W. Zandstra. A stepwise model of Reaction-Diffusion and Positional-Information governs self-organized human peri-gastrulation-like patterning. *Development*, 2017.

[31] Aryeh Warmflash, Benoit Sorre, Fred Etoc, Eric D. Siggia, and Ali H. Brivanlou. A method to recapitulate early embryonic spatial patterning in human embryonic stem cells. *Nature Methods*, 11(8):847–854, 2014.

[32] Thomas Graf and Matthias Stadtfeld. Heterogeneity of Embryonic and Adult Stem Cells. *Cell Stem Cell*, 3(5):480–483, 2008.

[33] Thomas F. Allison, Andrew J. H. Smith, Konstantinos Anastassiadis, Jackie Sloane-Stanley, Veronica Biga, Dylan Stavish, James Hackland, Shan Sabri, Justin Langerman, Mark Jones, Kathrin Plath, Daniel Coca, Ivana Barbaric, Paul Gokhale, and Peter W. Andrews. Identification and Single-Cell Functional Characterization of an Endodermally Biased Pluripotent Substate in Human Embryonic Stem Cells. *Stem Cell Reports*, 10(6):1895–1907, 2018.

[34] Céline Liu Bauwens, Raheem Peerani, Sylvia Niebruegge, Kimberly A. Woodhouse, Eugenia Kumacheva, Mansoor Husain, and Peter W. Zandstra. Control of Human Embryonic Stem Cell Colony and Aggregate Size Heterogeneity Influences Differentiation Trajectories. *Stem Cells*, 26(9):2300–2310, 2008.

[35] Uri Ben-David and Nissim Benvenisty. The tumorigenicity of human embryonic and induced pluripotent stem cells. *Nature Reviews Cancer*, 11(4):268–277, 2011.

[36] Evgenios Neofytou, Connor Galen O'Brien, Larry A. Couture, and Joseph C. Wu. Hurdles to clinical translation of human induced pluripotent stem cells. *The Journal of Clinical Investigation*, 125(7), 2015.

[37] Yan Shi, Jeong Tae Do, Caroline Desponts, Heung Sik Hahm, Hans R. Schöler, and Sheng Ding. A Combined Chemical and Genetic Approach for the Generation of Induced Pluripotent Stem Cells. *Cell Stem Cell*, 2(6), 2008.

[38] Peter W. Andrews, Ivana Barbaric, Nissim Benvenisty, Jonathan S. Draper, Tenneille Ludwig, Florian T. Merkle, Yoji Sato, Claudia Spits, Glyn N. Stacey, Haoyi Wang, and Martin F. Pera. The consequences of recurrent genetic and epigenetic variants in human pluripotent stem cells. *Cell Stem Cell*, 29(12), 2022.

[39] Kurt Jacobs, Afroditi Mertzanidou, Mieke Geens, Ha Thi Nguyen, Catherine Staessen, and Claudia Spits. Low-grade chromosomal mosaicism in human somatic and embryonic stem cell populations. *Nature Communications*, 5(1):4227, 2014.

[40] Alexander Keller, Laurentijn Tilleman, Dominika Dziedzicka, Filippo Zambelli, Karen Sermon, Filip Van Nieuwerburgh, Claudia Spits, and Mieke Geens. Uncovering low-level mosaicism in human embryonic stem cells using high throughput single cell shallow sequencing. *Scientific Reports*, 9(1):14844, 2019.

[41] Owen Laing, Jason Halliwell, and Ivana Barbaric. Rapid PCR Assay for Detecting Common Genetic Variants Arising in Human Pluripotent Stem Cell Cultures. *Current Protocols in Stem Cell Biology*, 49(1):e83, 2019.

[42] Victoria Steventon-Jones, Dylan Stavish, Jason A. Halliwell, Duncan Baker, and Ivana Barbaric. Single Nucleotide Polymorphism (SNP) Arrays and Their Sensitivity for Detection of Genetic Changes in Human Pluripotent Stem Cell Cultures. *Current Protocols*, 2(11):e606, 2022.

[43] Wing Hing Wong, Sima Bhatt, Kathryn Trinkaus, Iskra Pusic, Kevin Elliott, Nitin Mahajan, Fei Wan, Galen E. Switzer, Dennis L. Confer, John DiPersio, Michael A. Pulsipher, Nirali N. Shah, Jennifer Sees, Amelia Bystry, Jamie R. Blundell, Bronwen E. Shaw, and Todd E. Druley. Engraftment of rare, pathogenic donor hematopoietic mutations in unrelated hematopoietic stem cell transplantation. *Science Translational Medicine*, 12(526):eaax6249, 2020.

[44] Hilde Van de Velde, Greet Cauffman, Herman Tournaye, Paul Devroey, and Inge Liebaers. The four blastomeres of a 4-cell stage human embryo are able to develop individually into blastocysts with inner cell mass and trophectoderm. *Human Reproduction*, 23(8):1742–1747, 2008.

[45] Guillaume Blin, Catherine Picart, Sally Lowell, Darren Wisniewski, Manuel Thery, and Michel Puceat. Geometrical confinement controls the asymmetric patterning of brachyury in cultures of pluripotent cells. *Development*, 2018.

[46] Ashley R. G. Libby, Demarcus Briers, Iman Haghighi, David A. Joy, Bruce R. Conklin, Calin Belta, and Todd C. McDevitt. Automated Design of Pluripotent Stem Cell Self-Organization. *Cell Systems*, 9(5):483–495.e10, 2019.

[47] Jonathon M. Muncie, Nadia M. E. Ayad, Johnathon N. Lakins, Xufeng Xue, Jianping Fu, and Valerie M. Weaver. Mechanical Tension Promotes Formation of Gastrulation-like Nodes and Patterns Mesoderm Specification in Human Embryonic Stem Cells. *Developmental Cell*, 55(6):679–694.e11, 2020.

[48] Tiankai Zhao, Yubing Sun, Xin Li, Mehdi Baghaee, Yuenan Wang, and Hongyan Yuan. A contraction-reaction-diffusion model for circular pattern formation in embryogenesis. *bioRxiv*, page 2021.05.14.444097, 2021.

[49] Fred Etoc, Jakob Metzger, Albert Ruzo, Christoph Kirst, Anna Yoney, M. Zeeshan Ozair, Ali H. Brivanlou, and Eric D. Siggia. A Balance between Secreted Inhibitors and Edge Sensing Controls Gastruloid Self-Organization. *Developmental Cell*, 39(3), 2016.

[50] Himanshu Kaul, Nicolas Werschler, Ross D. Jones, M. Mona Siu, Mukul Tewary, Andrew Hagner, Joel Ostblom, Daniel Aguilar-Hidalgo, and Peter W. Zandstra. Virtual cells in a virtual microenvironment recapitulate early development-like patterns in human pluripotent stem cell colonies. *Stem Cell Reports*, 18(1):377–393, 2023.

[51] Içvara Barbier, Rubén Perez-Carrasco, and Yolanda Schaerli. Controlling spatiotemporal pattern formation in a concentration gradient with a synthetic toggle switch. *Molecular Systems Biology*, 16(6):e9361, 2020.

[52] David A. Rand, Archishman Raju, Meritxell Saez, Francis Corson, and Eric D. Siggia. Geometry of Gene Regulatory Dynamics. 2021.

[53] Andrew E. Teschendorff and Tariq Enver. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nature Communications*, 8, 2017.

[54] Anna Bove, Daniel Gradeci, Yasuyuki Fujita, Shiladitya Banerjee, Guillaume Charras, and Alan R. Lowe. Local cellular neighborhood controls proliferation in cell competition. *Molecular Biology of the Cell*, 28(23):3215–3228, 2017.

[55] Christopher J. Soelistyo, Giulia Vallardi, Guillaume Charras, and Alan R. Lowe. Learning the rules of cell competition without prior scientific knowledge. *bioRxiv*, page 2021.11.24.469554, 2021.

[56] Ian Chambers and Austin Smith. Self-renewal of teratocarcinoma and embryonic stem cells. *Oncogene*, 23(43):7150–7160, 2004.

[57] Hitoshi Niwa. How is pluripotency determined and maintained? *Development*, 134(4):635–646, 2007.

[58] Hitoshi Niwa. Molecular Mechanism to Maintain Stem Cell Renewal of ES Cells. *Cell Structure and Function*, 26(3):137–148, 2001.

[59] Laurie A. Boyer, Tong Ihn Lee, Megan F. Cole, Sarah E. Johnstone, Stuart S. Levine, Jacob P. Zucker, Matthew G. Guenther, Roshan M. Kumar, Heather L. Murray, Richard G. Jenner, David K. Gifford, Douglas A. Melton, Rudolf Jaenisch, and Richard A. Young. Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells. *Cell*, 122(6):947–956, 2005.

[60] Guangjin Pan and James A. Thomson. Nanog and transcriptional networks in embryonic stem cell pluripotency. *Cell Research*, 17(1):42–49, 2007.

[61] José Silva, Ian Chambers, Steven Pollard, and Austin Smith. Nanog promotes transfer of pluripotency after cell fusion. *Nature*, 441(7096):997–1001, 2006.

[62] Jennifer Nichols and Austin Smith. *Naive and Primed Pluripotent States*, volume 4. 2009.

[63] Leehee Weinberger, Muneef Ayyash, Noa Novershtern, and Jacob H. Hanna. Dynamic stem cell states: Naive to primed pluripotency in rodents and humans. *Nature Reviews. Molecular Cell Biology*, 17(3):155–169, 2016.

[64] Victor Olariu, Neil J. Harrison, Daniel Coca, Paul J. Gokhale, Duncan Baker, Steve Billings, Visakan Kadirkamanathan, and Peter W. Andrews. Modeling the evolution of culture-adapted human embryonic stem cells. *Stem Cell Research*, 4(1):50–56, 2010.

[65] Tariq Enver, Shamit Soneji, Chirag Joshi, John Brown, Francisco Iborra, Torben Orntoft, Thomas Thykjaer, Edna Maltby, Kath Smith, Raed Abu Dawud, Mark Jones, Maryam Matin, Paul Gokhale, Jonathan Draper, and Peter W. Andrews. Cellular differentiation hierarchies in normal and culture-adapted human embryonic stem cells. *Human Molecular Genetics*, 14(21):3129–3140, 2005.

[66] A. A. Markov. An Example of Statistical Investigation of the Text Eugene Onegin Concerning the Connection of Samples in Chains. *Science in Context*, 19(4):591–600, 2006.

[67] Amar M. Singh, Takashi Hamazaki, Katherine E. Hankowski, and Naohiro Terada. A Heterogeneous Expression Pattern for Nanog in Embryonic Stem Cells. *Stem Cells*, 25(10):2534–2542, 2007.

[68] T Kalmar, C Lim, P Hayward, S Muñ Oz-Descalzo, and J Nichols. Regulated Fluctuations in Nanog Expression Mediate Cell Fate Decisions in Embryonic Stem Cells. *PLoS Biology*, 7(7):1000149, 2009.

[69] Murat Acar, Jerome T. Mettetal, and Alexander van Oudenaarden. Stochastic switching as a survival strategy in fluctuating environments. *Nature Genetics*, 40(4):471–475, 2008.

[70] Gürol M. Süel, Rajan P. Kulkarni, Jonathan Dworkin, Jordi Garcia-Ojalvo, and Michael B. Elowitz. Tunability and noise dependence in differentiation dynamics. *Science*, 315(5819):1716–1719, 2007.

[71] David J. Rodda, Joon-Lin Chew, Leng-Hiong Lim, Yuin-Han Loh, Bei Wang, Huck-Hui Ng, and Paul Robson. Transcriptional Regulation of Nanog by OCT4 and SOX2. *Journal of Biological Chemistry*, 280(26):24731–24737, 2005.

[72] Yuin-Han Loh, Qiang Wu, Joon-Lin Chew, Vinsensius B. Vega, Weiwei Zhang, Xi Chen, Guillaume Bourque, Joshy George, Bernard Leong, Jun Liu, Kee-Yew Wong, Ken W. Sung, Charlie W. H. Lee, Xiao-Dong Zhao, Kuo-Ping Chiu, Leonard Lipovich, Vladimir A. Kuznetsov, Paul Robson, Lawrence W. Stanton, Chia-Lin Wei, Yijun Ruan, Bing Lim, and Huck-Hui Ng. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nature Genetics*, 38(4):431–440, 2006.

[73] Elsa Abranches, Ana M. V. Guedes, Martin Moravec, Hedia Maamar, Petr Svoboda, Arjun Raj, and Domingos Henrique. Stochastic NANOG fluctuations allow mouse embryonic stem cells to explore pluripotency. *Development*, 141(14):2770–2779, 2014.

[74] Ben D. MacArthur, Colin P. Please, and Richard O. C. Oreffo. Stochasticity and the Molecular Mechanisms of Induced Pluripotency. *PLoS ONE*, 3(8):e3086, 2008.

[75] Clinton H. Hansen and Alexander van Oudenaarden. Allele-specific detection of single mRNA molecules in situ. *Nature Methods*, 10(9):869–871, 2013.

[76] Lynn E. Heasley and Bryon E. Petersen. Signalling in stem cells. *EMBO Reports*, 5(3):241–244, 2004.

[77] Richard J. McMurtrey. Roles of Diffusion Dynamics in Stem Cell Signaling and Three-Dimensional Tissue Development. *Stem Cells and Development*, 26(18):1293–1303, 2017.

[78] Hyung Chul Lee, Cato Hastings, Nidia M. M. Oliveira, Rubén Pérez-Carrasco, Karen M. Page, Lewis Wolpert, and Claudio D. Stern. 'Neighbourhood watch' model: Embryonic epiblast cells assess positional information in relation to their neighbours. *Development*, 149(10):dev200295, 2022.

[79] Shailan B. Shah, Isaac Skromne, Clifford R. Hume, Daniel S. Kessler, Kevin J. Lee, Claudio D. Stern, and Jane Dodd. Misexpression of chick Vg1 in the marginal zone induces primitive streak formation. *Development*, 124(24):5127–5138, 1997.

[80] Albert F. Candia, Tetsuro Watabe, Stephanie H. B. Hawley, Darya Onichtchouk, Ying Zhang, Rik Derynck, Christof Niehrs, and Ken W. Y. Cho. Cellular interpretation of multiple TGF-β signals: Intracellular antagonism between activin/BVg1 and BMP-2/4 signaling mediated by Smads. *Development*, 124(22):4467–4480, 1997.

[81] John A. Fozard, Glen R. Kirkham, Lee DK Buttery, John R. King, Oliver E. Jensen, and Helen M. Byrne. Techniques for analysing pattern formation in populations of stem cells and their progeny. *BMC Bioinformatics*, 12(1):396, 2011.

[82] Simon Schardt and Sabine C Fischer. Adjusting the range of cell–cell communication enables fine-tuning of cell fate patterns from checkerboard to engulfing. *Journal of Mathematical Biology*, 87(4):54, 2023.

[83] Sabine Fischer, Simon Schardt, Joaquin Lilao-Garzon, and Silvia Munoz-Descalzo. The salt-and-pepper pattern in mouse blastocysts is compatible with signalling beyond the nearest neighbours. *bioRxiv*, pages 2023–05, 2023.

[84] Sun-Jung Kim, Jae Kyoo Lee, Jin Won Kim, Ji-Won Jung, Kwangwon Seo, Sang-Bum Park, Kyung-Hwan Roh, Sae-Rom Lee, Yun Hwa Hong, Sang Jeong Kim, Yong-Soon Lee, Sung June Kim, and Kyung-Sun Kang. Surface modification of polydimethylsiloxane (PDMS) induced proliferation and neural-like cells differentiation of umbilical cord blood-derived mesenchymal stem cells. *Journal of Materials Science: Materials in Medicine*, 19(8):2953–2962, 2008.

[85] Yadolah Dodge. *Kolmogorov–Smirnov Test*, pages 283–287. Springer Science & Business Media, New York, 2008.

[86] Shefali Talwar, Nikhil Jain, and G. V. Shivashankar. The regulation of gene expression during onset of differentiation by nuclear mechanical heterogeneity. *Biomaterials*, 35(8):2411–2419, 2014.

[87] Lois Annab, Carl Bortner, Maria Sifre, Jennifer Collins, Ruchir Shah, Darlene Dixon, H Karimi Kinyamu, and Trevor Archer. Differential Responses to Retinoic Acid and Endocrine Disruptor Compounds of Subpopulations within Human Embryonic Stem Cell Lines. *Differentiation; research in biological diversity*, 84, 2012.

[88] K. A. Schafer. The Cell Cycle: A Review. *Veterinary Pathology*, 35(6):461–478, 1998.

[89] Kevin Andrew Uy Gonzales, Hongqing Liang, Yee-Siang Lim, Yun-Shen Chan, Jia-Chi Yeo, Cheng-Peow Tan, Bin Gao, Beilin Le, Zi-Ying Tan, Kok-Yao Low, Yih-Cherng Liou, Frederic Bard, and Huck-Hui Ng. Deterministic Restriction on Pluripotent State Dissolution by Cell-Cycle Pathways. *Cell*, 162(3):564–579, 2015.

[90] WH Fleming, EJ Alpern, N Uchida, K Ikuta, GJ Spangrude, and IL Weissman. Functional heterogeneity is associated with the cell cycle status of murine hematopoietic stem cells. *Journal of Cell Biology*, 122(4):897–902, 1993.

[91] Adam A. Filipczyk, Andrew L. Laslett, Christine Mummery, and Martin F. Pera. Differentiation is coupled to changes in the cell cycle regulatory apparatus of human embryonic stem cells. *Stem Cell Research*, 1(1):45–60, 2007.

[92] Siim Pauklin and Ludovic Vallier. The Cell-Cycle State of Stem Cells Determines Cell Fate Propensity. *Cell*, 155(1):135–147, 2013.

[93] Thermo Fisher. Countess automated cell counter, cell data sheet, hesc. https://bionumbers.hms.harvard.edu/bionumber.aspx?s=n&v=0&id=10-8885, 2023.

[94] Linda Harkness, Xiaoli Chen, Marianne Gillard, Peter Paul Gray, and Anthony Mitchell Davies. Media composition modulates human embryonic stem cell morphology and may influence preferential lineage differentiation potential. *PLoS ONE*, 14(3):e0213678, 2019.

[95] Shuai Liu, Mengye Lu, Hanshuang Li, and Yongchun Zuo. Prediction of Gene Expression Patterns With Generalized Linear Regression Model. *Frontiers in Genetics*, 10, 2019.

[96] Oliver Stegle, Sebastian V. Fallert, David J. C. MacKay, and Soren Brage. Gaussian Process Robust Regression for Noisy Heart Rate Data. *IEEE Transactions on Biomedical Engineering*, 55(9):2143–2151, 2008.

[97] Ambrish G, Bharathi Ganesh, Anitha Ganesh, Chetana Srinivas, Dhanraj, and Kiran Mensinkal. Logistic regression technique for prediction of cardiovascular disease. *Global Transitions Proceedings*, 3(1):127–130, 2022.

[98] Jürgen Groβ. *Linear Regression.* Springer Science & Business Media, 2003.

[99] Carl Edward Rasmussen. Gaussian Processes in Machine Learning. In Olivier Bousquet, Ulrike von Luxburg, and Gunnar Rätsch, editors, *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*, Lecture Notes in Computer Science, pages 63–71. Springer, 2004.

[100] Alexei Botchkarev. Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology. *Interdisciplinary Journal of Information, Knowledge, and Management*, 14:045–076, 2019.

[101] P. a. P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1-2):17–23, 1950.

[102] Da-Woon Jung, Woong-Hee Kim, and Darren Reece Williams. Reprogram or Reboot: Small Molecule Approaches for the Production of Induced Pluripotent Stem Cells and Direct Cell Reprogramming. *ACS Chemical Biology*, 9(1):80–95, 2014.

[103] Richard J. McMurtrey. Analytic Models of Oxygen and Nutrient Diffusion, Metabolism Dynamics, and Architecture Optimization in Three-Dimensional Tissue Constructs with Applications and Insights in Cerebral Organoids. *Tissue Engineering Part C: Methods*, 22(3):221–249, 2016.

[104] Guilherme M. Oliveira, Attila Oravecz, Dominique Kobi, Manon Maroquenne, Kerstin Bystricky, Tom Sexton, and Nacho Molina. Precise measurements of chromatin diffusion dynamics by modeling using Gaussian processes. *Nature Communications*, 12(1):6184, 2021.

[105] Rubén Peréz-Carrasco, Pilar Guerrero, James Briscoe, and Karen M. Page. Intrinsic Noise Profoundly Alters the Dynamics and Steady State of Morphogen-Controlled Bistable Genetic Switches. *PLoS Computational Biology*, 12(10):e1005154, 2016.

[106] Richard H. Byrd, Jean Charles Gilbert, and Jorge Nocedal. A trust region method based on interior point techniques for nonlinear programming. *Mathematical Programming*, 89(1):149–185, 2000.

[107] Jorge Nocedal, Figen Öztoprak, and Richard A. Waltz. An interior point method for nonlinear programming with infeasibility detection capabilities. *Optimization Methods and Software*, 29(4):837–854, 2014.

[108] Hiroshi Ochiai, Takeshi Sugawara, Tetsushi Sakuma, and Takashi Yamamoto. Stochastic promoter activation affects Nanog expression variability in mouse embryonic stem cells. *Scientific Reports*, 4(1):7125, 2014.

[109] Philipp Thomas. Intrinsic and extrinsic noise of gene expression in lineage trees. *Scientific Reports*, 9(1), 2019.

[110] Richard N. Wang, Jordan Green, Zhongliang Wang, Youlin Deng, Min Qiao, Michael Peabody, Qian Zhang, Jixing Ye, Zhengjian Yan, Sahitya Denduluri, Olumuyiwa Idowu, Melissa Li, Christine Shen, Alan Hu, Rex C. Haydon, Richard Kang, James Mok, Michael J. Lee, Hue L. Luu, and Lewis L. Shi. Bone Morphogenetic Protein (BMP) signaling in development and human diseases. *Genes & Diseases*, 1(1):87–105, 2014.

[111] Ren He Xu, Xin Chen, Dong S. Li, Rui Li, Gregory C. Addicks, Clay Glennon, Thomas P. Zwaka, and James A. Thomson. BMP4 initiates human embryonic stem cell differentiation to trophoblast. *Nature Biotechnology*, 20(12):1261–1264, 2002.

[112] James Adjaye, John Huntriss, Ralf Herwig, Alia BenKahla, Thore C. Brink, Christoph Wierling, Claus Hultschig, Detlef Groth, Marie-Laure Yaspo, Helen M. Picton, Roger G. Gosden, and Hans Lehrach. Primary differentiation in the human blastocyst: Comparative molecular portraits of inner cell mass and trophectoderm cells. *Stem Cells*, 23(10), 2005.

[113] R. Marisol Herrera-Perez, Christian Cupo, Cole Allan, Alicia B. Dagle, and Karen E. Kasza. Tissue Flows Are Tuned by Actomyosin-Dependent Mechanics in Developing Embryos. *PRX Life*, 1(1):013004, 2023.

[114] Sui Huang, Gabriel Eichler, Yaneer Bar-Yam, and Donald E. Ingber. Cell Fates as High-Dimensional Attractor States of a Complex Gene Regulatory Network. *Physical Review Letters*, 94(12):128701, 2005.

[115] C. H. Waddington. *The Strategy of the Genes.* Routledge, 2014.

[116] Sui Huang, Yan-Ping Guo, Gillian May, and Tariq Enver. Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Developmental Biology*, 305(2):695–713, 2007.

[117] Guy Karlebach and Ron Shamir. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, 9(10):770–780, 2008.

[118] Sui Huang. Cell Lineage Determination in State Space: A Systems View Brings Flexibility to Dogmatic Canonical Rules. *PLoS Biology*, 8(5):e1000380, 2010.

[119] Julien Dubrulle and Olivier Pourquié. Coupling segmentation to axis formation. *Development*, 131(23):5783–5793, 2004.

[120] Mijo Simunovic, Jakob J Metzger, Fred Etoc, Anna Yoney, Albert Ruzo, Iain Martyn, Gist Croft, Dong Shin You, Ali H Brivanlou, and Eric D Siggia. A 3D model of a human epiblast reveals BMP4-driven symmetry breaking. *Nature Cell Biology*, 2019.

[121] Mijo Simunovic, Jakob J. Metzger, Fred Etoc, Anna Yoney, Albert Ruzo, Iain Martyn, Gist Croft, Ali H. Brivanlou, and Eric D. Siggia. Molecular mechanism of symmetry breaking in a 3D model of a human epiblast. bioRxiv, 2018.

[122] Idse Heemskerk, Kari Burt, Matthew Miller, Sapna Chhabra, M Cecilia Guerra, Lizhong Liu, and Aryeh Warmflash. Rapid changes in morphogen concentration control self-organized patterning in human embryonic stem cells. *eLife*, 8:e40526, 2019.

[123] Moshe Sheintuch and Olga Nekhamkina. Reaction-diffusion patterns on a disk or a square in a model with long-range interaction. *The Journal of Chemical Physics*, 107(19):8165–8174, 1997.

[124] Mukul Tewary, Dominika Dziedzicka, Joel Ostblom, Laura Prochazka, Nika Shakiba, Tiam Heydari, Daniel Aguilar-Hidalgo, Curtis Woodford, Elia Piccinini, David Becerra-Alonso, Alice Vickers, Blaise Louis, Nafees Rahman, Davide Danovi, Mieke Geens, Fiona M. Watt, and Peter W. Zandstra. High-throughput micropatterning platform reveals Nodal-dependent bisection of peri-gastrulation–associated versus preneurulation-associated fate patterning. *PLoS Biology*, 17(10), 2019.

[125] Ian Stewart, Martin Golubitsky, and Marcus Pivato. Symmetry Groupoids and Patterns of Synchrony in Coupled Cell Networks. *Society for Industrial and Applied Mathematics*, 2(4):609–646, 2003.

[126] Boris Greber, Guangming Wu, Christof Bernemann, Jin Young Joo, Dong Wook Han, Kinarm Ko, Natalia Tapia, Davood Sabour, Jared Sterneckert, Paul Tesar, and Hans R. Schöler. Conserved and divergent roles of FGF signaling in mouse epiblast stem cells and human embryonic stem cells. *Cell Stem Cell*, 6(3), 2010.

[127] Ludovic Vallier, Thomas Touboul, Zhenzhi Chng, Minodora Brimpari, Nicholas Hannan, Enrique Millan, Lucy E. Smithers, Matthew Trotter, Peter Rugg-Gunn, Anne Weber, and Roger A. Pedersen. Early Cell Fate Decisions of Human Embryonic Stem Cells and Mouse Epiblast Stem Cells Are Controlled by the Same Signalling Pathways. *PLoS ONE*, 4(6):e6082, 2009.

[128] David A. Turner, Pau Rué, Jonathan P. Mackenzie, Eleanor Davies, and Alfonso Martinez Arias. Brachyury cooperates with Wnt/β-catenin signalling to elicit primitive-streak-like behaviour in differentiating mouse embryonic stem cells. *BMC Biology*, 12(1):63, 2014.

[129] S. Boccaletti, Louis M. Pecora, and A. Pelaez. Unifying framework for synchronization of coupled dynamical systems. *Physical Review E*, 63(6):066219, 2001.

[130] Alexander G. Fletcher, James M. Osborne, Philip K. Maini, and David J. Gavaghan. Implementing vertex dynamics models of cell populations in biology within a consistent computational framework. *Progress in Biophysics and Molecular Biology*, 113(2):299–326, 2013.

[131] Wei Gao, Hualong Wu, Muhammad Kamran Siddiqui, and Abdul Qudair Baig. Study of biological networks using graph theory. *Saudi Journal of Biological Sciences*, 25(6):1212–1219, 2018.

[132] Willard Miller. *Symmetry Groups and Their Applications*. Academic Press, 1973.

[133] H. Brandt. Über eine verallgemeinerung des gruppenbegriffes. *Mathematische Annalen*, 96(1):360–366, 1927.

[134] P. J. Higgins. Presentations of groupoids, with applications to groups. *Mathematical Proceedings of the Cambridge Philosophical Society*, 60(1):07–20, 1964.

[135] Paul S. Jensen. Finite difference techniques for variable grids. *Computers & Structures*, 2(1):17–29, 1972.

[136] Gordon D. Smith. *Numerical Solution of Partial Differential Equations: Finite Difference Methods.* Clarendon Press, 1985.

[137] Suraj Shankar. Diffusion in 1d and 2d. https://www.mathworks.com/matlabcentral/fileexchange/38088-diffusion-in-1d-and-2d, 2023.

[138] Kunihiko Kaneko. Period-Doubling of Kink-Antikink Patterns, Quasiperiodicity in Antiferro-Like Structures and Spatial Intermittency in Coupled Logistic Lattice: Towards a Prelude of a "Field Theory of Chaos". *Progress of Theoretical Physics*, 72(3):480–486, 1984.

[139] James P. Crutchfield. Space-time dynamics in video feedback. *Physica D: Nonlinear Phenomena*, 10(1):229–245, 1984.

[140] Irene Waller and Raymond Kapral. Spatial and temporal structure in systems of coupled nonlinear oscillators. *Physical Review A*, 30(4):2047–2055, 1984.

[141] Kunihiko Kaneko. Coupled Map Lattice. In Roberto Artuso, Predrag Cvitanović, and Giulio Casati, editors, *Chaos, Order, and Patterns*, NATO ASI Series, pages 237–247. Springer US, 1991.

[142] Stephen Wolfram. Statistical mechanics of cellular automata. *Reviews of Modern Physics*, 55(3):601–644, 1983.

[143] Yakov Pesin and Alex Yurchenko. Some physical models of the reaction-diffusion equation, and coupled map lattices. *Russian Mathematical Surveys*, 59, 2004.

[144] Michelle Starz-Gaiano, Mariana Melani, Hans Meinhardt, and Denise Montell. Interpretation of the UPD/JAK/STAT morphogen gradient in Drosophila follicle cells. *Cell cycle (Georgetown, Tex.)*, 8(18), 2009.

[145] H. Sedaghat. *Nonlinear Difference Equations: Theory with Applications to Social Science Models.* Springer Science & Business Media, 2013.

[146] James D Meiss. *Differential dynamical systems.* SIAM, 2007.

[147] David Angeli, James E. Ferrell, and Eduardo D. Sontag. Detection of multistability, bifurcations, and hysteresis in a large class of biological positive-feedback systems. *Proceedings of the National Academy of Sciences*, 101(7):1822–1827, 2004.

[148] J. L. Hindmarsh, R. M. Rose, and Andrew Fielding Huxley. A model of neuronal bursting using three coupled first order differential equations. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 221(1222):87–102, 1997.

[149] Saber N Elaydi. *Discrete chaos: with applications in science and engineering*. CRC press, 2007.

[150] Sierpinski W. Sur une courbe dont tout point est un point de ramification. *C. R. Acad. Sci.*, 160:302–305, 1915.

[151] Edward Bormashenko, Irina Legchenkova, Mark Frenkel, Nir Shvalb, and Shraga Shoval. Voronoi Entropy vs. Continuous Measure of Symmetry of the Penrose Tiling: Part I. Analysis of the Voronoi Diagrams. *Symmetry*, 13(9):1659, 2021.

[152] F. A. C. Martins and D. E. Rival. A robust Voronoi-tessellation-based approach for detection of coherent structures in sparsely-seeded flows.

[153] Atsuyuki Okabe, Barry Boots, and Kokichi Sugihara. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. John Wiley & Sons, Inc., 1992.

[154] Jakob Sievers. Voronoilimit(varargin). https://www.mathworks.com/matlabcentral/fileexchange/34428-voronoilimit-varargin, 2023.

[155] Patrick S. Stumpf, Fumio Arai, and Ben D. MacArthur. Heterogeneity and 'memory' in stem cell populations. *Physical Biology*, 17(6):065013, 2020.

[156] Irene Otero-Muras, Rubén Pérez-Carrasco, Julio R Banga, and Chris P Barnes. Automated design of gene circuits with optimal mushroom-bifurcation behavior. *Iscience*, 26(6), 2023.

[157] Klaus A. Becker, Prachi N. Ghule, Jaclyn A. Therrien, Jane B. Lian, Janet L. Stein, Andre J. van Wijnen, and Gary S. Stein. Self-renewal of human embryonic stem cells is supported by a shortened G1 cell cycle phase. *Journal of Cellular Physiology*, 209(3):883–893, 2006.

[158] Ivana Barbaric, Veronica Biga, Paul J. Gokhale, Mark Jones, Dylan Stavish, Adam Glen, Daniel Coca, and Peter W. Andrews. Time-Lapse Analysis of Human Embryonic Stem Cells Reveals Multiple Bottlenecks Restricting Colony Formation and Their Relief upon Culture Adaptation. *Stem Cell Reports*, 3(1), 2014.

[159] F. Cecconi, M. Cencini, M. Falcioni, and A. Vulpiani. Brownian motion and diffusion: From stochastic processes to chaos and beyond. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 15(2):026102, 2005.

[160] Han Qin, Tianxin Yu, Tingting Qing, Yanxia Liu, Yang Zhao, Jun Cai, Jian Li, Zhihua Song, Xiuxia Qu, Peng Zhou, Jiong Wu, Mingxiao Ding, and Hongkui Deng. Regulation of Apoptosis and Differentiation by p53 in Human Embryonic Stem Cells*. *Journal of Biological Chemistry*, 282(8):5842–5852, 2007.

[161] Lewis Wolpert. Positional information and the spatial pattern of cellular differentiation. *Journal of Theoretical Biology*, 25(1):1–47, 1969.

[162] Risi Imre Kondor and John Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proceedings of the 19th international conference on machine learning*, volume 2002, pages 315–322, 2002.

[163] Ralucca Gera, L. Alonso, Brian Crawford, Jeffrey House, J. A. Mendez-Bermudez, Thomas Knuth, and Ryan Miller. Identifying network structure similarity using spectral graph theory. *Applied Network Science*, 3(1):2, 2018.

[164] Dorina Thanou, Xiaowen Dong, Daniel Kressner, and Pascal Frossard. Learning Heat Diffusion Graphs. *IEEE Transactions on Signal and Information Processing over Networks*, 3(3):484–499, 2017.

[165] Hamzah Aweidah, Chen Matsevich, Hanita Khaner, Masha Idelson, Ayala Ejzenberg, Benjamin Reubinoff, Eyal Banin, and Alexey Obolensky. Survival of Neural Progenitors Derived from Human Embryonic Stem Cells Following Subretinal Transplantation in Rodents. *Journal of Ocular Pharmacology and Therapeutics*, 39(5):347–358, 2023.

[166] Elmira Rezaei Zonooz, Zahra Ghezelayagh, Azadeh Moradmand, Hossein Baharvand, and Yaser Tahamtani. Protocol-Dependent Morphological Changes in Human Embryonic Stem Cell Aggregates during Differentiation toward Early Pancreatic Fate. *Cells Tissues Organs*, pages 1–12, 2022.

[167] Peter W. Andrews, Uri Ben-David, Nissim Benvenisty, Peter Coffey, Kevin Eggan, Barbara B. Knowles, Andras Nagy, Martin Pera, Benjamin Reubinoff, Peter J. Rugg-Gunn, and Glyn N. Stacey. Assessing the Safety of Human Pluripotent Stem Cells and Their Derivatives for Clinical Applications. *Stem Cell Reports*, 9(1):1–4, 2017.

[168] Alexander Keller and Claudia Spits. The Impact of Acquired Genetic Abnormalities on the Clinical Translation of Human Pluripotent Stem Cells. *Cells*, 10(11):3246, 2021.

[169] M. A. Kingsbury, B. Friedman, M. J. McConnell, S. K. Rehen, A. H. Yang, D. Kaushal, and J. Chun. Aneuploid neurons are functionally active and integrated into brain circuitry. *Proceedings of the National Academy of Sciences*, 102(17):6143–6147, 2005.

[170] Duncan Baker, Adam J. Hirst, Paul J. Gokhale, Miguel A. Juarez, Steve Williams, Mark Wheeler, Kerry Bean, Thomas F. Allison, Harry D. Moore, Peter W. An-

drews, and Ivana Barbaric. Detecting Genetic Mosaicism in Cultures of Human Pluripotent Stem Cells. *Stem Cell Reports*, 7(5):998–1012, 2016.

[171] Jonathan S. Draper, Kath Smith, Paul Gokhale, Harry D. Moore, Edna Maltby, Julie Johnson, Lorraine Meisner, Thomas P. Zwaka, James A. Thomson, and Peter W. Andrews. Recurrent gain of chromosomes 17q and 12 in cultured human embryonic stem cells. *Nature Biotechnology*, 22(1):53–54, 2004.

[172] International Stem Cell Initiative, Katherine Amps, Peter W. Andrews, George Anyfantis, Lyle Armstrong, Stuart Avery, Hossein Baharvand, Julie Baker, Duncan Baker, Maria B. Munoz, Stephen Beil, Nissim Benvenisty, Dalit Ben-Yosef, Juan-Carlos Biancotti, Alexis Bosman, Romulo Martin Brena, Daniel Brison, Gunilla Caisander, María V. Camarasa, Jieming Chen, Eric Chiao, Young Min Choi, Andre B. H. Choo, Daniel Collins, Alan Colman, Jeremy M. Crook, George Q. Daley, Anne Dalton, Paul A. De Sousa, Chris Denning, Janet Downie, Petr Dvorak, Karen D. Montgomery, Anis Feki, Angela Ford, Victoria Fox, Ana M. Fraga, Tzvia Frumkin, Lin Ge, Paul J. Gokhale, Tamar Golan-Lev, Hamid Gourabi, Michal Gropp, Guangxiu Lu, Ales Hampl, Katie Harron, Lyn Healy, Wishva Herath, Frida Holm, Outi Hovatta, Johan Hyllner, Maneesha S. Inamdar, Astrid Kresentia Irwanto, Tetsuya Ishii, Marisa Jaconi, Ying Jin, Susan Kimber, Sergey Kiselev, Barbara B. Knowles, Oded Kopper, Valeri Kukharenko, Anver Kuliev, Maria A. Lagarkova, Peter W. Laird, Majlinda Lako, Andrew L. Laslett, Neta Lavon, Dong Ryul Lee, Jeoung Eun Lee, Chunliang Li, Linda S. Lim, Tenneille E. Ludwig, Yu Ma, Edna Maltby, Ileana Mateizel, Yoav Mayshar, Maria Mileikovsky, Stephen L. Minger, Takamichi Miyazaki, Shin Yong Moon, Harry Moore, Christine Mummery, Andras Nagy, Norio Nakatsuji, Kavita Narwani, Steve K. W. Oh, Sun Kyung Oh, Cia Olson, Timo Otonkoski, Fei Pan, In-Hyun Park, Steve Pells, Martin F. Pera, Lygia V. Pereira, Ouyang Qi, Grace Selva Raj, Benjamin Reubinoff, Alan Robins, Paul Robson, Janet Rossant, Ghasem H. Salekdeh, Thomas C. Schulz, Karen Sermon, Jameelah Sheik Mohamed, Hui Shen, Eric Sherrer, Kuldip Sidhu, Shirani Sivarajah, Heli Skottman, Claudia Spits, Glyn N. Stacey, Raimund Strehl, Nick Strelchenko, Hirofumi Suemori, Bowen Sun, Riitta Suuronen, Kazutoshi Takahashi, Timo Tuuri, Parvathy Venu, Yuri Verlinsky, Dorien Ward-van Oostwaard, Daniel J. Weisenberger, Yue Wu, Shinya Yamanaka, Lorraine Young, and Qi Zhou. Screening ethnically diverse human embryonic stem cells identifies a chromosome 20 minimal amplicon conferring growth advantage. *Nature Biotechnology*, 29(12):1132–1144, 2011.

[173] Nathalie Lefort, Maxime Feyeux, Cécile Bas, Olivier Féraud, Annelise Bennaceur-Griscelli, Gerard Tachdjian, Marc Peschanski, and Anselme L. Perrier. Human

embryonic stem cells reveal recurrent genomic instability at 20q11.21. *Nature Biotechnology*, 26(12):1364–1366, 2008.

[174] Stuart Avery, Adam J. Hirst, Duncan Baker, Chin Yan Lim, Sharmini Alagaratnam, Rolf I. Skotheim, Ragnhild A. Lothe, Martin F. Pera, Alan Colman, Paul Robson, Peter W. Andrews, and Barbara B. Knowles. BCL-XL Mediates the Strong Selective Advantage of a 20q11.21 Amplification Commonly Found in Human Embryonic Stem Cell Cultures. *Stem Cell Reports*, 1(5):379–386, 2013.

[175] Andrew S. Lee, Chad Tang, Mahendra S. Rao, Irving L. Weissman, and Joseph C. Wu. Tumorigenicity as a clinical hurdle for pluripotent stem cell therapies. *Nature Medicine*, 19(8):998–1004, 2013.

[176] Daisuke Doi, Asuka Morizane, Tetsuhiro Kikuchi, Hirotaka Onoe, Takuya Hayashi, Toshiyuki Kawasaki, Makoto Motono, Yoshiki Sasai, Hidemoto Saiki, Masanori Gomi, Tatsuya Yoshikawa, Hideki Hayashi, Mizuya Shinoyama, Mohamed M. Refaat, Hirofumi Suemori, Susumu Miyamoto, and Jun Takahashi. Prolonged Maturation Culture Favors a Reduction in the Tumorigenicity and the Dopaminergic Function of Human ESC-Derived Neural Cells in a Primate Model of Parkinson's Disease. *Stem Cells*, 30(5):935–945, 2012.

[177] Ninette Amariglio, Abraham Hirshberg, Bernd W. Scheithauer, Yoram Cohen, Ron Loewenthal, Luba Trakhtenbrot, Nurit Paz, Maya Koren-Michowitz, Dalia Waldman, Leonor Leider-Trejo, Amos Toren, Shlomi Constantini, and Gideon Rechavi. Donor-Derived Brain Tumor Following Neural Stem Cell Transplantation in an Ataxia Telangiectasia Patient. *PLoS Medicine*, 6(2):e1000029, 2009.

[178] Ken Garber. RIKEN suspends first clinical trial involving induced pluripotent stem cells. *Nature Biotechnology*, 33(9):890–892, 2015.

[179] Nataliia V. Katolikova, Aleksandr A. Khudiakov, Daria D. Shafranskaya, Andrey D. Prjibelski, Alexey E. Masharskiy, Mikael S. Mor, Alexey S. Golovkin, Anastasia K. Zaytseva, Irina E. Neganova, Evgeniya V. Efimova, Raul R. Gainetdinov, and Anna B. Malashicheva. Modulation of Notch Signaling at Early Stages of Differentiation of Human Induced Pluripotent Stem Cells to Dopaminergic Neurons. *International Journal of Molecular Sciences*, 24(2):1429, 2023.

[180] G. Morata. Cell competition: A historical perspective. *Developmental Biology*, 476:33–40, 2021.

[181] Michael E. Baumgartner and Eugenia Piddini. Mechanical cell competition in human pluripotent stem cell cultures. *Developmental Cell*, 56(17):2401–2402, 2021.

[182] Shi Chen, Long Huang, Ge Li, Funan Qiu, Yaodong Wang, Can Yang, Jingjing Pan, Zhangwei Wu, Jiangzhi Chen, and Yifeng Tian. LncRNA STXBP5-AS1 suppresses stem cell-like properties of pancreatic cancer by epigenetically inhibiting neighboring androglobin gene expression. *Clinical Epigenetics*, 12(1):168, 2020.

[183] Jeannie T. Lee. Epigenetic Regulation by Long Noncoding RNAs. *Science*, 338(6113):1435–1439, 2012.

[184] Masatoshi Ohgushi, Maki Minaguchi, and Yoshiki Sasai. Rho-Signaling-Directed YAP/TAZ Activity Underlies the Long-Term Survival and Expansion of Human Embryonic Stem Cells. *Cell Stem Cell*, 17(4):448–461, 2015.

[185] Ju-Won Jang, Min-Kyu Kim, and Suk-Chul Bae. Reciprocal regulation of YAP/TAZ by the Hippo pathway and the Small GTPase pathway. *Small GTPases*, 11(4):280–288, 2020.

[186] Tito Panciera, Luca Azzolin, Michelangelo Cordenonsi, and Stefano Piccolo. Mechanobiology of YAP and TAZ in physiology and disease. *Nature Reviews Molecular Cell Biology*, 18(12):758–770, 2017.

[187] Francesca Zanconato, Michelangelo Cordenonsi, and Stefano Piccolo. YAP/TAZ at the Roots of Cancer. *Cancer Cell*, 29(6):783–803, 2016.

[188] Cheston Hsiao, Michael Lampe, Songkhun Nillasithanukroh, Wenqing Han, Xiaojun Lian, and Sean P. Palecek. Human pluripotent stem cell culture density modulates YAP signaling. *Biotechnology Journal*, 11(5):662–675, 2016.

[189] Christopher J. Price, Dylan Stavish, Paul J. Gokhale, Ben A. Stevenson, Samantha Sargeant, Joanne Lacey, Tristan A. Rodriguez, and Ivana Barbaric. Genetically variant human pluripotent stem cells selectively eliminate wild-type counterparts through YAP-mediated cell competition. *Developmental Cell*, 56(17):2455–2470.e10, 2021.

[190] Margarida Sancho, Aida Di-Gregorio, Nancy George, Sara Pozzi, Juan Miguel Sánchez, Barbara Pernaute, and Tristan A. Rodríguez. Competitive Interactions Eliminate Unfit Embryonic Stem Cells at the Onset of Differentiation. *Developmental Cell*, 26(1):19–30, 2013.

[191] Romain Levayer and Eduardo Moreno. *How to Be in a Good Shape? The Influence of Clone Morphology on Cell Competition*, volume 9. Taylor and Francis Inc., 2016.

[192] Hans Clevers. The cancer stem cell: Premises, promises and challenges. *Nature Medicine*, 17(3):313–319, 2011.

[193] Yuxuan Richard Xie, Varsha K. Chari, Daniel C. Castro, Romans Grant, Stanislav S. Rubakhin, and Jonathan V. Sweedler. Data-Driven and Machine Learning-Based Framework for Image-Guided Single-Cell Mass Spectrometry. *Journal of Proteome Research*, 22(2):491–500, 2023.

[194] Richard O Duda, Peter E Hart, and David G Stork. Part 1: Pattern Classification. *Pattern classification and scene analysis*, 3, 1973.

[195] M. Mohammady, H. R. Moradi, H. Zeinivand, and A. J. a. M. Temme. A comparison of supervised, unsupervised and synthetic land use classification methods in the north of Iran. *International Journal of Environmental Science and Technology*, 12(5):1515–1526, 2015.

[196] Guo Yiqiang, Wu Yanbin, Ju Zhengshan, Wang Jun, and Zhao Luyan. Remote sensing image classification by the Chaos Genetic Algorithm in monitoring land use changes. *Mathematical and Computer Modelling*, 51(11):1408–1416, 2010.

[197] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 144–152. Association for Computing Machinery, 1992.

[198] M. Eren Ahsen, Nitin K. Singh, Todd Boren, M. Vidyasagar, and Michael A. White. A new feature selection algorithm for two-class classification problems and application to endometrial cancer. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 2976–2982. IEEE, 2012.

[199] Yiwen Bao, Lufeng Wang, Fei Yu, Jie Yang, and Dongya Huang. Parkinson's Disease Gene Biomarkers Screened by the LASSO and SVM Algorithms. *Brain Sciences*, 13(2):175, 2023.

[200] C. Li, X. Dong, Q. Yuan, G. Xu, Z. Di, Y. Yang, J. Hou, L. Zheng, W. Chen, and G. Wu. Identification of novel characteristic biomarkers and immune infiltration profile for the anaplastic thyroid cancer via machine learning algorithms. *Journal of Endocrinological Investigation*, 2023.

[201] Sanmay Das. Filters, wrappers and a boosting-based hybrid for feature selection. In *Icml*, volume 1, pages 74–81. Citeseer, 2001.

[202] Maher Maalouf. Logistic regression in data analysis: An overview. *International Journal of Data Analysis Techniques and Strategies*, 3(3):281–299, 2011.

[203] G. S. Watson. Generalized Linear Models (P. Mccullagh and J. A. Nelder). *SIAM Review*, 28(1):128–130, 1986.

[204] Peter Karsmakers, Kristiaan Pelckmans, and Johan AK Suykens. Multi-class kernel logistic regression: a fixed-size implementation. In *2007 International Joint Conference on Neural Networks*, pages 1756–1761. IEEE, 2007.

[205] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[206] Víctor Gómez. Wiener–Kolmogorov Filtering and Smoothing for Multivariate Series With State–Space Structure. *Journal of Time Series Analysis*, 28(3):361–385, 2007.

[207] C.K.I. Williams and D. Barber. Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351, 1998.

[208] R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[209] Maja Pohar, Mateja Blas, and Sandra Turk. Comparison of logistic regression and linear discriminant analysis: a simulation study. *Metodoloski zvezki*, 1(1):143, 2004.

[210] J. Mathew, V. K. Jha, and G. S. Rawat. Application of binary logistic regression analysis and its validation for landslide susceptibility mapping in part of Garhwal Himalaya, India. *International Journal of Remote Sensing*, 28(10):2257–2275, 2007.

[211] Mei Chen, Zeming Wang, Shengpeng Jiang, Jian Sun, Li Wang, Narayan Sahoo, G. Brandon Gunn, Steven J. Frank, Cheng Xu, Jiayi Chen, Quynh-Nhu Nguyen, Joe Y. Chang, Zhongxing Liao, X. Ronald Zhu, and Xiaodong Zhang. Predictive performance of different NTCP techniques for radiation-induced esophagitis in NSCLC patients receiving proton radiotherapy. *Scientific Reports*, 12(1):9178, 2022.

[212] Sotiris B Kotsiantis, Ioannis Zaharakis, P Pintelas, et al. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1):3–24, 2007.

[213] Ruey-Hsia Li and Geneva G. Belford. Instability of decision tree classification algorithms. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 570–575. Association for Computing Machinery, 2002.

[214] Seungil Huh, Dai Fei Elmer Ker, Ryoma Bise, Mei Chen, and Takeo Kanade. Automated Mitosis Detection of Stem Cell Populations in Phase-Contrast Microscopy Images. *IEEE Transactions on Medical Imaging*, 30(3):586–596, 2011.

[215] Chia-Chen Hsu, Jiabao Xu, Bas Brinkhof, Hui Wang, Zhanfeng Cui, Wei E. Huang, and Hua Ye. A single-cell Raman-based platform to identify developmental stages of human pluripotent stem cell-derived neurons. *Proceedings of the National Academy of Sciences*, 117(31):18412–18423, 2020.

[216] Sakina M. Mota, Robert E. Rogers, Andrew W. Haskell, Eoin McNeill, Roland R. Kaunas, Carl A. Gregory, Maryellen L. Giger, and Kristen C. Maitland. Automated mesenchymal stem cell segmentation and machine learning-based phenotype classification using morphometric and textural analysis. *Journal of Medical Imaging*, 8(1):014503, 2021.

[217] Adam Witmer and Bir Bhanu. Generative Adversarial Networks for Morphological–Temporal Classification of Stem Cell Images. *Sensors*, 22(1):206, 2022.

[218] Dennis Gabor. Theory of Communication. *Journal of the Institute of Electrical Engineers*, 93(26):429–441, 1946.

[219] Benjamin X. Guan, Bir Bhanu, Prue Talbot, Sabrina Lin, and Nikki Weng. Comparison of texture features for human embryonic stem cells with bio-inspired multi-class support vector machine. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 4102–4106, 2014.

[220] Veronica Biga, Olívia M. Alves Coelho, Paul J. Gokhale, James E. Mason, Eduardo M. A. M. Mendes, Peter W. Andrews, and Daniel Coca. Statistical Texture-Based Mapping of Cell Differentiation Under Microfluidic Flow. In Andrea Bracciali, Giulio Caravagna, David Gilbert, and Roberto Tagliaferri, editors, *Computational Intelligence Methods for Bioinformatics and Biostatistics*, volume 10477, pages 93–106. Springer International Publishing, 2017.

[221] C. H. Waddington. Genetic Assimilation of the Bithorax Phenotype. *Evolution*, 10(1):1–13, 1956.

[222] Henry Joutsijoki, Markus Haponen, Jyrki Rasku, Katriina Aalto-Setälä, and Martti Juhola. Machine Learning Approach to Automated Quality Identification of Human Induced Pluripotent Stem Cell Colony Images. *Computational and Mathematical Methods in Medicine*, 2016:e3091039, 2016.

[223] Atena Zahedi, Vincent On, Sabrina C. Lin, Brett C. Bays, Esther Omaiye, Bir Bhanu, and Prue Talbot. Evaluating Cell Processes, Quality, and Biomarkers in Pluripotent Stem Cells Using Video Bioinformatics. *PLoS ONE*, 11(2):e0148642, 2016.

[224] Stuart Berg, Dominik Kutra, Thorben Kroeger, Christoph N. Straehle, Bernhard X. Kausler, Carsten Haubold, Martin Schiegg, Janez Ales, Thorsten Beier, Markus Rudy, Kemal Eren, Jaime I. Cervantes, Buote Xu, Fynn Beuttenmueller, Adrian Wolny, Chong Zhang, Ullrich Koethe, Fred A. Hamprecht, and Anna Kreshuk. Ilastik: Interactive machine learning for (bio)image analysis. *Nature Methods*, 16(12):1226–1232, 2019.

[225] Puri Catalina, Rosa Montes, Gertru Ligero, Laura Sanchez, Teresa de la Cueva, Clara Bueno, Paola E. Leone, and Pablo Menendez. Human ESCs predisposition to karyotypic instability: Is a matter of culture adaptation or differential vulnerability among hESC lines due to inherent properties? *Molecular Cancer*, 7(1):76, 2008.

[226] Casey O. DuBose, John R. Daum, Christopher L. Sansam, and Gary J. Gorbsky. Dynamic Features of Chromosomal Instability during Culture of Induced Pluripotent Stem Cells. *Genes*, 13(7), 2022.

[227] G.I. Webb and Z. Zheng. Multistrategy ensemble learning: Reducing error by combining ensemble learning techniques. *IEEE Transactions on Knowledge and Data Engineering*, 16(8):980–991, 2004.

[228] Peter Bühlmann. Bagging, Boosting and Ensemble Methods. In James E. Gentle, Wolfgang Karl Härdle, and Yuichi Mori, editors, *Handbook of Computational Statistics: Concepts and Methods*, Springer Handbooks of Computational Statistics, pages 985–1022. Springer, 2012.

[229] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[230] Yoav Freund, Robert Schapire, and Naoki Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

[231] Robert E Schapire et al. A brief introduction to boosting. In *Ijcai*, volume 99, pages 1401–1406. Citeseer, 1999.

[232] Siyamalan Manivannan, Wenqi Li, Shazia Akbar, Ruixuan Wang, Jianguo Zhang, and Stephen J McKenna. Hep-2 cell classification using multi-resolution local patterns and ensemble svms. In *2014 1st Workshop on Pattern Recognition Techniques for Indirect Immunofluorescence Images*, pages 37–40. IEEE, 2014.

[233] Zeynep Banu Ozger and Pınar Cihan. A novel ensemble fuzzy classification model in SARS-CoV-2 B-cell epitope identification for development of protein-based vaccine. *Applied Soft Computing*, 116:108280, 2022.

[234] Wei Fan, Haonan Peng, Siyin Luo, Chujie Fang, and Yuanyuan Li. SCEC: A Novel Single-Cell Classification Method Based on Cell-Pair Ensemble Learning. In *Intelligent Computing Theories and Application*, pages 433–444. Springer, Cham, 2021.

[235] Richard B. Dickinson and Robert T. Tranquillo. Optimal estimation of cell movement indices from the statistical analysis of cell tracking data. *AIChE Journal*, 39(12):1995–2010, 1993.

[236] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 2002.

[237] Michael W Browne. Cross-Validation Methods. *Journal of Mathematical Psychology*, 44(1):108–132, 2000.

[238] Yun Xu and Royston Goodacre. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *Journal of Analysis and Testing*, 2(3):249–262, 2018.

[239] Lu Xu, Ou Hu, Yuwan Guo, Mengqin Zhang, Daowang Lu, Chen-Bo Cai, Shunping Xie, Mohammad Goodarzi, Hai-Yan Fu, and Yuan-Bin She. Representative splitting cross validation. *Chemometrics and Intelligent Laboratory Systems*, 183:29–35, 2018.

[240] Jamal Ahmad and Maqsood Hayat. MFSC: Multi-voting based feature selection for classification of Golgi proteins by adopting the general form of Chou's PseAAC components. *Journal of Theoretical Biology*, 463:99–109, 2019.

[241] Yawei Li and Yuan Luo. Performance-weighted-voting model: An ensemble machine learning method for cancer type classification using whole-exome sequencing mutation. *Quantitative Biology*, 8(4):347–358, 2020.

[242] Matheus Henrique Dal Molin Ribeiro and Leandro dos Santos Coelho. Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. *Applied Soft Computing*, 86:105837, 2020.

[243] Elaine Fuchs, Tudorita Tumbar, and Geraldine Guasch. Socializing with the Neighbors: Stem Cells and Their Niche. *Cell*, 116(6):769–778, 2004.

[244] Anshuman Singh, C. B. Yadav, N. Tabassum, A. K. Bajpeyee, and V. Verma. Stem cell niche: Dynamic neighbor of stem cells. *European Journal of Cell Biology*, 98(2):65–73, 2019.

[245] Stephen A Bustin and Tania Nolan. Pitfalls of quantitative real-time reverse-transcription polymerase chain reaction. *Journal of biomolecular techniques: JBT*, 15(3):155, 2004.

[246] Gary Warnes. Cell cycle analysis. http://www.icms.qmul.ac.uk/flowcytometry/uses/cellcycleanalysis/cellcycle/, 2019.

[247] David Sinclair. S-hull: A fast radial sweep-hull routine for Delaunay triangulation. *arXiv*, 2016.

[248] Dylan Stavish, Charlotta Böiers, Christopher Price, Thomas J. R. Frith, Jason Halliwell, Ingrid Saldaña-Guerrero, Jason Wray, John Brown, Jonathon Carr, Chela James, Ivana Barbaric, Peter W. Andrews, and Tariq Enver. Generation and trapping of a mesoderm biased state of human pluripotency. *Nature Communications*, 11(1):4989, 2020.

[249] Robert M. May. Simple mathematical models with very complicated dynamics. *Nature*, 261(5560):459–467, 1976.

Appendix A

Methods for Cell Biology

A.1 Culture of Human Pluripotent Stem Cells

A.1.1 Maintaining Human Pluripotent Stem Cells in Culture

Human pluripotent stem cells were fed either daily with Essential 8 (E8) or mTesR cell culture media, or every other day with Stable Essential 8 (S8) cell culture media. Prior to cell feeding, cell culture media was warmed to 37°C in a water bath. Between 2-5ml of media was added to cell cultures depending on confluency. Old media was aspirated using a glass stripette prior to adding new media. Flasks of cells were kept at 37°C with 5% CO_2 in the air (mimicking conditions within the human body) using an incubator.

A.1.2 Passaging of Human Pluripotent Stem Cells

Culture media was aspirated from flasks of hPSCs, then 0.5ml ReLeSR (STEMCELL Technologies, 05873) was added for 30 seconds. After aspirating excess ReLeSR, hPSC flasks were left at room temperature for 4 minutes. Then, warm cell culture media (S8 or mTeSR) was added to the flasks of hPSCs. Flasks were tapped until cells began to detach from the surface, then the cell solution was taken up and down in a 10ml stripette to mix the cells evenly throughout the media. Care was taken to ensure colonies were not broken apart. The cell solution was then split between flasks that had been pre-coated with Vitronectin (VTN) recombinant human protein (ThermoFisher Scientific A14700) at desired split ratios. Flasks were kept horizontal and shaken both up/down and left/right to improve cell distribution.

A.1.3 Dissociating Human Pluripotent Stem Cells to Single Cells

Culture media was aspirated from flasks of hPSCs, then cells were washed with Phosphate Buffer Saline (PBS). 1ml of Accutase was then added to the hPSC flasks which were incubated at 37°C for 4-5 minutes depending on cell confluency. The flask was kept upright to facilitate cell detachment. Cells were removed from the incubator and 4-5ml warm DMEM-F12 was added to the flask. Flask contents were transferred to a 15ml falcon tube. A 10µm sample was extracted from the cell solution and a cell count was taken using a haemocytometer. The remaining cells were spun in a balanced centrifuge at 800rpm for 3 minutes. After this, the tube containing cells was returned to the hood and the supernatent was aspirated. The remaining cell pellet was gently agitated to break up any clumps of cells. 1ml warm DMEM-F12 was added to the tube.

A.1.4 Seeding Human Pluripotent Stem Cells

Culture vessels were coated with 1% Geltrex in DMEM-F12, then left at room temperature for a minimum of 1 hour. Informed by the cell count, media supplemented with 10% Y-27632 was added to the tube to yield the desired cell concentration. Cells were then were incubated at 37°C with 5% CO2 for 24h. After 24h, cells were fed with culture media.

A.1.5 Transfecting Human Pluripotent Stem Cells

hPSCs were transfected with the pCAG-H2B-RFP-IRES-PURO plasmid at a concentration of $2.27\,\mu g/\mu L$ using electroporation. A total of 2.5 $\times 10^6$ cells were used. Prior to electroporation, 3 wells of a 6-well plate were coated with Geltrex. Transfections were performed using the Neon Transfection system (Thermo Fisher Scientific, MPK10025). First, 3mL of E2 buffer was added to the neon tube and placed in the electroporation device. The cells were then dissociated to single cells (Appendix A.1.3) (Thermo Fisher Scientific, 12504013), and a cell count was taken using a haemocytometer. The cells were centrifuged for 3 minutes at 1100RPM and the cell pellet resuspended to a concentration of 2.5 $\times 10^6$ cells/$120\,\mu L$ of resuspension buffer. Next, $120\,\mu L$ of cell suspension was mixed with up to $5\,\mu g$ of plasmid, keeping the volume of the plasmid to less than $5\,\mu L$. Pre-warmed mTeSR media (STEMCELL Technologies, 85850) supplemented with $10\,\mu m$ Y-27632 was added to the 3 coated wells of the 6-well plate in preparation of cell seeding. To perform the electroporation, $100\,\mu L$ of the plasmid/cell solution was taken using a Neon $100\,\mu L$ tip. The neon tip was placed into the E2 buffer and electroporated using the following conditions: 1600V, 20msec, 1 pulse. Cells were then resuspended in the pre-warmed cell culture media and returned to 37°C incubation.

Nuclear transfection was performed on H9TVD3 hESCs, passage number P29+4+ 4+4+3+3 at a 1:10 ratio. The purpose of this was to enable cell nuclei to be easily seen at all times using the RFP fluorophore. Then, computational cell segmentation would be straightforward. The antibiotic puromycin (InvivoGen ant-pr-1) was used for cell selection at 10mg/ml. For cloning the transfected cell line, 1:10 CloneR (STEMCELL Technologies, 05888) was used in mTeSR media, plus gentamicin (Thermo Fisher Scientific, 15710064) at 1:1000. Transfected cells were fed with mTeSR media for 2 weeks, then transitioned to S8 media gradually. Quantitative polymerase chain reaction (qPCR) analysis revealed no abnormalities in the transfected cell DNA. For more sensitive chromosomal analysis, P+5 cells were sampled and sent for karyotyping at the Sheffield Teaching Hospital on 12/03/20. Unfortunately, the onset of the first covid-19 pandemic lockdown shortly after this date meant that karyotyping results were not returned.

A.2 Immuno-Histochemistry

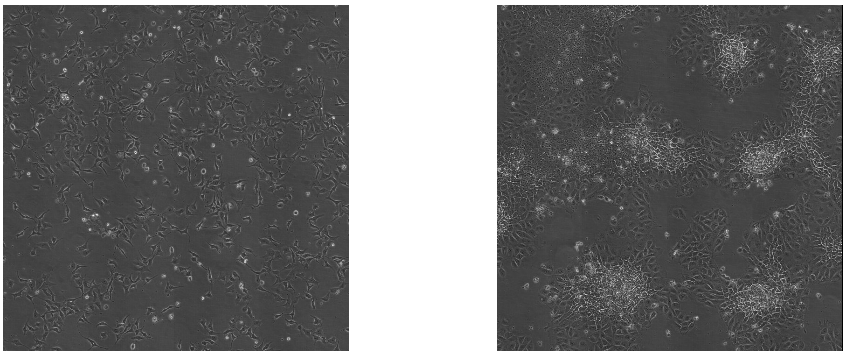
A.2.1 Fixing Cells with Paraformaldehyde

Prior to fixing, media was aspirated from the culture flask and washed 5 times with 1-2ml of PBS depending on flask size. Paraformaldehyde (PFA) was used as the fixative for cell cultures. 1ml of 4% PFA in PBS was added to the culture flask, which was applied to cells for 10-15 minutes and then aspirated. The culture flask was again washed 5 times with with 1-2ml of PBS. 1ml PBS was added and left in the cell culture to prevent dehydration.

A.2.2 Cell Staining for Surface Antigens

Cells were blocked for 1h by adding wash buffer, consisting of PBS supplemented with 10% fetal calf serum (FCS). If permeabilisation of cells was needed, 0.3% of Triton was also added to the wash buffer. Wash buffer was stored at 4°C. After this, the desired primary antibody (AB) was added to the cell culture at the necessary ratio with the wash buffer. This was kept overnight at 4°C. The next day, the primary AB was aspirated, and the culture was washed 4 times with wash buffer. The corresponding secondary AB was added to the cell culture at the necessary ratio with the wash buffer, plus Hoechst at the ratio 1:5000 with the wash buffer. Cells were incubated at 4°C in the dark for at least 2h. The cell culture was gently rocked by a rocker to homogeneously distribute the antibody. Upon aspiration of the secondary antibody, cells were washed 4 times with wash buffer in the dark. 1ml wash buffer was added and left in the cell culture to prevent dehydration. For fluorophore wavelength λ, the following primary/secondary AB combinations were used for cell staining, with given ratios of the AB:wash buffer solution:

- Hoechst: 1st N/A, 2nd N/A, $\lambda = 452$ (DAPI),
- Brachyury: 1st N/A, 2nd N/A, $\lambda = 507$ (FITC),
- Nanog: 1st rabbit (1:200), 2nd goat anti-rabbit (1:200), $\lambda = 647$ (Cy5),
- Sox2: 1st mouse (1:500), 2nd goat anti-mouse (1:200), $\lambda = 594$ (Cy3),



(a) 4h after seeding.

(b) 20h after seeding.

Figure A.1: Brightfield images of H9TVD3 hESCs time-lapsed using a BioStation.

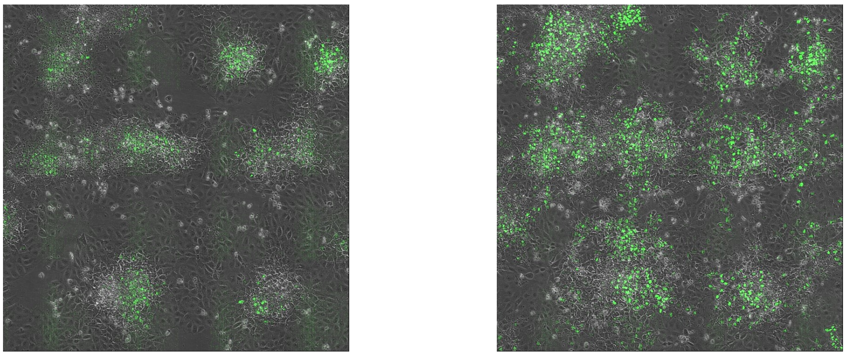
- Sox17: 1st goat (1:500), 2nd donkey anti-goat (1:200), $\lambda = 594$ (Cy3).

Note that Brachyury staining was not necessary for the H9TVD3 reporter line.

A.3 Time-Lapse Imaging of Live Cells

Time-lapse imaging was performed using a BioStation (Nikon). Protocols were optimised for computational object segmentation and tracking for single cells. Phototoxicity of imaging lasers was accounted for. Cell cultures were imaged every 4 minutes at 20x magnification. For Cy3 and Cy5 immunofluorescence images, 30 exposure was used. For FITC, 60 exposure was used. Media was supplemented with 10% Y-27632. If fluorescence imaging was needed from the start of the experiment, cells were imaged for only 24h due to photo-toxicity of imaging lasers. Otherwise, cells were time-lapse imaged for 48h.

To determine single-cell evolution of Brachyury expression, the H9TVD3 hESC line was used, as Brachyury expression is reported via nuclear GFP. Dissociated cells were seeded in S8 media supplemented with 10% Y-27632 and grown for 20h. A BioStation was used for brightfield time-lapse imaging at 20x magnification. Then, media was swapped to S8 supplemented with CHIR99021 (CHIR) (STEMCELL Technologies, 72054) at the ratio 1:2000. CHIR is a Wnt pathway inhibitor that forces mesodermal differentiation; thus, cells were expected to express Brachyury. Cells were imaged for a further 48h with 10 exposure for the GFP channel. Figure A.1 shows frames of H9TVD3 cells at different intervals after seeding. After cells were differentiated with CHIR, Brachyury expression was investigated via GFP expression, as shown in Figure A.2.



(a) 24h after differentiation (44h after seeding).

(b) 48h after differentiation (68h after seeding).

Figure A.2: Fluorescence images of H9TVD3 hESCs time-lapsed using a BioStation. GFP shows nuclear Brachyury expression.

A.4 Differentiation of Human Pluripotent Stem Cells

A.4.1 Preparation of BMP4

Recombinant human bone morphogenetic protein 4 (BMP4) (ThermoFisher Scientific PHC9534) was centrifuged to bring contents to the bottom of the vessel. Next, BMP4 was reconstituted in 4mM hydrochloric acid (HCl) with 0.1% bovine serum albumin (BSA), adding BSA prior to adding HCl. The resultant solution was split into 200ul aliquots and stored at -18°C.

A.4.2 BMP4 Supplementation in E8, S8, and mTeSR Media

BMP4 stock was removed from the freezer and thawed at room temperature. Four different BMP4 concentrations were used for this work: 50, 100, 150, and 200ng/ml. If the cell culture was being seeded, or was being fed up to 48h after seeding, then media was supplemented with 10% Y-27632.

A.5 Hoechst Histogram for the Cell Cycle

Hoechst intensity of an individual stem cell can be used to identify the cell cycle phase [87, 88], with two of the four phases known to promote pluripotency [89]. Specifically, the S and G2 phases positively regulate the pluripotent state, but only cells in the G1 phase are receptive to molecular signalling, while cells in the M phase are 'at rest'. It has been shown that the differentiation propensity of stem cells is coupled to phases of the cell cycle [90, 91, 92].

A histogram showing Hoechst (DAPI) intensity across the cell cycle is given in Figure

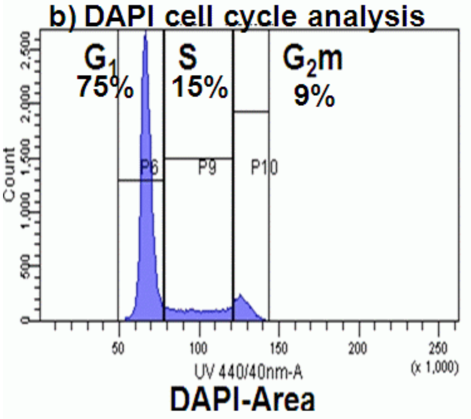


Figure A.3: Figure 4(b) from Warnes (2019): Cell line H357 were fixed and processed according to the DAPI labelling protocol listed on this webpage. Single cells were gated via DAPI Width and Area Signals. Courtesy of Luke Gammon, Centre for Cutaneous Research, ICMS, Barts & The Royal London, London University, 4 Newark Street, London E1 2AT.

A.3, which is taken from the literature [246]. Note that phases G2 and M are combined into G2m, as cells at rest in M phase show a 'flat' Hoechst profile.

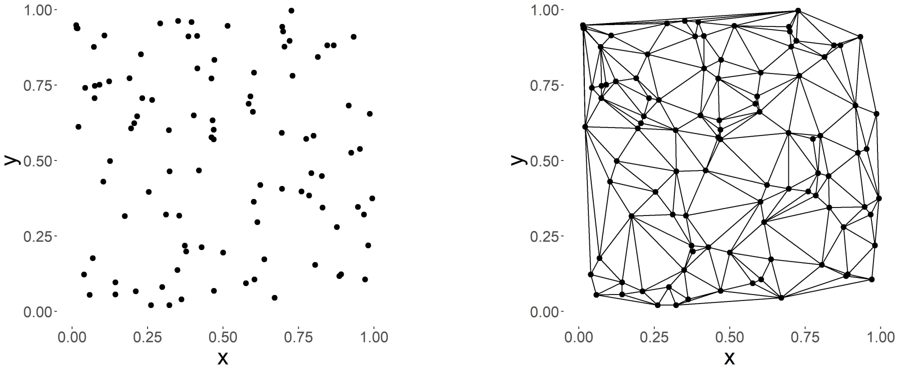
Appendix B

Delaunay Triangulation of Cell Nuclei in 2D

For a set of points $\mathbf{P}$ in 2-dimensional Euclidean space, a Delaunay triangulation $DT(\mathbf{P})$ is a triangulation such that no point in $\mathbf{P}$ lies within the circumcircle of any triangle in $DT(\mathbf{P})$. A Delaunay triangulation is not necessarily unique.

For this work, the algorithm for finding a Delaunay triangulation was the s-hull algorithm [247]. This leverages previous methods by using a radially propagating sweep-hull, along with triangle 'flipping' as the final step. For a set of unique points x_i in $\mathbb{R}^2$, the algorithm is as follows:

1. From x_i, select a seed point x_0.
2. Order the vector of points according to $\mid x_i - x_0 \mid^2$.
3. Find the point x_j which minimises the distance between x_i and x_j.
4. Find the point x_k that minimises the area of the circumcircle with x_0 and x_j, i.e. the unique circle which passes through the points x_0, x_j, x_k. Assign C as the centre of this circumcircle.
5. Order the points x_0, x_j, x_k to give a right-handed system. This is the initial seed convex hull.
6. Re-order the complete vector of points according to $\mid x_i - C \mid^2$. The output is the points s_i.
7. Sequentially add the points s_i to the propagating 2-dimensional convex hull that is seeded with the triangle formed by x_0, x_j, x_k. Each time a new point is added, the facets of the 2-dimensional hull that are visible to the point form new triangles.
8. A non-overlapping triangulation of the set of points x_i is now created.



(a) Scatter plot of 2-dimensional points.

(b) The same scatter plot as in (a), with additional Delaunay tiles generated through s-hull Delaunay triangulation.

Figure B.1: A randomly generated set of 100 points with (x, y) coordinates in the range (0,1). The purpose of this figure is to show the application of Delaunay triangulation to 2-dimensional spatial data, giving a Delaunay tiling of the plane.

9. Adjacent pairs of triangles must be 'flipped' to generate a Delaunay triangulation from the triangulation created in the previous step.

The triangle 'flipping' step can be described by considering two triangles ABC and BCD. Let these triangles share a common edge BD. Then the common edge must be 'flipped' such that it is instead the edge AC. The reason for this is that, continuing the example, the sum of the angles DAB and DCB must be less than or equal to $180°$ to meet Delaunay conditions. Additionally, a point must not lie within the interior of the circumcircle of ABC or BCD to meet Delaunay conditions. The s-hull algorithm first generates a triangulation which does not meet these conditions, but finally ensures that they are met by 'flipping' the common edge of triangles.

The centre of each cell's nucleus, represented as (x,y) coordinates, were used as vertices for Delaunay triangulation. An illustration of how Delaunay triangulation can be applied to tile spatial data in 2-dimensions - representative of cell nuclei in culture - is given by Figure B.1.

Appendix C

Predicting Binary Sox10 Expression via the Local Cellular Neighbourhood

Approaches to single-cell classification generally rely on auto-cellular features such as motility and morphology. In this chapter, an SVM classifier is trained to predict binary single-cell expression of the Sox10 gene in differentiated hESCs using only three neighbourhood features: (1) the cell's number of neighbours, (2) local cellular density, and (3) the linearly weighted sum of Sox10 expression from directly neighbouring cells. Using this small feature set, an excellent model accuracy of 92.7% is achieved. These results suggest that the local cellular environment should not be neglected when modelling gene expression, as single-cell fate behaviour cannot be decoupled from neighbourhood states. In particular, the impact of neighbouring cells appears based simply on linear distance.

C.1 Neural Crest Differentiation of H9TVD3 Cells in PRIMO Media

C.1.1 Data Overview

H9TVD3 hESCs were cultured in PRIMO Media (formulated by D. Stavish *et. al*) which promotes and 'traps' cells in a state primed for mesodermal fate [248]. PRIMO media consists of E8 basal medium plus 0.1% bovine serum albumin, 2μM cholesterol, 3μM CHIR, 1μM Wnt pathway inhibitor IWP-2, and 0.48μM lysophophatidic acid, the latter of which promotes pluripotency. Cells were then differentiated by I.M. Saldaña-Guerrero to yield neural crest. Subsequently, the cells were fixed (Appendix A.2.1) and stained (Appendix A.2.2) to investigate expression of Sox10 and Brachyury (T). Sox10 is a marker of neuro-ectoderm, whilst T marks mesodermal fate.

Four experimental conditions were used, varying whether Rock inhibitor (Y-27632) and

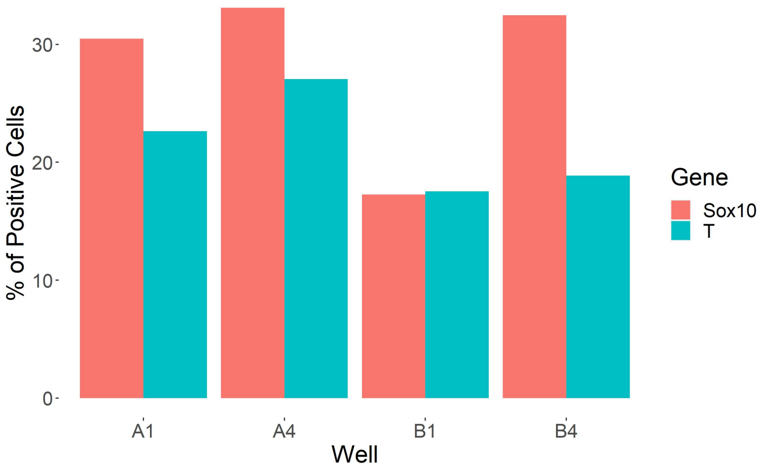


Figure C.1: Bar chart showing percentages of Sox10 and T positive cells per experimental condition, given by each well.

IWP-2 were supplemented in PRIMO Media. Each condition corresponded to a plate well:

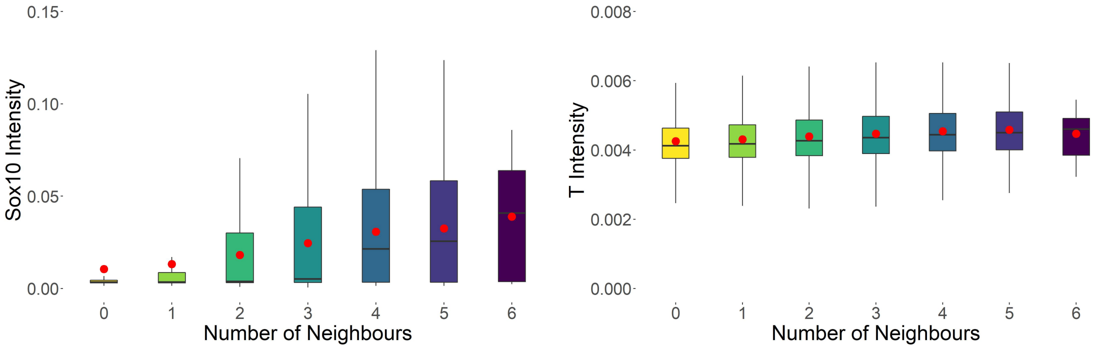
1. Well A1: with Y-27632 for 2 days and without IWP-2,
2. Well A4: without Y-27632 and without IWP-2,
3. Well B1: with Y-27632 and with IWP-2, both for 2 days,
4. Well B4: without Y-27632 and with IWP-2 for 2 days.

Cells were imaged using InCell technology, processed in ImageJ, then passed to CellProfiler for object segmentation and quantifying basic cell statistics. The output .csv file contained 86441 rows, each corresponding to a single cell. Counts of Sox10 and T positive cells per experimental condition are given in Figure C.1.

C.2 Feature Extraction from Raw Data

To calculate the number of neighbours, a radial distance measure was used. The distance threshold was taken as $d = 16.25\,\mu\text{m}$, as detailed in Chapter 2, Section 2.4.1. Importantly, if the central coordinates of cell's nucleus were within distance d of the image boundary, the cell was removed. This was done to reduce the possibility that neighbouring cells outside the image boundaries could confound results.

Using this method, the respective total sums of neighbours' Sox10 and T expression were calculated for each cell. Bar charts with means showing the relationship between normalised Sox10/T expression and number of neighbours are shown in Figure C.2. Intensity outliers were removed using the method given in Equations(4.2-4.4). This omission combined with removal of cells at image borders preserved 91.7% of the original data.



(a) Sox10.

(b) T.

Figure C.2: Boxplots showing Sox10 and T expression at the single-cell level vs. the cells' number of neighbours. Plots show results across all experimental conditions. Mean values are shown in red. Outliers were removed.

Local density was calculated per image according to the triangulation methodology given in Appendix B combined with Equation 4.1.

Next, four different weighting functions were calculated for the contribution of neighbouring cells' gene expression. Functions were defined such that weights were proportional to the distance from the cell of interest to its neighbour. Weights were calculated as following:

1. Weights all equal to 1 (unweighted),
2. Linear decay as distance increases,
3. Gaussian decay as distance increases,
4. Exponential decay as distance increases.

Weighting condition (1) corresponds to the single-cell sums of neighbours' Sox10 expression described above.

For calculating the weights corresponding to model weighting condition (2), i.e. linear decay with inter-cellular distance, a line with equation

$$y(x) = 1 - \left(\frac{1}{d+\epsilon}\right)x, \quad x = 0, ..., (d+\epsilon) \tag{C.1}$$

was defined, taking $\epsilon = 0.0001$. The parameter ϵ was incorporated to ensure that $y(d) \neq 0$, but rather $y(d+\epsilon) = 0$, as neighbours that have distance less than or equal to d of the cell of interest should not have zero weight. Importantly, this does not allow cells with greater distance than d from the cell to have non-zero weight.

For model weighting condition (3), i.e. Gaussian decay, the right plane of a normal dis-

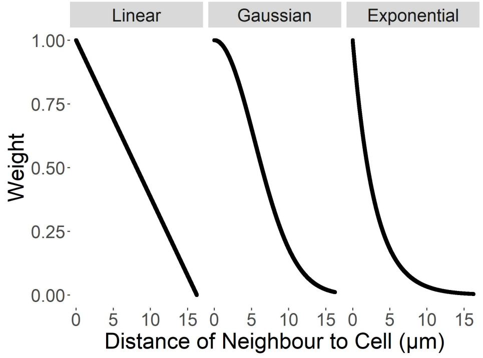


Figure C.3: Graphs of the linear, Gaussian, and exponential decay weighting functions used for weighting each cell's sum of Sox10 intensity from directly neighbouring cells.

tribution given by

$$y(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, \quad x = 0, ..., (d+\epsilon) \tag{C.2}$$

was used, with mean $\mu = 0$ and standard deviation $\sigma = d/3$. This distribution was then normalised in the range $[0, 1]$.

Exponential decay with equation

$$y(x) = Ae^{-\alpha x}, \quad x = 0, ..., (d+\epsilon) \tag{C.3}$$

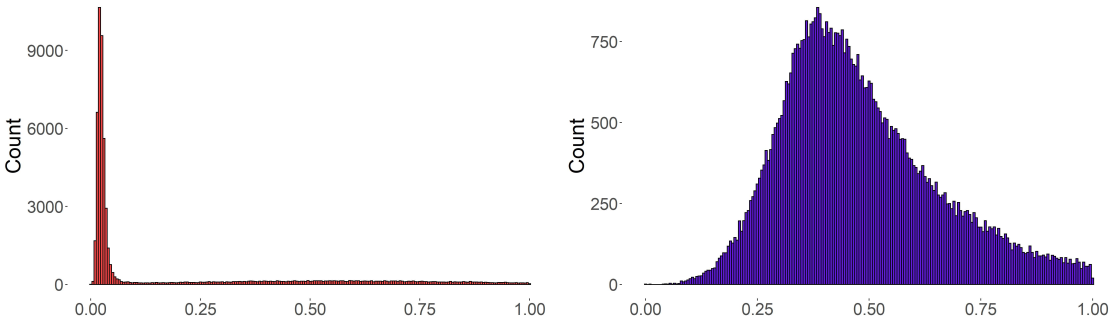
was used for model weighting condition (4), taking $A = 1$ and $\alpha = (-4\ln(0.25))/d$. This value of α was chosen such that $(1/4)\max(x) = (1/4)\max(y)$. In other words, the x-axis and y-axis achieve one quarter of their maximums at the same value.

See Figure C.3 for graphs of the linear, Gaussian, and exponential decay weighting functions used for this work.

C.2.1 Data Pre-Processing

From Figure C.2, it can be seen that there is very little variation in T expression with respect to the cell's number of neighbours. However, Sox10 expression increases on average proportionally to neighbour count. This is the inverse of the phenomena discovered for Nanog expression in hESCs described in Chapter 2. Interestingly, this relationship is seen across all data, i.e. across all experimental conditions. The Sox10/neighbours relationship was particularly prevalent for cells in Well A1, which were cultured with Y-27632 and without IWP-2. For this reason, the classification model solely focused on predicting Sox10 expression for cells in Well A1.

Cells with no neighbours were filtered from the data (as the models depend on non-empty information about each cell's local neighbourhood), thus preserving 68.5% of data



(a) Sox10.

(b) T.

Figure C.4: Histogram of intensities in the range $[0, 1]$ for H9TVD3 hESCs differentiated to neural crest. Outliers and cells with 0 neighbours were removed.

for Well A1. Intensities for Sox10 and T were re-scaled in the range $[0, 1]$, with the corresponding histograms given by Figure C.4.

C.3 Results of Sox10 Classification Modelling

C.3.1 Feature Selection by Preliminary Model Training

Features that were considered for model training were:

1. Number of neighbours ($\mathbb{Z} \geq 0$),
2. Local density ($\mathbb{R} \geq 0$),
3. Unweighted sum of neighbours' Sox10 intensities ($\mathbb{R} \geq 0$),
4. Linear weighted sum of neighbours' Sox10 intensities ($\mathbb{R} \geq 0$),
5. Gaussian weighted sum of neighbours' Sox10 intensities ($\mathbb{R} \geq 0$),
6. Exponential weighted sum of neighbours' Sox10 intensities ($\mathbb{R} \geq 0$),
7. Unweighted sum of neighbours' T intensities ($\mathbb{R} \geq 0$),
8. Linear weighted sum of neighbours' T intensities ($\mathbb{R} \geq 0$),
9. Gaussian weighted sum of neighbours' T intensities ($\mathbb{R} \geq 0$),
10. Exponential weighted sum of neighbours' T intensities ($\mathbb{R} \geq 0$).

The binary response variable for the classification model was whether a cell is Sox10 positive or negative. MATLAB was used to fit fine, medium, and coarse tree models and k-nearest neighbour (KNN) models to the data. These models had the advantage of fast

Features	Best Model	Accuracy (%)
All	Coarse KNN	93.1
(1),(2)	Medium Tree	69.2
(1),(2),(3)	**Coarse KNN**	**93.2**
(1),(2),(4)	Medium Tree	90.3
(1),(2),(5)	Coarse KNN	90.9
(1),(2),(6)	Coarse KNN	92.2
(1),(2),(7)	Coarse KNN	70.7
(1),(2),(8)	Coarse KNN	69.2
(1),(2),(9)	Coarse KNN	69.1
(1),(2),(10)	Medium Tree	69.2

Table C.1: Table displaying % accuracy of the best performing classification models. The model with maximal accuracy is shown in bold.

computing times, making it possible to run many different model set-ups. To reduce the problem of over-fitting, randomised 10-fold cross-validation was used for each model. The training set consisted of a randomly sampled 80% of the data, thus leaving 20% for testing.

For all but one combination of predictors, the best performing model was the coarse KNN. Using the number of neighbours and local density alone as training features resulted in a much lower percentage accuracy than for models incorporating neighbours' Sox10 expression. Models trained using neighbourhood T expression also fared comparatively poorly, but still produced reasonable accuracies with the limited model architectures used. See Table C.1 for the accuracy of all best performing models. Overall, the model with the highest accuracy was the model that used the cell's number of neighbours, local density, and the unweighted sum of neighbours' Sox10 expression to predict the response.

C.3.2 A Gaussian Kernel Support Vector Machine Model for Accurate Classification of Sox10 Expression

Informed by the above results, a Gaussian (non-linear) kernel support vector machine (SVM) was trained using features (1)-(3). An SVM model was selected as these are generally powerful for classification tasks, demonstrated in Chapter 5. As before, 10-fold cross-validation was implemented. At each fold split, 80% of the data was used for model training, and 10% for validating the model at that fold. Table C.2 shows these results, including the mean accuracy of 93.2% across all folds. Finally, after cross-validation was completed, the model was trained on 80% of the full dataset, and tested on the remaining 10% of data that had not been used at any point during cross-validation. The overall model accuracy achieved was 92.7%, with precision 0.89, recall 0.88, and F1-score 0.88.

Training Fold	Accuracy (%)
1	93.3
2	94.3
3	90.6
4	92.6
5	93.3
6	95.0
7	94.0
8	91.3
9	94.0
10	93.3
Mean	**93.2**

Table C.2: Sox10 classification results using an SVM model with Gaussian kernel. Cross-validation was implemented with 10 training folds. Accuracies are reported for the validation set at each fold.

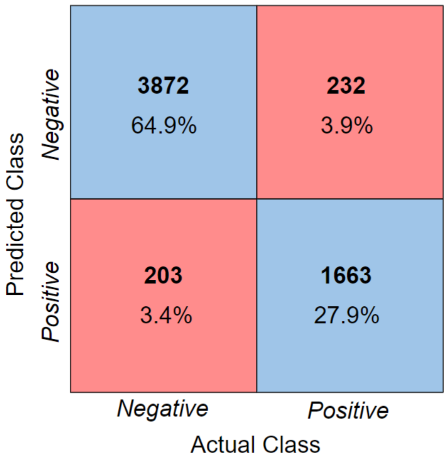


Figure C.5: Confusion matrix for the Gaussian kernel SVM classifier trained to predict whether hESCs are Sox10 positive/negative. Model accuracy = 92.7%, F1-score = 0.88.

Figure C.5 gives the confusion matrix for SVM model. To test significance of the model accuracy, the response variable was randomly shuffled, and the same SVM model was applied, using the shuffled response variable as the target. This process was repeated 5 times. The output models had a mean accuracy of 56.5%, suggesting that the original model accuracy of 92.7% is indeed meaningful.

C.4 Summary and Conclusions

In this appendix, a support vector machine (SVM) using a Gaussian kernel was trained to classify differentiated hESCs as either positive or negative for the expression of the Sox10 gene. Spatial features regarding the cellular neighbourhood were used for model training. Specifically, the following three features were used: (1) the cell's number of neighbours, (2) local cellular density, and (3) the linearly weighted sum of Sox10 expression from directly neighbouring cells. The classifier achieved an excellent accuracy of 92.7%.

These results show that expression of Sox10 in differentiated hESC cultures is heavily regulated by the local cellular environment. In particular, several weighting functions representing the contribution of neighbouring cells' Sox10 expression were calculated, which were used for model training prior to obtaining the final SVM. Both Gaussian and exponential decay functions (with respect to inter-cellular distance) performed worse for predicting Sox10 expression labels, as the resultant models yielded a lower accuracy than that of the final model. Additionally, using an unweighted sum of neighbours' Sox10 intensity for model training yielded poorer results than using the linear decay weighting function. Therefore, this may suggest that cell signalling responsible for orchestrating gene expression patterning in differentiating hESC populations occurs across linear distances between cells. This could be explained by both juxtacrine and diffusive signalling; it is likely that both of these signalling mechanisms are influential in single-cell gene regulation.

Appendix D

Custom Micro-Patterned Chip Designs for Cell Culture

To validate predictions for cell fate patterning output by the model presented in Chapter 3, three custom micro-patterned chips were designed. These were intended to be produced by the company CYTOO, who manufacture the commonly used CYTOOchip™Arena. During design, a good balance of geometries previously employed by experimental researchers and novel geometries was proposed. In particular, it was of interest to explore hollow shapes, fractals, and symmetric shapes that would force very simplistic neighbour relations between cells, such as the cross in Figure D.1. Additionally, varying the size of micro-patterned cultures, without altering the shape, has been demonstrated to impact resultant fate patterning. Therefore, various sizes of the same shape were incorporated into the designs to test this phenomenon.
Future work remains of interest to manufacture these micro-patterns, which are shown over the next three pages, and use them for hPSC differentiation experiments.

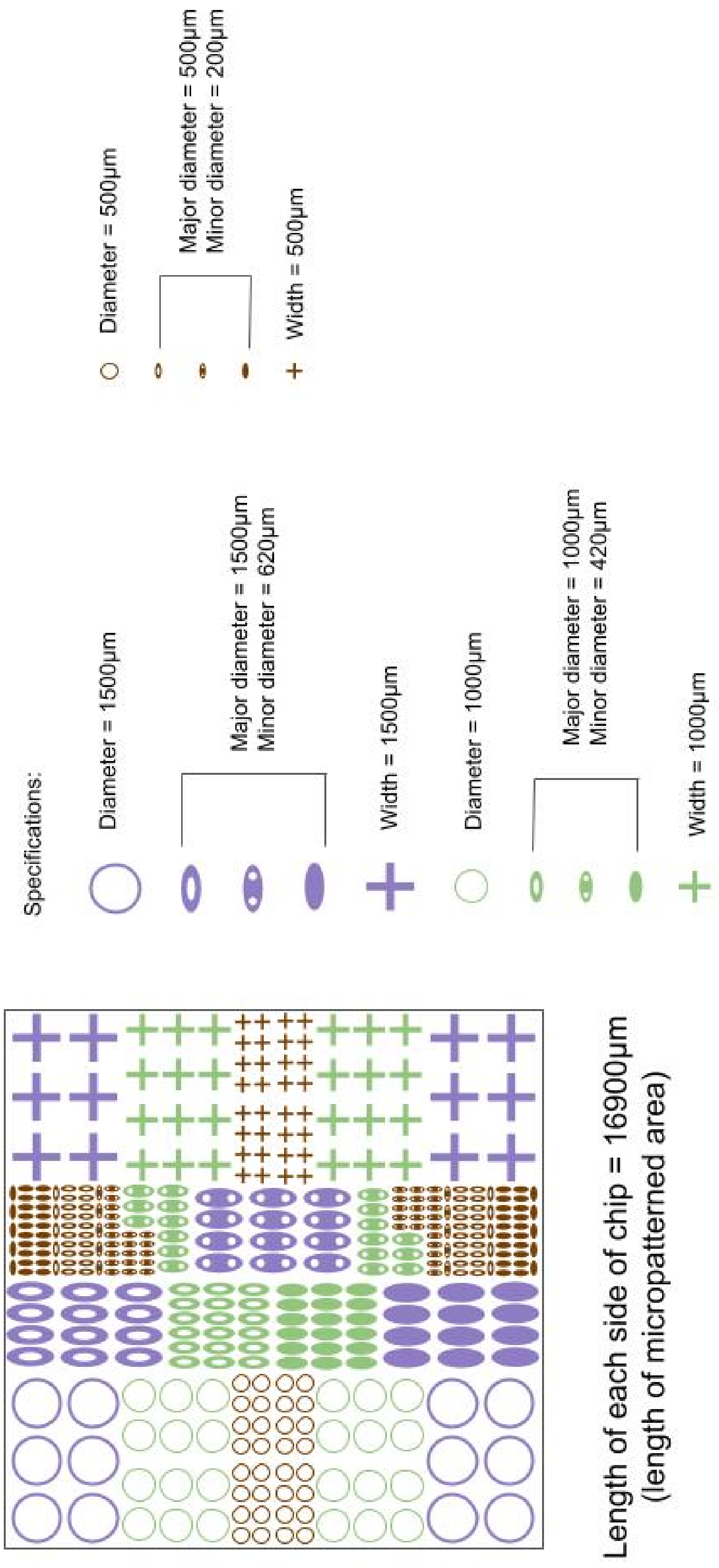


Figure D.1: Custom micro-patterned chip design 1.

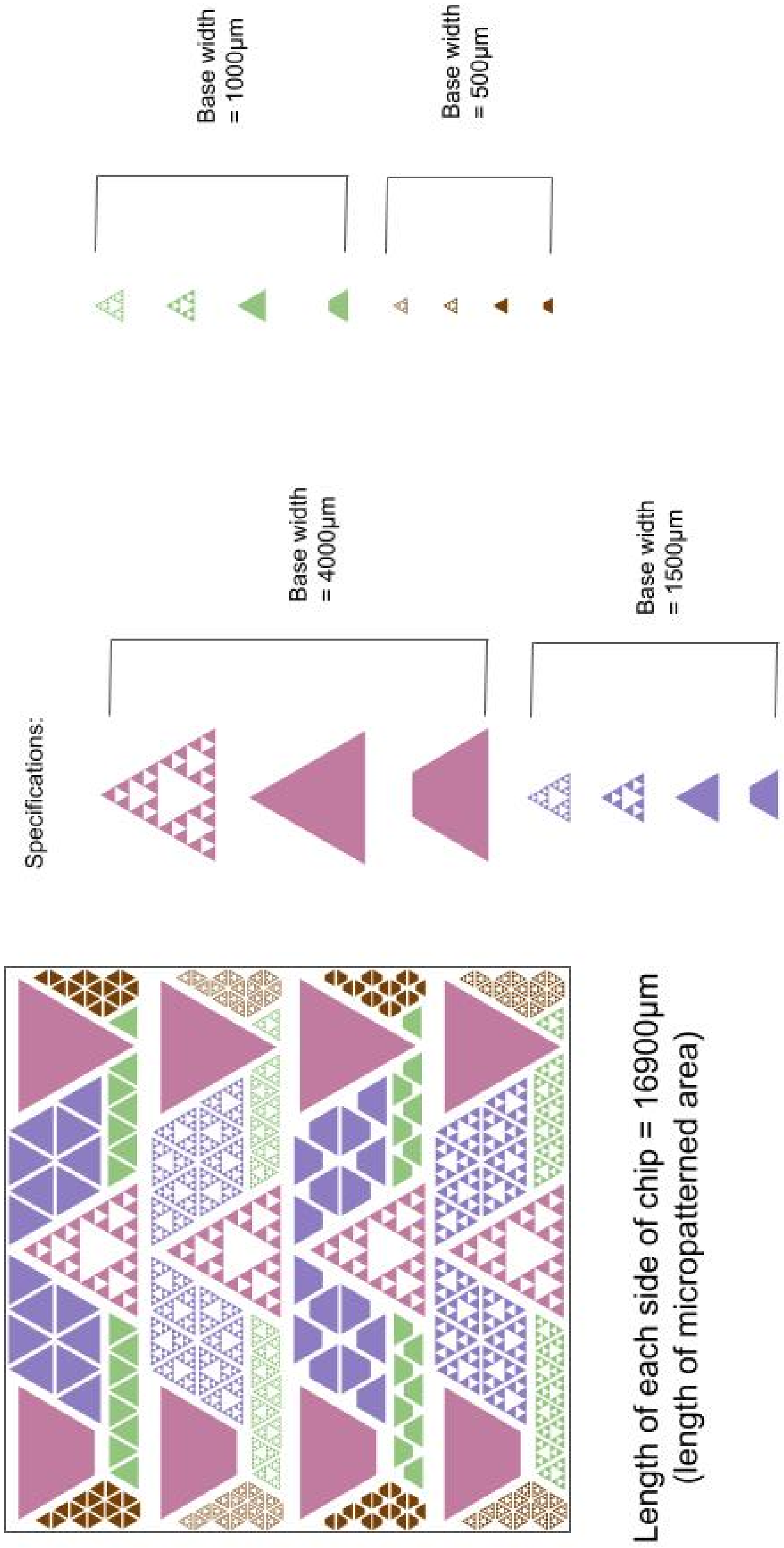


Figure D.2: Custom micro-patterned chip design 2.

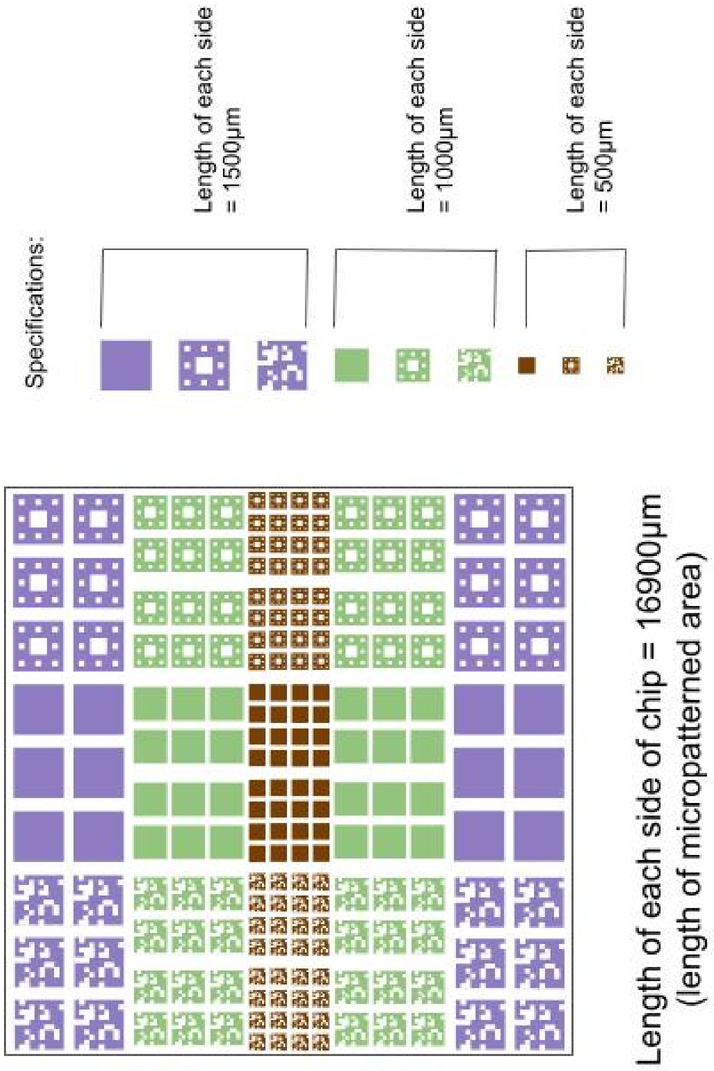

Specifications:
Length of each side = 1500µm
Length of each side = 1000µm
Length of each side = 500µm
Length of each side of chip = 16900µm (length of micropatterned area)

Figure D.3: Custom micro-patterned chip design 3.