THE UNIVERSITY
of EDINBURGH

New insights into probabilistic pattern formation of embryonic stem cells using agent-based modelling

Minhong Wang

Thesis in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Supervisors: Prof. Dave Robertson, Dr. Athanasios Tsanas, Dr. Guillaume Blin, and Dr. Saturnino Luz

The University of Edinburgh

2020

To my family, with love

To my teachers, with gratitude

Agent-based Modelling of Pattern Formation in Embryonic Stem Cells

Minhong Wang

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

Embryonic stem cells (ESCs) hold great potential for developing future therapies for a wide range of diseases. However, the mechanisms of pattern formation during embryonic development remain poorly understood. ESCs in culture self-organise to form spatial patterns of gene expression upon geometrical confinement indicating that patterning is an emergent phenomenon that results from the many interactions between the cells.

Here, we applied an agent-based modelling approach to identify biologically plausible rules acting at the mesoscale within stem cell collectives that may explain spontaneous patterning. We tested different models involving differential motile behaviours including exploring effects due to neighbour interactions. We introduced a new metric, the stem cell aggregate pattern distance (SCAPD), to assess the deviation between the probabilistic experimental pattern formation (used as ground truth) and the probabilistic simulated outcome. We demonstrated our models can produce broadly realistic pattern formation (when compared to experimental data) with a quantified level of uncertainty. The best of our models improve fitness, evaluated by SCAPD, by 70% and 77% over the random models for a discoidal or an ellipsoidal stem cell confinement, respectively. Collectively, our findings provide compelling arguments that a parsimonious mechanism that involves differential motility is sufficient to explain the spontaneous patterning of the cells upon confinement. Furthermore, our work also defines a region of the parameter space that is compatible with patterning, which assists future studies in the field of cell engineering. We envisage that the novel approaches explored within this work will be applicable to many biological systems and will contribute towards facilitating progress by reducing the need for extensive and costly experiments.

Acknowledgements

Throughout my PhD, I have received a great deal of support and assistance. I would like to thank the following people, without whom I would not have been able to complete this research.

First and foremost, I would like to express my deepest gratitude to all my supervisors Professor Dave Robertson, Dr Athanasios Tsanas, Dr Guillaume Blin and Dr Saturnino Luz. Dave, my primary supervisor, guided me throughout this study. He is always supportive with his valuable advice and immense knowledge. I am extremely grateful for his patience and guidance throughout my PhD journey. Athanasios is always helpful whenever I come to him. Even though we met in my second year, he continuously provided encouragement and was always willing and enthusiastic to assist in any way he could. He provided a lot of excellent advice for my study, especially in mathematical fields. I would like to say a big thank you to him from the bottom of my heart. I would also like to express my sincere gratitude to Guillaume, who supported the data for this study as well as provided tremendous advice in regards to biological aspects. I would like to thank him for his support and time.

I would like to extend my thanks to Dr Honghan Wu and Dr Areti Manataki for offering me opportunities for different skills training, including data analysis and teaching. I would also like to acknowledge the support received from my colleagues at Usher Institute. I enjoyed the great times we had and always find our great discussions inspiring. Special thanks to Dr Lowell Edgar for his help with proofreading my publications.

Finally, I am extremely grateful to my parents, my family, and all my friends. Without their enormous support and understanding, it would be impossible for me to complete my study.

Abbreviations

ABM	Agent-based Modelling
CA	Cellular Automata
CPM	Cellular Potts Model
CRPS	Continuous Ranked Probability Score
CSV	Comma Separated Values
DAH	Differential Adhesion Hypothesis
EBs	Embryoid Bodies
ECM	Extracellular Matrix
EMD	Earth Mover's Distance
ESCs	Embryonic Stem Cells
HDA	High-Density Area
hESCs	Human Embryonic Stem Cells
ICM	Inner Cell Mass
KDE	Kernel Density Estimation
KL divergence	Kullback-Leibler Divergence
mESCs	Mouse Embryonic Stem Cells
Nessys	Nuclear Envelop Segmentation System
SCAPD	Stem Cell Aggregate Pattern Distance
TST	Tissue Surface Tension

Frequently used notation

The following mathematical notational conventions are used throughout this thesis:

Scalars are written in italic lower case letters, for example x; random variables are written in italic capital letters, for example X. P stands for probability. In Horn clauses, the normal logical operators are being used. For example, the operators $\leftarrow$, $\wedge$, and $\vee$ are the connectives for implication, conjunction and disjunction. In addition, we also use [ ] to represent an empty list and $\{\}$ to represent an empty set.

List of figures

List of tables

Contents

Chapter 1

1 Introduction

Cells, the smallest unit capable of independent reproduction named by Robert Hooke in 1665, are the basic unit of all living things (Hooke, 1665). They have the metabolism to produce energy and keep them survive. They communicate with each other and sensing the signals from the environment, which allows them to collect the information and respond accordingly. Cells increase the number by division, the process that a parent cell divides into two or more daughter cells. As the human body is composed of trillions of cells, many different types of specialised cells carry out different functions for the body. Specialised cells (e.g. nerve cells, muscle cells, blood cells etc.,) make up tissues, tissues make up organs, and organs make up the systems that work together to make up human bodies (Bianconi *et al.*, 2013).

In contrast to specialised cells, stem cells are unspecialised cells, which means they are naïve cells as they have not yet developed to perform particular functions. Hence, stem cells hold the potential to differentiate to specialised cells. Cellular differentiation is the process of changing a cell from one cell type to another. For stem cells, *differentiation* is the procedure where a cell changes its key properties to specialize in a specific task (Slack, 2007).

Stem cells are divided into several categories according to their potential to differentiate. Adult stem cells (e.g. hematopoietic stem cells, mesenchymal stem cells, neural stem cells, epithelial stem cells etc.,) are undifferentiated cells found living within specific differentiated tissues. They can renew themselves and only can differentiate into limited cell types (Ramalho-Santos *et al.*, 2002). Different from adult stem cells, *pluripotent stem cells* have the ability to undergo self-renewal and differentiate into all cell types of the tissues

of the body (Romito and Cobellis, 2016). There are two types of pluripotent stem cells, *embryonic stem cells* (ESCs) and *induced pluripotent stem cells* (dedifferentiation of adult somatic cells in vitro through cell reprogramming). The overview of ESCs will be provided in Section 1.1 as well as more detailed existing biological studies will be described in Section 2.1.

Since ESCs can self-renew and can differentiate, they have enormous potential for developing many new treatments. However, with existing knowledge of stem cells, we cannot fully control their differentiation (Tewary, Shakiba and Zandstra, 2018) – which is a precondition of applying stem cell technology to the development of therapeutic treatments. One of the challenges of controlling stem cells to achieve new therapies is that we do not know the key cell behaviours which lead to the desired pattern formation (Kamm *et al.*, 2018).

In this study, we focus on studying the observed pattern formation in a specific type of stem cells to investigate their social behaviours. We generated a series of models to reproduce the pattern formation in order to take a step toward extending the current understanding of ESCs pattern formation. We developed an approach, which will be applicable to many biological systems, contributing to facilitating biological study progress by reducing the need for extensive and costly experiments.

1.1 Overview of embryonic stem cells

ESCs are defined as cells that can self-renew and differentiate into mature cells of any particular tissue. ESCs are stem cells derived from the undifferentiated inner cell mass (ICM) of an embryo at the blastocyst stage. The derivation of the ESCs was firstly achieved in mouse back in 1981 (Evans and Kaufman, 1981; Martin, 1981). After the derivation of mouse embryonic stem cells (mESCs), the derivation of ESCs from different species was achieved (Thomson *et al.*, 1995) including the isolation of human embryonic

stem cells (hESCs) (Thomson *et al.*, 1998). In 1998, Thomson et al. has allowed new comparative studies and have helped to reveal the conserved mechanisms that control *pluripotency* across species (Thomson *et al.*, 1998). Different from adult stem cells, ESCs are *pluripotent*. More specifically, they are able to differentiate into all derivatives of the three primary germ layers: ectoderm, endoderm and mesoderm and therefore they can develop into all cells of the adult body. Hence, ESCs hold a greater potential compared to adult stem cells, which can differentiate into limited types of cells. Furthermore, induced pluripotent stem cells, a type of embryonic-like cells reprogrammed from differentiated cells, were generated from mouse cells in 2006 (Takahashi and Yamanaka, 2006). Afterwards, the induction of pluripotent stem cells from human adult cells was achieved in 2007 (Takahashi *et al.*, 2007). Induced pluripotent stem cells provide a positive alternative to ESCs since it enables the development of unlimited source from skin or blood cells instead of obtaining ESCs from embryos. These promising findings strongly support the notion of regenerative medicine since the embryo is not the only source for pluripotent stem cells.

Due to the special properties of pluripotent stem cells and their tremendous potential for treating human diseases, regenerating desired tissues or organs from stem cells is undergoing intense study. For example, multiple types of stem cells have been used for the growth of new blood vessels in vascular regeneration, including ESCs and induced pluripotent stem cells (Leeper, Hunter and Cooke, 2010); pluripotent stem cell-based therapeutic strategies were also proposed for neural injury (e.g. spinal cord injury) as researchers are working toward the development of neuroprotective and regenerative interventions (Ronaghi *et al.*, 2009); researchers are also working on inducing differentiation of pluripotent stem cells into cardiomyocytes and these cells may lead to treatments of different types of heart disease (Freund and Mummery, 2009). In addition, people with different diseases might benefit from stem cell therapies, including people with cardiovascular diseases (Okano *et al.*, 2013), Parkinson's disease (Politis and Lindvall, 2012), diabetes (Meier,

Bhushan and Butler, 2006; Aguayo-Mazzucato and Bonner-Weir, 2010) and many other diseases as well (Wu and Hochedlinger, 2011).

In addition to regenerative medicine, ESCs are useful for understanding the molecular events that underlying stemness and cell lineage commitment (Gan *et al.*, 2007; Ivey *et al.*, 2008). Moreover, stem cell studies can help with the pathophysiological understanding of disease onset (Davis *et al.*, 2011; Avior, Sagi and Benvenisty, 2016). Stem cell studies provide opportunities to observe the process that stem cells mature into different types of cells and study how diseases and conditions develop. For example, hESCs may provide new insights into cancer research by revealing the genetic and epigenetic changes that occur during normal development, and therefore facilitate understanding of the oncogenic process (Nishikawa, Goldstein and Nierras, 2008). Besides discovering disease developments, stem cells can be used for testing new drugs for safety and effectiveness. For instance, stem cells can be programmed into tissue-specific cells or acquire specific properties for testing new drugs (Davidson, Ware and Khetani, 2015). To realise their great potential, the effectiveness of programming stem cells would be key to this achievement.

The differences between hESCs and mESCs is an area of immense research interest (Ginis *et al.*, 2004; Rao, 2004; Cheng *et al.*, 2017). Even though hESCs are different from mESCs in many respects, mESCs provide an opportunity to carry out stem cell studies at a lower cost and overcoming ethical concerns on the use of hESCs. Many studies were first performed in mESCs before being tested on hESCs (Okita, Ichisaka and Yamanaka, 2007; Aoi *et al.*, 2008). Hence, there is a large number of studies focusing on mESCs to investigate the pluripotency during early embryonic development, fundamental characteristics of ESCs and pluripotency factors (Pauklin, Pedersen and Vallier, 2011).

1.2 Statement of the problem

Even though ESCs hold great potential for investigating future therapies and advanced disease studies, many challenges in ESCs studies remain. Insufficient understanding of cell dynamic behaviours (e.g. self-organisation and fate decision) still hinders the development of targeted stem cells therapies (Tabar and Studer, 2014). ESCs have been shown to self-organise to form spatial patterns in vitro that resemble in vivo developmental processes (Brink *et al.*, 2014). These processes of spatial patterns forming are essential for establishing mammalian body plan. Capitalizing on the intrinsic ability of cells to self-assemble and self-organize into complex and functional tissues and organs, embryoids, organoids and gastruloids have recently been generated in vitro (Simunovic and Brivanlou, 2017). Understanding these processes would provide new insights into cell behaviours as well as controlling the ESCs differentiation, which is the precondition for effective stem cell therapies.

There have been many studies focusing on understanding the mechanisms of cell behaviours at the molecular level (the level under cellular level) (Young, 2011). Since the discovery of the structure of DNA, the genome has often been thought of as the overriding architect: a given combination of genes that determines the phenotype through a linear chain of causal events (Rossant and Joyner, 1989). The problem is that in addition to genetic approaches which have revealed important aspects of spatial pattern formation (Rossant and Joyner, 1989), other sophisticated mechanisms are involved as well (Beccari *et al.*, 2018). Embryogenesis and dynamic cell forms and functions emerge from multiple molecular interactions and interconnected regulatory feedback loops as discussed in Section 2.1.1. Moreover, many additional characteristics, such as physical constraints, collective behaviours, interactions between cells and extracellular matrix (ECM) are not under the direct control of the genome. Therefore, we cannot hope to explain cell morphogenesis, for example, by invoking simple linear chains of causal events that link genes to phenotypes (Karsenti, 2008). However, due to the complexity of cell-cell and cell-ECM

interactions, as well as the signalling network, we are still a long way from having a full understanding of the underlying mechanisms (Prasad *et al.*, 2016).

A key practical question is whether deep biological modelling of the cells is essential to predict their pattern formation, or whether there is sufficient predictive power in simply modelling their behaviours and interactions at a higher level (Kamm *et al.*, 2018). Similarly to diverse fields in biology and physiology, we can develop mathematical models focusing at different levels of investigating biological mechanisms, e.g. at the cellular level to organ/systems level; this is a major field that has attracted considerable attention in the last 30 years. An authoritative resource is Keener and Sneyd's two books on mathematical physiology which focus on cellular physiology modelling (Keener and Sneyd, 2009a) and systems physiology modelling (Keener and Sneyd, 2009b), respectively. Arguably, higher-level studies can provide a mechanistic understanding of the key underlying mechanisms that impact ESCs pattern formation, which is beneficial for obtaining the desired pattern required for tissue regeneration. Hence, in this study, we focus on the social interactions and behaviours of ESCs at a high-level to predict aggregate crowd behaviours within a level of uncertainty.

1.3 Scope and structure of the thesis

Given that the underlying mechanisms of pattern formation in ESCs at the molecular level is very complicated, we focus on investigating pattern formation at a higher level. We hypothesize that the pattern formation in ESCs can be modelled from their high-level features and social interactions at a cell population level.

This thesis aims to address the following key questions:

1) Can we generate a minimal model to reproduce the pattern formation in ESCs on the population level by using minimal rules?

2) What are the essential underlying cell motility rules to achieve the observed pattern formation?

3) How do we evaluate our model by assessing the differences between model results and empirical data?

To answer these key questions, we applied *agent-based model* strategies to investigate specific observed pattern formation (Section 6.1) in ESCs (more details about agent-based modelling are provided in Section 2.3.3), and tested a wide range of motility rules (Section 5.3) to obtain a parsimonious set of rules that can reproduce the pattern formation observed in experimental data.

This study aims to reproduce pattern formation with a minimal set of behavioural rules. The rules in our models might not be consistent with the mechanisms in reality as our models represent the idealisations of the physical and chemical system. However, our minimal models provide valuable output because they can be used as tools for predicting pattern formation in related contexts. These models, which can be re-parameterised easily, hold the potential to reduce the need for extensive and costly experiments by providing opportunities to test the effect on different individual social and dynamic cell behaviours and their combinations.

The framework of this study is presented in Figure 1-1. The biological lab based in the Centre for Regenerative Medicine led by Guillaume Blin supported experimental data collected from the wet lab as well as the data collected from the images. The experimental data we used in this study are static cell images of cell colonies collected 48 hours after cell seeding. Cells fully colonised the geometrical confined area and formed dome-shaped colonies after 48 hours. After visualising and analysing our experimental data, we proposed high level but biologically plausible rules of cell behaviours based on previous studies and experimental observations. We evaluated our models by comparing model outputs to the experimental data. Based on our models, we proposed new testable rules of cell behaviours.

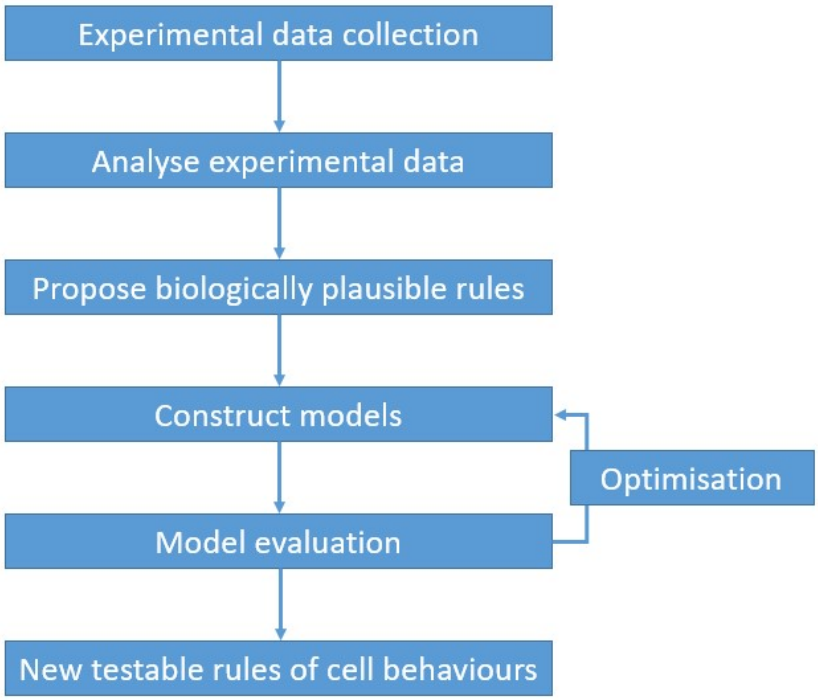


Figure 1-1: Framework of this study.

In Chapter 2, we provide a comprehensive literature review, including biological studies and different types of cell models. Furthermore, the review illustrates the gap that our study fills in and clarifies the reasons for choosing the modelling strategy. In Chapter 3, we describe our experimental data collected from the wet lab which is used as the ground truth for evaluating the developed models. The visualisation of experimental data is provided in Chapter 3 as well. In Chapter 4, we provide background information on the mathematical algorithms we used in this study, including existing algorithms and introduce a novel metric to assess ESC pattern formation model estimates. In Chapter 5, we describe our models in natural language as well as providing a formal definition of the model structures. Besides, the descriptions of the biologically plausible rules we proposed are also provided in Chapter 5. In Chapter 6, we present detailed analysis results of the experimental data,

including the description of the patterns we observed in experimental data. Subsequently, we provide our model outputs and evaluation results in Chapter 7. Finally, in Chapter 8, we summarise the key findings, implications and limitations of this work, and provide pointers towards future work.

1.4 Novel contributions

This study makes some key contributions towards understanding ESC pattern formation properties, towards providing new insights which may translate into novel therapies. Specifically:

- We proposed a novel framework to study the social behaviours of ESCs through agent-based modelling. We introduced biologically plausible rules (we focus on motility rules in this study) which may lead to the resulting pattern formation. The rules are described in Section 5.3.

- We constructed a set of 16 novel algorithmic models to test all combinations of these plausible rules and evaluated the models. The models were optimised both in terms of identifying the most parsimonious set of rules and also by internally optimising specific model hyper-parameters (see Section 5.6 for details regarding parameter optimisation and Section 7.3 for results with model optimisation). By dissecting the most successful modelling approaches, we gain some mechanistic insights into cell behaviours and propose additional biologically plausible rules of engineering cells to achieve desired patterns.

- We developed a new evaluation approach to assess our models by calculating the distance between the aggregate results from models and our experimental data. We found that the basic methods we applied initially for evaluation did not follow the visual impression we had from observing the resulting patterns (more details are provided in

Section 7.1), which led to exploring new methods and developing a new evaluation approach that is tailored specifically to this application. The novel evaluation metric is described in Section 4.4.3. This new evaluation metric allows us to select the best model with resulting patterns that is closest to our experimental data, which also is consistent with our visual impression.

- The new framework provides new insights into cell behaviours from pattern formation in ESCs on a population level, including providing a tool for analysing the relationship between pattern formation and cell behaviours through simulations. The information collected through this framework could assist other biological studies in the future. In addition, modelling is beneficial for biological studies since it accelerates the studies and reduces the financial cost in wet labs.

Overall, we generated a series of models to study cell behaviours at a cellular level. These models, which do not include the complicated underlying molecular mechanisms, reproduced the pattern formation in ESCs with biologically plausible rules of cell behaviours. This is a step towards accelerating the development of clinical stem cell therapies by providing information for engineering cells to probabilistically control pattern formation. These models have the potential to facilitate future work on understanding underlying cell behaviours. Ultimately, these models can accelerate the development of clinical stem cell therapies by providing key new insights contributing to designing new stem cell therapies.

These novel contributions resulted in the following peer-reviewed publications:

- **M. Wang**, A. Tsanas, G. Blin, and D. Robertson, “Predicting pattern formation in embryonic stem cells using a minimalist, agent-based probabilistic model,” *Scientific Reports*, vol. 10, e16209, 2020.

- **M. Wang**, A. Tsanas, G. Blin, and D. Robertson, “Assessing preferred proximity between different types of embryonic stem cells,” 13th Int.

Conf. Bio-Inspired Syst. Signal Process. Proceedings. (BIOSTEC), pp. 377–381, 2020.

- **M. Wang**, A. Tsanas, G. Blin, and D. Robertson, “Investigating motility and pattern formation in pluripotent stem cells through agent-based modeling,” 19th IEEE International Conference on Bioinformatics and Bioengineering (BIBE), pp. 909–913, 2019

- **M. Wang**, D. Robertson, G. Blin, S. Lowell, and T. Tsanas, “Agent-based modelling of pattern formation in pluripotent stem cells : initial experiments and results,” in 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2018.

- H. Wu, **M. Wang,** Q. Zeng, W. Chen, T. Nind, E. Jefferson, M. Bennie, C. Black, J. Z. Pan, C. Sudlow, and D. Robertson, “Knowledge driven phenotyping,” Stud. Health Technol. Inform., vol. 270, pp. 1327–1328, 2020.

Chapter 2

2 Literature review

This chapter provides a literature review that covers two principal topics: biological studies on ESCs and related mathematical modelling approaches. We reviewed biological studies on cell behaviours and self-organisation as well as summarised the existing knowledge driven from these studies. We also reviewed multiple modelling approaches that are commonly used in modelling pattern formation in cells and explained the advantages and disadvantages of different approaches.

2.1 Biological studies on stem cells

Since understanding stem cell behaviours is a precondition of fully controlling stem cells differentiation to generate desired tissues or organs, many extensive studies are working on understanding stem cell behaviours. The mechanisms of cells self-organisation in an embryo are remarkably complex, from the mechanisms on the molecular level focusing on the units smaller than cells (e.g. chemical elements) to the population level focusing on collective dynamic and social behaviours of cells (Phadnis *et al.*, 2015). Some examples of molecular levelled studies are the studies focusing on the genetic and epigenetic level (Guo *et al.*, 2013), analysing signal sensing (Chacón-Martínez, Koester and Wickström, 2018), and working on the interactions with extracellular matrices (Guilak *et al.*, 2009; Gattazzo, Urciuolo and Bonaldo, 2014; Vining and Mooney, 2017). Collective behaviours on a population level are a result of collective molecular-level mechanisms. One step further, the ability to self-organising gives rise to the pattern formation at a population level, which is a result of collective interactions of individual cells between

themselves and the extracellular environment (Kamm *et al.*, 2018). Based on studies from past years, we have a large amount of knowledge-driven from biological studies on stem cells, which enables knowledge-based modelling and also provides the clues of our biological plausible hypothesis (Zhou *et al.*, 2007; Phadnis *et al.*, 2015; Vining and Mooney, 2017).

2.1.1 Biological studies on stem cell behaviours

Among the biological studies on stem cells in the past years, many biomarkers were found that regulates cell behaviours. For example, according to the knowledge-driven from biological studies on stem cells, it is known that the pluripotent state in ESCs is mainly regulated by the core transcription factor trio of Oct4, Sox2 and Nanog (Young, 2011). They are transcription factors required to maintain the pluripotency and self-renewal of ESCs (Wang *et al.*, 2006). Many other studies also uncovered additional novel ESC regulators besides the trio. The importance of some mESC factors such as Esrrb, Tbx3 and Tcl1 (Ivanova *et al.*, 2006)(Zhou *et al.*, 2007), as well as the chromatin regulators Tip60-p400 (Fazzio, Huff and Panning, 2008) and SetDB1 (Bilodeau *et al.*, 2009), has been discovered. More recently, Semrau and colleagues measured the gene expression dynamics in mESC differentiation from pluripotency to lineage commitment (Semrau *et al.*, 2017). They provided a comprehensive analysis of the exit from pluripotency and lineage commitment at the single-cell level, which is a potential stepping stone to improved lineage manipulation through the timing of differentiation cues.

When ESCs exit pluripotency in vitro, they undergo the same lineage transitions as they would in the embryo. Hence, ESCs can be used to mimic the early stages of embryonic development. Gastrulation is a phase early in embryonic development which is apparent by embryonic day (E) 6.5. During this process, gastrula is formed from reorganizing pluripotent epiblast into a multi-layered structure with the formation of the primitive streak (PS). The PS is a transient structure whose formation marks the start of gastrulation. The

formation of the PS is one of the earliest signs of anteroposterior polarity. The positioning of the streak depends on a complex interplay between several signalling molecules including Nodal, Wnt and BMP that are initially expressed and secreted by extraembryonic structures (Brennan *et al.*, 2001; Ben-Haim *et al.*, 2006; Marikawa *et al.*, 2009). More recently, symmetry breaking has been described in vitro with ESC that cannot generate extraembryonic lineages suggesting that symmetry breaking can happen spontaneously in the absence of extraembryonic structures (Brink *et al.*, 2014; Blin *et al.*, 2018). However, the mechanisms of this spontaneous symmetry breaking are currently unknown.

The PS is characterised by the expression of early mesendodermal markers such as brachyury (T). As reported by (Smith, 1997), the Brachyury (Greek for 'short tail'), or T (tail), the mutation was first described in 1927 (Dobrovolska'ia-Zavadskaa, 1927). Brachyury (T) is the founder member of a family of transcription factors that share the so-called T-box – a 200 amino acid DNA-binding domain. In vivo, Brachyury (T) marks the onset of gastrulation (Beddington, Rashbass and Wilson, 1992), during which the three primary germ layers and the basic body plan are established. This marker is also found to be expressed within a subpopulation of mESCs in culture (Tsakiridis *et al.*, 2014). mESC T+ cells harbour a gene expression profile reminiscent of the primitive streak (Suzuki *et al.*, 2006), a structure that gives rise to both mesodermal and endodermal derivatives during development (Rossant and Joyner, 1989).

In this project, we focus on studying the pattern formation in mESCs colonies of cells marked by Brachyury (T). Cells marked by Brachyury (T) are called T+ cells; while T- cells are the cells not marked. Hence, T+ cells are early differentiated cells; while T- cells are naïve cells, which means they have not started differentiation yet. In cell colonies with geometrical confinements, asymmetric patterning of T+ cells were observed as T+ cells show polarised patterning. Detailed descriptions of cell colonies with geometrical

confinements will be provided in Chapter 3. The illustration of T+ cell patterning will be provided in Section 6.1.

As the mechanisms of spontaneous symmetry breaking are still unknown, many possible mechanisms guide cell behaviours. In addition to intrinsic factors, many extrinsic factors affect and control the fate of stem cells. For example, coordinated interactions with soluble factors, other cells, and extracellular matrices define a local biochemical and mechanical niche with complex and dynamic regulation that stem cells sense, would play a role in controlling stem cells (Discher, Mooney and Zandstra, 2009). Through integrin-mediated focal adhesions, cells can anchor onto the underlying substrate, sense the surrounding microenvironment, and react to its properties. Substrate-cell and cell-cell interactions activate specific mechanotransduction pathways that regulate stem cell fate (Nava, Raimondi and Pietrabissa, 2012). Mechanical factors, including substrate stiffness, surface nanotopography, microgeometry, and extracellular forces can all have a significant influence on regulating stem cell activities (Nava, Raimondi and Pietrabissa, 2012). ECM is a dynamic and complex environment characterized by biophysical, mechanical and biochemical properties specific for each tissue and able to regulate cell behaviour. Since ECM can directly or indirectly modulate the maintenance, proliferation, self-renewal and differentiation of stem cells, there are plenty of studies working on the mechanisms of how ECM affects stem cell behaviours (Gattazzo, Urciuolo and Bonaldo, 2014).

In addition to the biological studies focusing on a molecular level, existing studies are working on quantifying cell behaviours on a cell level, which assist us to propose biologically plausible rules for modelling. It is known that stem cells can interact with their microenvironment (Fuchs, Tumbar and Guasch, 2004), and have social interactions with neighbouring cells (Gong *et al.*, 2008). Understanding the protocols of these interactions would assist in the controlling of stem cell differentiation.

Phadnis et al. delivered some quantitative analysis on dynamic and social behaviours of human pluripotent stem cells (Phadnis *et al.*, 2015). They

revealed that the density of the colony affects stem cell behaviours, including cells survival rate, cell velocity, and cell size. They reported that cell behave differently depends on their types as they measured that differentiated cells have a higher speed than undifferentiated cells, which is consistent with the results reported by (Turner, Rue, *et al.*, 2014). Blin and colleagues demonstrated that T+ and T- cells tend to have different numbers of neighbouring cells. On average, T+ cells have less neighbouring cells within the same radius compared to T- cells (Blin *et al.*, 2018). However, even though we know it is likely that the state of the cells is influenced by interactions with neighbouring cells, to the best of our knowledge no studies are quantifying and demonstrating cells community effect (how their behaviours effected by neighbouring cells) and little is known about how interactions between ESCs influence their dynamic and social behaviours.

2.1.2 Biological studies on stem cell self-organisation and patterning

Self-organisation was first defined by the philosopher Kant as a characteristic of living systems implying the existence of a loop between organisation and function (Van De Vijver, Van Speybroeck and Vandevyvere, 2003). Recent work has provided a simpler definition of self-organisation as a *dynamic organisation emerging from the collective behaviours of 'agents'* (Karsenti, 2008). These individual agents have properties that cannot account for the properties of the final dynamic pattern. Thus, from observing cells on cell-level, patterns in the living cells can emergent from collective behaviours from individual cells. From previous biological studies, there are many approaches are being used to answer fundamental questions about cells.

One of the approaches is studying cell self-organisation by providing an engineered environment. Micropatterns, the spatially confined areas that are adhesive to cells (Falconnet *et al.*, 2006), provide the environments of studying cells in vitro with controlling the spreading of attached cells through engineered

surfaces. Micropattern is very important for designing cell culture for tissue engineering and studying the relationship between cell patterning and spatial confinement. Micropatterns play an important role in stem cell studies because controlling microenvironments through engineering surfaces is key to guiding the differentiation of stem cells.

Previously, existing studies focusing on the relationship between geometric confinement and pattern formation in ESCs showed that geometric confinement is sufficient to trigger self-organized patterning in hESCs (Warmflash *et al.*, 2014). In their study, they confined cells on circular micropatterns and marked the differentiated cells with bone morphogenetic protein 4 (BMP4), which is a factor playing a role in self-renewing and differentiation. In vivo BMP4 is secreted by the extraembryonic ectoderm and triggers gastrulation. Afterwards, they produced an ordered array of germ layers along the radial axis of the colony. They demonstrated that hESCs will self-organise to generate embryonic patterns with being given minimal geometric and signalling cues. (Deglincerti *et al.*, 2016) described a protocol of micropattern approach from differentiation patterns. In their study, hESCs are confined on disc-shaped micropatterns and form patterns in concentric radial domains, which express specific markers associated with the embryonic germ layers, reminiscent of gastrulating embryos. More recently, (Britton *et al.*, 2019) developed an in vitro model of human ectodermal patterning, in which hESCs self-organise to form robust and quantitatively reproducible patterns by using micropatterns to provide geometric confinements.

Cell sorting is the process that cells are separated according to their type due to their different properties (Rosental *et al.*, 2017), which can occur between cells that express different cadherins with promiscuous binding specificities (Niessen and Gumbiner, 2002), and sometimes occurs with the equally strong heterophilic and homophilic association as assayed in vitro (Prakasam, Maruthamuthu and Leckband, 2006). These studies suggested that adhesion specificity and the strength of cell association are not fully captured by the extracellular binding of cadherins and that kinetic parameters might also

control cell adhesion and cell sorting. With these suggestions in mind, the differential adhesion hypothesis has been applied to study cell cultures for understanding self-organisation (more details in Section 2.3.1).

In addition to cell shape change, cell proliferation, and differential cohesion or tension can lead to cell patterning in tissues. Mori et al. revealed that the differences in cell motility can also lead to cell sorting within tissues (Mori *et al.*, 2009). They investigated how the tree-like branching pattern is regulated in the mammary gland by using mosaic engineered mammary epithelial tubules. They found that epithelial cells sorted on rectangle-shaped mammary epithelial tubules as cells marked by MMP14 (cells responsible for initiating branching express Metalloproteinases) sort to the leading edges of these ducts. With the findings showing that cell speed and persistence time were enhanced by MMP14 expression, they constructed agent-based models (see Section 2.3.3 for further details) to investigate how cellular differential motility leads to cell sorting. According to their models, only cell persistence time (the time length of the cell sticking with one direction) was required for sorting. These results indicated that differential directional persistence could be one of the key factors that give rise to patterns within model development tissues.

Cells have a remarkable capacity to self-organise and self-assemble into complex and functional structures. As discussed before, cell behaviours are governed by many factors. Hence, many factors play roles in pattern formation during embryogenesis. For example, the physical and morphological properties of cells, the signal they receive, and the mechanical properties of tissues (Simunovic and Brivanlou, 2017). Moreover, many parameters, such as physical constraints and collective behaviours, are not under the direct control of the genome (Karsenti, 2008). Hence, we cannot hope to explain cell self-organisation simply by the linkage between genes and phenotypes. With all existing studies on patterning and cell behaviours in ESCs, there is still a lack of studying on cells social behaviours and community effects on a quantitative level. It is still unknown how cells' different behaviours affect the

pattern formation in cell colonies, and what the key behaviours to achieve the pattern formation are.

Hence, we focus on studying the pattern formation from the cell-level, we do not reason cell behaviours from the molecular level. More specifically, we focus on analysing the relationship between pattern formations and collective cell behaviours based on different behaviour protocols. From previous biological studies, many approaches are being used to answer fundamental questions about cells self-organising in embryology and to study how we reproduce pattern formation.

2.2 Stem cell engineering

Over the past several decades, bioengineers, biophysicists, and biologists have made steady progress toward the creation of systems that are composed of living cells and tissues organized in a way that produces novel functionalities by design (Kamm *et al.*, 2018). In this section, we summarise the relevant stem cell engineering technologies. These novel systems hold a broad range of potential uses including disease treating and discovering since they allow biologist to edit cell behaviours to obtain desired cell patterning. One of the popular approaches to engineer stem cells is working on the cellular microenvironment, which is composed of both physical and chemical signals, including extracellular matrix proteins, neighbouring cells, soluble and immobilized growth factors, and small molecules (Flaim, Chien and Bhatia, 2005; Jang and Schaffer, 2006). Since stem cell differentiation and self-renewing is determined by this complex collection of factors present in the local environment of the cells, many studies are focusing on the biomaterials that can be employed to regulate cell behaviours for tissue engineering applications, such as adhesion, proliferation, and differentiation (Lutolf and Hubbell, 2005).

The majority of these studies working on synthetic biomaterials to mimic the regulatory characteristics of natural extracellular matrix (ECM) and ECM-bound growth factors. The developments of this approach include nanofibrillar network formed by self-assembly of small building blocks, artificial ECM networks from protein polymers or peptide-conjugated synthetic polymers that present bioactive ligands and respond to cell-secreted signals to enable proteolytic remodelling. These materials have already found application in differentiating stem cells into neurons, repairing bone and inducing angiogenesis (Lutolf and Hubbell, 2005).

Hydrogel platforms have been developed to regulate stem cell fate by controlling micro-environmental parameters including matrix mechanics, degradability, cell-adhesive ligand presentation, local microstructure, and cell-cell interactions. A recent study (Madl and Heilshorn, 2018) summarised the approach of modulating hydrogel microenvironments properties to recapitulate the stem cell niche. They reviewed the effects of micro-environmental parameters on maintaining stemness and controlling differentiation for a variety of stem cell types. In addition to mimicking ECMs, mechanical forces have been reported to induce proliferation and/or differentiation in ESCs. (Saha *et al.*, 2006) reported that the differentiation of hESCs could be inhibited by a mechanical strain.

These approaches focus on providing the synthetic materials that contain the necessary signals to recapitulate developmental processes in specific differentiation. Even though we can control the distribution of the cells by controlling the concentration of specific chemicals (e.g. generating gradient patterns), we do not have control over cells local effects. As discussed before, cell-cell signalling is an important component of the stem cell microenvironment. It affects both differentiation and self-renewal of stem cells. With traditional cell-culture techniques, we do not have precise control over cell-cell interactions. (Rosenthal, Macdonald and Voldman, 2007) created a microfabricated polymer chip to trap down to a single stem cell or pattern small

groups of cells with or without cell-cell contact. This new tool provides the opportunity of engineering a single stem cell.

Many studies have provided insights into the transcriptional control of embryonic stem cell state, including the regulatory circuitry underlying pluripotency (Young, 2011). Besides, by using synthetic biology techniques to engineer simple genetic or cellular systems, it is possible to test principles of patterning, differentiation and morphogenesis to see whether they perform as expected (Davies, 2017). For example, engineering cell speed and level of adherence of cells can be achieved by using synthetic biology (Cachat *et al.*, 2014).

Armed with existing knowledge, biologists can control many aspects of the environment of the cells and can thus steer self-organisation to produce bio-medically relevant products. However, quantitative studies of cell behaviours are needed to identify the key parameters that influence pattern formation.

2.3 Mathematical modelling of pattern formation in stem cells

In morphogenesis, a key aim is to understand the mechanisms underlying spatio-temporal pattern formation (Wolpert, 1969). Although genes play a crucial role, a study of genetics alone cannot provide a mechanistic understanding of how physical and chemical processes within a developing system conspire to produce the complex multi-scaled factors and signals to which respond to or interact with. These problems extend beyond biology and require multi-disciplinary analysis due to the complexity of the systems (Maini, 2004). Such systems are amenable to mathematical modelling and the role of the modeller is to suggest explanations, based on biologically plausible mechanisms, of observed behaviours and make experimentally testable predictions.

2.3.1 Theoretical models

The goal of mathematical modelling is to provide mechanistic insights into emerging problems in biology and allied fields by offering quantitative techniques, analysis and solutions (Keener and Sneyd, 2009a, 2009b; Rabajante *et al.*, 2015). Most of the mathematical models in ESCs are knowledge-based (bottom-up) models, which were built for predicting the behaviours and cell fate from the molecular level (Pir and Novère, 2015). For example, the Waddington model quantifies the dynamics of cell-fate specification based on gene regulatory networks (Ladewig, Koch and Brüstle, 2013; Rabajante and Babierra, 2015). However, these knowledge-based models can be tedious to build as the relevant complex molecular interactions have to be collected from the literature. In addition to these models, many models are working on a higher level to get insights into pattern formation in ESCs.

Turing was the first to realize that the interaction of two substances with different diffusion rates can cause pattern formation (Turing, 1952). In his hypothesis, a spatial pre-pattern in biochemical (where he coined the term *morphogens*) causes the patterns we observe during embryonic development. Therefore, cells would respond to this pre-pattern by differentiating in a threshold-dependent way. He proposed that these pre-patterns are generated by reaction-diffusion (Maini *et al.*, 2012). The Turing model (reaction-diffusion model) is one of the best-known theoretical models used to explain self-organised pattern formation in embryogenesis. Based on Turing's model, Gierer and Meinhardt proposed a theory of biological pattern formation in which concentration maxima of pattern forming substances are generated through local self-enhancement in conjunction with long-range inhibition (Gierer and Meinhardt, 1972). Hence, there are two types of morphogens, one acting as an activator, and one with an inhibitory effect. In this way, compared to Turing's model, the model is more biological interpretable as there are molecular candidates for activators and inhibitors (Meinhardt and Gierer, 2000).

Turing's theoretical model explains self-regulated pattern formation in developmental biology, however, it is long debated that how relevant Turing's model is to the real-world as the mechanisms of cell dispersal is not diffusion. Before the observed pattern formation in ESCs, all cells are equivalent to begin with. Afterwards, pattern emergent in ESCs based on cell rearrangement, for example, cells immigration, differentiation or apoptosis. And it is meaningful to understand the protocols of cells self-organisation.

In the 1960s, Steinberg proposed the differential adhesion hypothesis (DAH). He hypothesized that the interaction between two cells involved an adhesion surface energy that varied according to the cell types. DAH has been applied to polarization problems in ESCs. As each type of embryonic has a unique "tissue surface tension" (TST) that governs how tissues sort and differences in cell-cell adhesion result in differences in TST. Hence, patterns arise from these different tensions (Steinberg, 1970). They showed that, without exception, a cell aggregate of lower surface tension tends to envelop one of higher surface tension to which it adheres (Foty and Steinberg, 2005). Based on the theory, in the next step, we need to develop tools to precisely quantify interfacial tensions inside tissues. We need to develop a model that predicts how mobile adhesive molecules affect cell shapes and/or forces and vice versa. We also must extend these models to account for a broader range of mechanical effects at boundaries, such as differences in cell protrusivity and oriented cell divisions. Another challenge is to interpolate between short-time scale mechanical interactions that are important in tissues with highly mobile cells and long-timescale interactions that are important for dense tissues with mature contacts (Amack and Manning, 2012).

2.3.2 Cellular automata and cellular potts modelling

One popular and intuitive approach to study self-organisation in cells is modelling the phenomenon, the aggregated pattern, in terms of the interactions or behaviours of the individuals (cells). There are multiple similar

methodologies in this domain. For example, cellular automata, cellular potts modelling, and agent-based modelling. In these approaches, we take autonomous computational individual or object as *agents* and model their interactions. However, by definitions, there are some differences between these modelling approaches. After explaining cellular and cellular potts modelling in this section, we will describe agent-based modelling comprehensively in the next section.

Cellular automata (CA) (Wolfram, 1983), sometimes known as a subset of agent-based modelling (Van Liedekerke *et al.*, 2015), have been defined as "*discrete spatio-temporal dynamic systems based on local rules*" (Miller, 2009). CA are the simplest modelling framework that allows simulating extraordinarily complex behaviour and demonstrating the emergence of patterns (M Batty, 2000). CA contain a grid of cells with initial conditions and a finite number of states and a set of rules which govern changing the states of the cells or exchanging information among the neighbouring cells. Many studies applied CA to cell biology. Different approaches with CA models were used to model cancer growth. (Monteagudo and Santos, 2015) applied CA to model the growth of cancer stem cells. (Poleszczuk and Enderling, 2014) proposed a CA model of tumour growth with high-performance. Besides cancer growth, (Garijo *et al.*, 2012) proposed a stochastic CA model for muscle satellite cells.

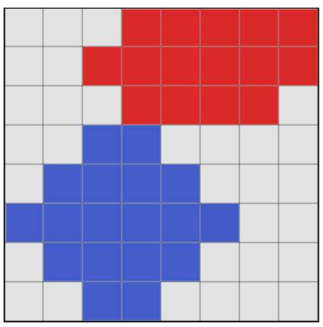


Figure 2-1: An illustrative figure of CPM with a grid-based environment and two cells (red and blue cells) are defined by covering multiple grids.

Similar to CA, the Cellular Potts Model (CPM) is a spatial grid-based model that a cell is defined over a region composed of multiple lattice sites (Figure 2-1 shows an example of CPM). The CPM formalism is very suitable for modelling biological cells as it takes cells as deformable objects by taking their shapes from a combination of internal and external forces which act upon it (which is different from CA) (Marée, Grieneisen and Hogeweg, 2007). By mapping the parameters of the basic CPM formalism to physical and biological properties of cells, CPM is a powerful tool for investigating a large range of biological questions, including biophysical properties of single cells, tissue-level properties, and understanding the full embryogenesis and morphogenesis of an organism's life-cycle.

Because of the special features of CPM, it has been applied to problems in stem cells. (Libby *et al.*, 2019) constructed computational replication of the self-organised hPSCs pattern by using an extended CPM enabled machine learning-driven optimization of parameters that yield the pattern emergence including cell immigration velocity. They also demonstrated that morphogenic dynamics can be accurately predicted through a model-driven exploration of hPSC behaviours via machine learning. Besides pluripotent stem cells, CPM has also been applied to other problems in biomedicine. For example, epidermal stem cells, tumour growth and invasion, and blood vessel growth (Savill and Merks, 2007). However, since CPM captures the irregular shapes of individual cells, CPM has a high computational cost and limitations in representing the mechanical integrity of large-scale structures. Hence, CPM is not the primary formalism of choice for modelling multi-cell systems in ESCs and it is unlikely to become the primary formalism due to its high computational cost and limitations in representing the mechanical integrity of large-scale structures (Kamm *et al.*, 2018).

2.3.3 Agent-based modelling in stem cells

The core idea behind agent-based modelling (ABM) is that many phenomena in the world can be effectively modelled with *agents*, an *environment*, and a description of *agent-agent and agent-environment interactions*. An agent is an autonomous individual or object with particular properties, actions, and possibly goals. Figure 2-2 gives an example of two agents in a grid-based environment. The size of agents can be adjusted as an agent can occupy one grid or multiple grids. The environment is the landscape on which agents interact and can be geometric, network-based, or drawn from real data (Wilensky and Rand, 2015). The models consist of a series of *states* (often time-stamped) and agents can update their internal *state* (timestamp) and also update their additional actions. Agents can interact with other agents or with the environment based on defined rules. The interactions, including the exchange of information, can be complex and change in time (through state change). Hence, agent-based models are often state-based and rule-based.

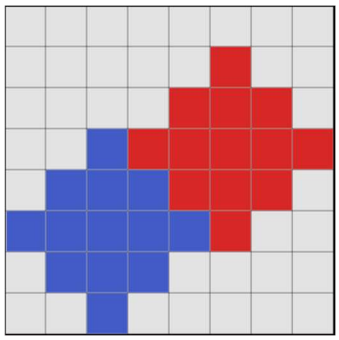


Figure 2-2: An illustrative figure of agent-based modelling including a grid-based environment and two agents occupying grids.

The history of the ABM can be traced back to the proposing of the Von Neumann machine, a theoretical machine capable of reproduction, in the 1960s (Von Neumann and Burks, 1966). In the 1970s and 1980s, some earlies ABM were constructed, including Thomas Schelling's segregation model

published in 1971 (Schelling, 1971). In the 1990s, ABM experienced an expansion, which came along with the availability of collecting large scale data and assessing powerful computation (Wilensky and Rand, 2015). With the aid of the computation ability we have today, we can simulate complex patterns and better understand how the patterns arise in many domains, both nature and society. Since agent-based modelling provides an intuitive view to study complex systems and emergence by aggregating individual behaviours and local effects, ABM is a suitable computational methodology that facilitates modelling of complex systems and investigates the rules at the micro-level that lead to the ordered pattern at the macro-level. Nowadays ABM is widely used in many scientific domains including biology, ecology and social science (Niazi and Hussain, 2017). In recent years, a large number of studies carried out by applying ABM in different domains. For example, in supply chains (Utomo, Onggo and Eldridge, 2018), social experiments (Tong *et al.*, 2018) and synthetic biology (Gorochowski, 2016).

ABM has been particularly successful in settings where the primary scientific question is about the control of differentiation through discrete cell states or fates (Setty, 2012). Over the past few decades, many different agent-based models were developed to mimic the multicellular organization. Many models were constructed to investigate self-organisation in ESCs. (Briers *et al.*, 2016) applied ABM for generating different patterns in ESCs forming embryoid bodies (EBs). They modelled cells proliferation and differentiation to produce the patterns observed in 3-dimensional spheroids, including differentiated cells localise at the border of the spheroids. They classified the patterns observed and simulated different types of patterns based on the formal specification of patterns. (White *et al.*, 2013) also have applied rules and agent-based modelling to ESCs forming EBs. Their results indicate that the rules dominate the emergence of patterns independent of EB structure, size, or cell division.

Besides ESCs, ABM has been applied to a wide range of cell modelling. (Poleszczuk, Macklin and Enderling, 2016) described the design and implementation of a lattice-based agent-based model of cancer stem cell-

driven tumour growth. (Wang *et al.*, 2015) provide a review that introduced some agent-based models that simulated cancer growth. (Walker *et al.*, 2004) applied ABM to simulate growth characteristics of epithelial cells. Rule-based ABM has also been used to model cellular signalling by describing biological interactions in terms of rules (Danos *et al.*, 2008).

Even though the majority of agent-based models in ESCs are grid-based models as the environment are composed of grids. Compared to CA, ABM has a higher degree of freedom as the agents are allowed to move around freely. Different from CA, the purpose of ABM is often the exploration of variants in system behaviour due to agent characteristics (such as different behaviours of different types of agents) or rules, rather than resulting aggregate structures. Many agent-based models are multi-agent models, which include more than one type of agents. (D'Inverno and Saunders, 2005) demonstrated some case studies of modelling stem cells with an agent-based approach as opposed to a CA approach. They illustrated that the multi-agent approach to modelling is appropriate because of its higher degree of freedom with modelling the cells as agents that can move on the top of the environment and react to their environment. With the multi-agent approach, we can build simple models of agents and environment to test biologically plausible simple rules that give rise to complex global behaviour.

Compared to equational models that are constructed from mathematical terms, ABM can be more intuitive and easier to interpret. Instead of mathematical symbols, accessible agent-based models are constructed out of objects that humans can readily relate to real-world entities and for which we can define intuitively simple rules for their behaviours or interactions. When describing these rules, the language and concepts we use in ABM are ideally much closer to natural language and our natural thinking. Hence, in the right circumstances, agent-based representations can be easier to understand than more abstract mathematical representations of the same phenomenon.

As described in this Chapter, previous studies reported that spontaneous patterning in ESCs can be triggered by geometrical confinement. While many

studies are focusing on understanding the mechanisms of cell behaviours from the molecular level, the mechanisms are still not fully understood due to the complexity of this problem. Hence, the mechanisms of this spontaneous patterning remain unknown. The agent-based modelling approach was applied in many biological settings to study cell behaviours and has been demonstrated as one of the suitable approaches to study pattern formation at a popular level. In this study, we aim at applying agent-based modelling to reproduce the pattern formation with a minimal set of rules. Our models fill in the gap of proposing new testable potential cell motility rules to achieve specific pattern formation in geometrically confined ESCs colonies.

Chapter 3

3 Experimental data collection, preparation and visualisation

Following the literature review of related biological studies and modelling approaches, we describe the collection, preparation and visualisation of the experimental data in this chapter. Based on the information of biomarker and data types provided in Chapter 2, in this chapter, we deliver the detailed descriptions of cell seeding and growing. We illustrate the framework of collecting data from cell images and visualise the experimental data. The process of experimental data collection is provided in Figure 3-1. Biological experiments were achieved by the team led by Guillaume Blin based in the Centre for Regenerative Medicine. All static images we used in this study were supported by Blin's lab as well as the image processing software and the procedures of extracting information from images. The methodologies of processing experimental data will be provided in Chapter 4 and the analysis results from experimental data will be provided in Chapter 6.

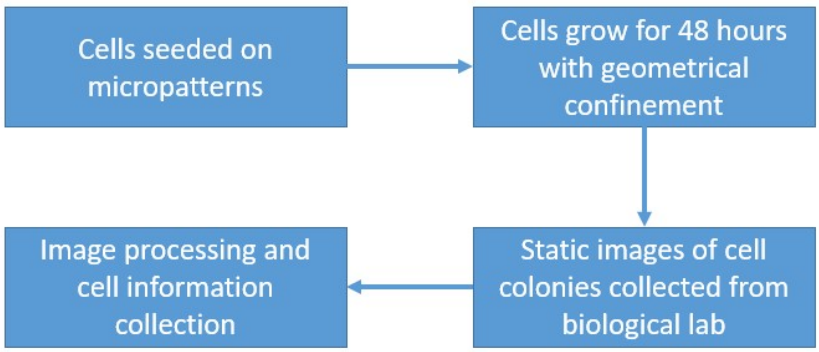


Figure 3-1: The process of experimental data collection

3.1 Cell seeding and growing

We have the disc and ellipse-shaped micropatterns and the dimensions are shown in Figure 3-2. The sizes of disc and ellipse micropatterns are both approximately 30,000 μm^2 . Micropatterned chips were fabricated using untreated IbiTreat plastic slides (Ibidi, IB-10813) as the base substrate. More details of ESC micropatterning was provided in (Blin *et al.*, 2018).

Initially, on average 6 cells were randomly seeded on disc and ellipse-shaped micropatterns. 7% of the initial cells are T+ cells while the rest are T- cells. Cells fully occupied the confined area and formed a dome shape 48 hours after seeding. Figure 3-3 gives an example of the 3D image of the cell colony after 48 hours. Nuclei are marked by Lamin B1. Red cells are the cells marked by Brachyury (T). Even though theoretically cells can only survive within these confined areas, it may be possible for a few cells to migrate out of these confined areas in practice. Multiple causes might be responsible for this phenomenon. For example, the degrading of the hydrophobic substrate, proteins attaching to the hydrophobic regions, or matrix secreted by cells.

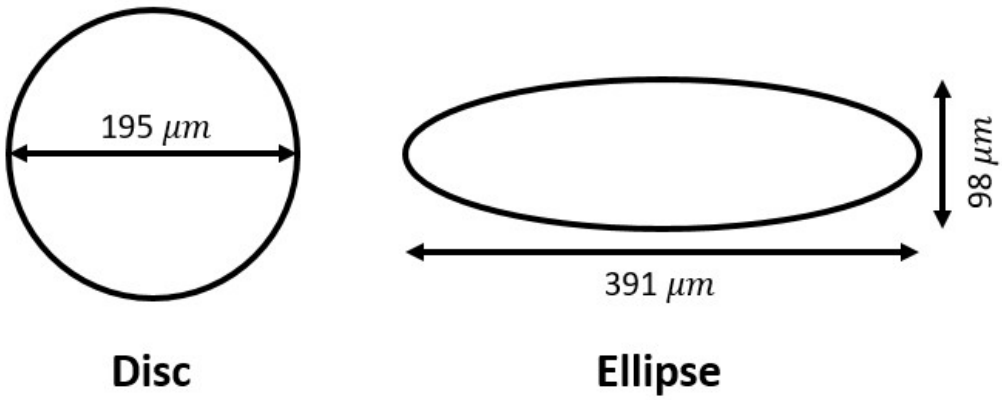


Figure 3-2: The dimensions of disc and ellipse micropatterns

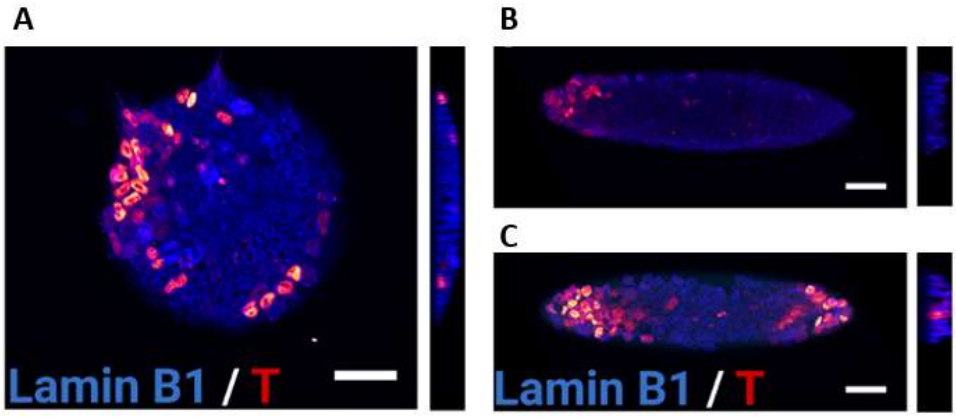


Figure 3-3: An example of the 3D image of ESCs grown on A) disc and B) C) ellipse. B) shows a colony with T+ cells on one side; C) shows a colony with T+ cells on two sides.

3.2 Image processing and cell selection

16-bit images were acquired using a Leica Sp8 inverted scanning confocal microscope using HyD detectors in 'normal' mode. Figure 3-4 illustrates the processes of collecting information from cell images for further data analysis.

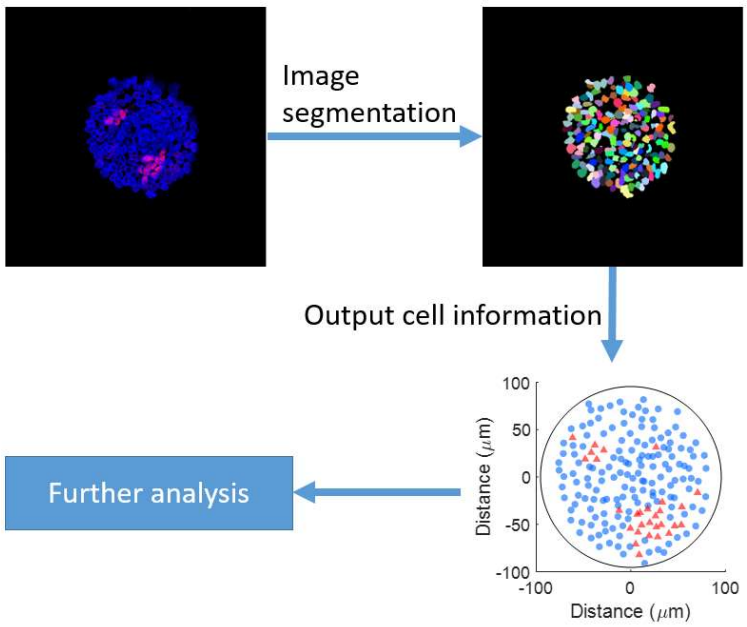


Figure 3-4: An illustration of extracting cell information from images for further processing.

Images were imported inside a custom Java-based application PickCells to perform the following tasks: nuclei segmentation, as well as manual correction of the segmentation, computation of nuclei 3D coordinates, computation of average intensities in colour channels. PickCells is an image analysis platform developed by the Centre for Regenerative Medicine at The University of Edinburgh. Image segmentation was achieved by the Nuclear Envelop Segmentation System (Nessys) (Blin *et al.*, 2019), a tool for the automated segmentation of nuclei in fluorescence images, as well as manual checking. The methodology of collecting data from cell images was summarised in (Wisniewski, Lowell and Blin, 2019).

Imaging settings and image analysis parameters were set for each experiment individually and kept identical for all samples within a specific experiment (Blin *et al.*, 2018). The tables of feature vectors for each cell within each experiment were then exported as comma-separated values (CSV) and stored for future analysis in R/Matlab. After aggregating results from different experiments, we have two CSV files holding cell information for disc and ellipse experiments separately. Cell information includes unique identification for each cell, cell location (x, y and z-axis value), the intensity of the Brachyury (T) channel, and image identification (same identification for cells from the same colony).

We focus on the x and y-axis value in the CSV files we exported as we are interested in the 2D pattern formation with geometrical confinements instead of the 3D dome shape caused by cells overlapping. Cells are marked as T+/T- cells by thresholding the intensity of the Brachyury (T) channel. We selected cells within 95.5 μm from the centre of the disc; for ellipse experiments, we selected cells within the ellipse with the semi-major axis as 193.5 μm and the semi-minor axis as 47 μm. We selected cells within the micropatterns that are more than 2 μm inside the border to obtain a clear cut at the border of the micropatterns. Cells outside these defined constrained areas were considered as random noise and were discarded. Finally, we got two CSV files contains all cell information we interested in this study for all cells on disc and ellipse micropatterns separately. For each cell, we have a unique cell ID, cell location

(x and y-axis), cell type (T+ or T- cell) and an index of which image (colony) it is from.

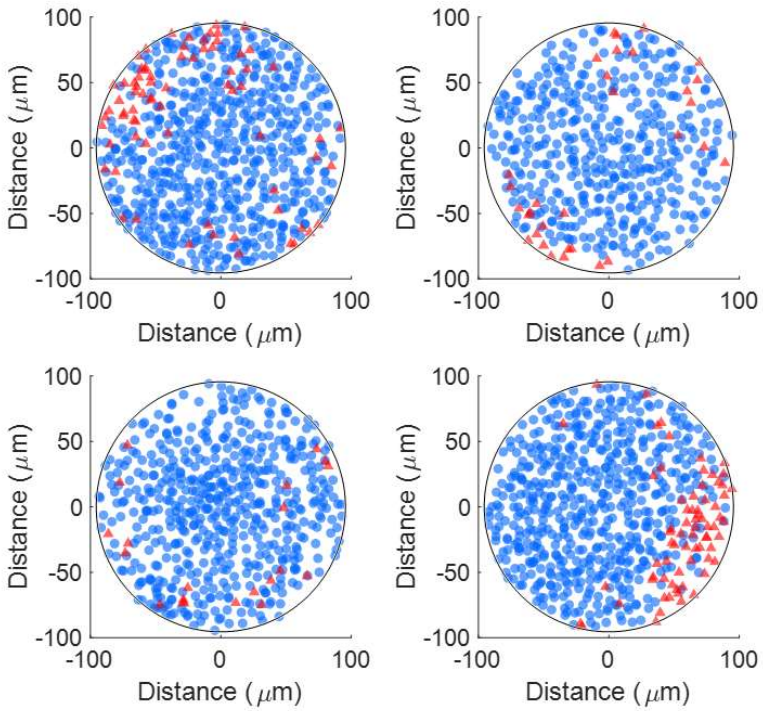


Figure 3-5: Four indicative, randomly selected examples of cell colonies on disc micropatterns. Red triangle markers stand for T+ cells; blue circle markers stand for T- cells.

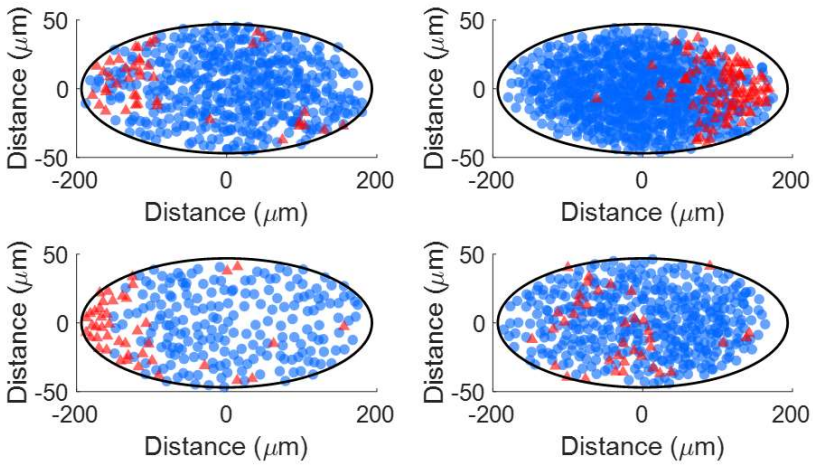


Figure 3-6: Four indicative, randomly selected examples of cell colonies on ellipse micropatterns. Red triangle markers stand for T+ cells; blue circle markers stand for T- cells.

3.3 Visualising the experimental data

Biological colleagues have led the experimental work and have provided us with 186 images (colonies) for disc micropatterns and 152 images (colonies) for ellipse micropatterns.

Figure 3-5 and Figure 3-6 gives four examples of cell colonies on disc and ellipse micropatterns separately, which indicates the high variability in our experimental data regarding cell numbers and patterns. More examples of cell colonies in experimental data are provided in Appendix A. Furthermore, there are not T+ cells on 18 images of disc micropatterns and 4 images of ellipse micropatterns (as shown in Figure 3-7).

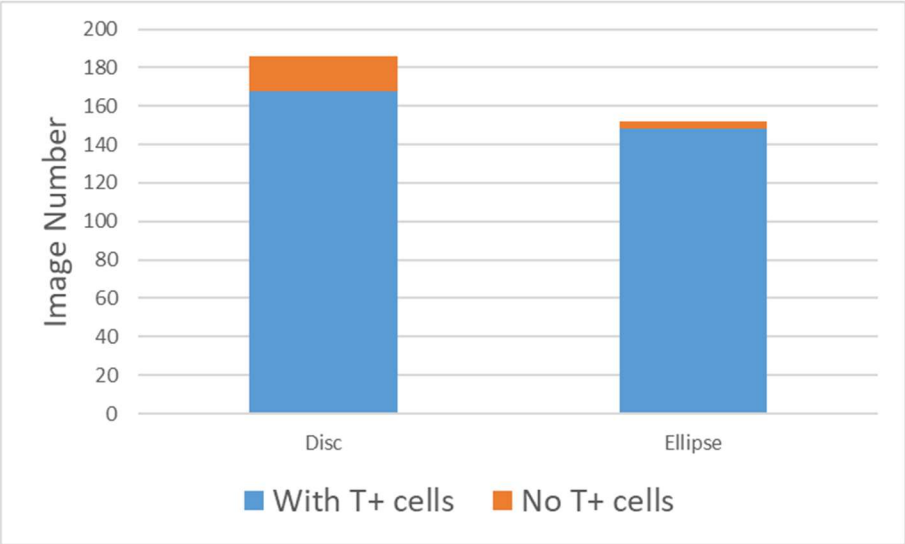


Figure 3-7: Image numbers of disc and ellipse micropatterns with or without T+ cells.

Figure 3-8 shows the distribution of the cell number of T+/T- cells on disc and ellipse micropatterns. On average there are 354 T- cells and 32 T+ cells on disc micropatterns; while for ellipse micropatterns, there are 367 T- cells and 43 T+ cells on average. The mean percentage of T+ cells on disc and ellipse micropatterns is 8.58% and 11.94%. We also observed high variability of cell number for both T- and T+ cells in Figure 3-8.

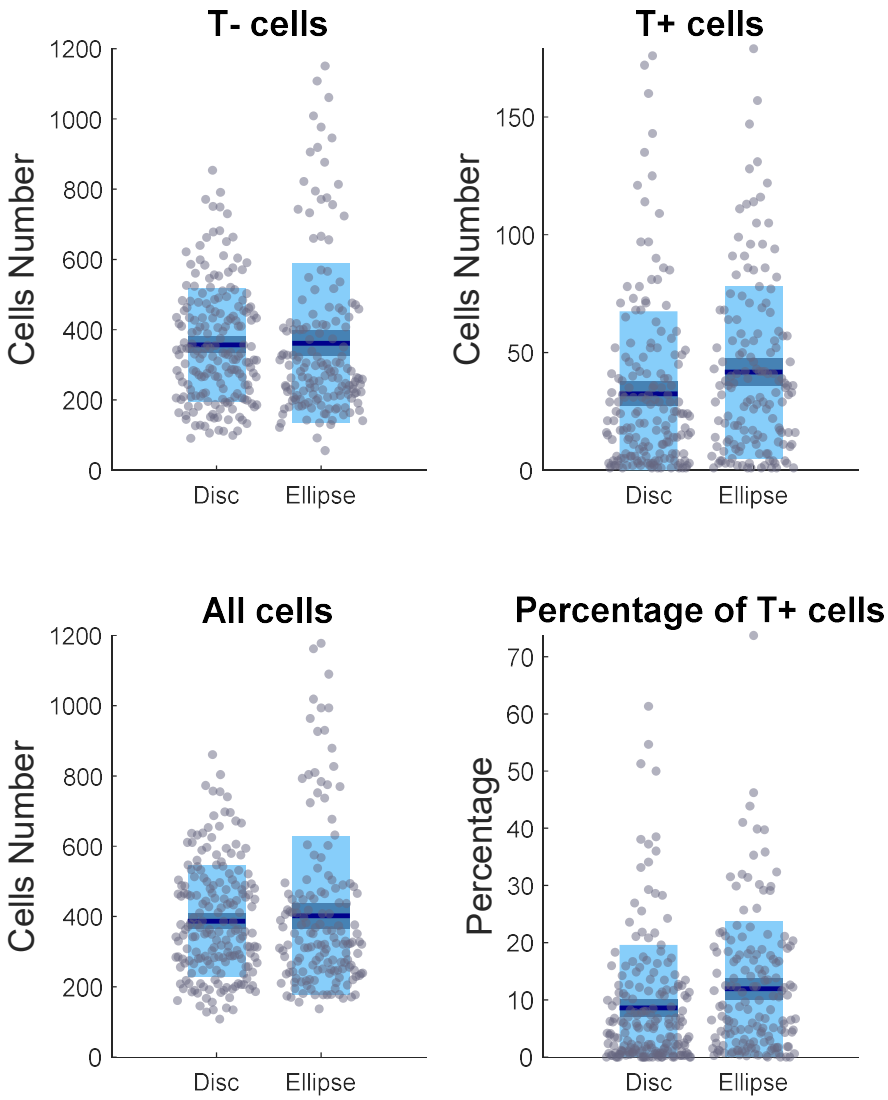


Figure 3-8: Plots of cell number of T- cells, T+ cells, all cells and the percentage of T+ cells on disc and ellipse micropatterns. Scattered points in grey represent the raw data. The dark blue lines stand for the mean of the grouped data, light blue shows 95% confidence interval, 1 standard deviation is also shown with grey-blue. Images were generated by Matlab function notBoxPlot.

In this chapter, we described the seeding and growing conditions of our experimental data supported by Blin's lab. We illustrated the procedures of extracting cell information from the static images collected from the wet lab. By

visualising our experimental data, we observed a high variety of cell numbers as well as cell patterns in both disc and ellipse micropatterns. The scatter plots of 10 cell colonies examples on disc and ellipse micropatterns are shown in Appendix A. More quantitative analysis results of experimental data are provided in Chapter 6.

Chapter 4

4 Data processing methodology

We provided the descriptions of experimental data collection in the former chapter. This chapter provides the algorithmic background on multiple mathematical algorithms which were applied in this study (for either experimental data or simulation outputs), including well-known algorithms and novel approaches. Furthermore, we provide the required background for mathematical algorithms which we tested but not adopted in the end. The detailed implementation and parameter settings of these algorithms will be provided in the section where these algorithms applied.

4.1 Kernel density estimation

As described in Chapter 3, we have a number of images of cells on disc or ellipse-shaped micropatterns. We are interested in different cells preference of localisation, hence, our target is to obtain and present the underlying probability distribution of different types of cells according to their locations. A simple way to achieve this is by generating histograms. A histogram divides the data into discrete bins, counts the number of points that fall in each bin, and then intuitively visualise the results. While the histogram algorithm simply maps each data point to a stack nearby with a fixed area, several issues occur when using histograms to provide an overall density estimate. One is that the choice of the bin size and location of the range can lead to misrepresenting the true underlying data distribution. Another issue is that by taking all points with the same weight in fixed bins, the stacks cannot reflect on the actual density of points nearby but reflects on the coincidences of how the bins align with the data points. Besides, the results of histograms are not smooth.

A better alternative to computing a smoother estimate, which we believe may be more realistic in practice, is with the application of *kernels*. A kernel is a function that specifies the shape of the distribution placed at each point. Therefore, we impose a bell-shaped distribution on each data point and then integrate over these individual data representations to obtain a smoothened distribution.

Our experimental dataset comprises discrete variables which indicate the cell 2D locations. We are primarily focusing on estimating the underlying continuous density function that describes well the randomness of our experimental data. Kernel density estimation (KDE), also known as the Parzen's window (Parzen, 1962), is one of the most well-known approaches to estimate the underlying probability density function of each of the variables in the dataset (Chen, 2017). KDE is a non-parametric density estimator that learns the shape of the density from the continuous real values data automatically. Compared to histograms, KDE is smooth and removes the dependence on the endpoints of the bins, the equal sub-intervals that we use to combine values to get the frequency. Due to the flexibility of KDE, it is a very popular approach for drawing probability density from a complicated distribution.

Let $x_1, \dots, x_n \in \mathbb{R}^d$ be an independent, identically distributed random sample from an unknown distribution P with density function p. Formally, KDE can be expressed as

$$\hat{p}_n(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{|x - x_i|}{h}\right) \tag{4.1}$$

where K is the *kernel*, a smooth non-negative function. $h > 0$ is a smoothing parameter called *bandwidth*, which controls the amount of smoothing. The Gaussian kernel, as described in Equation (4.2), is one of the most commonly used kernel functions since Gaussianity has been assessed over diverse settings (Hastie, Tibshirani and Friedman, 2009).

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \tag{4.2}$$

The bandwidth of the kernel is a free parameter that exhibits a strong influence on the resulting density estimate (Silverman, 1986; Sheather, 1992). However, there is always a trade-off problem of bandwidth selection. If the bandwidth is too small, the resulting estimate would be under-smoothed since it contains too many spurious data; similarly, if the bandwidth is very large, the resulting curve would be over-smoothed since it obscures much of the underlying structure. Hence, bandwidth selection is a classical research topic in nonparametric statistics.

Silverman's rule of thumb is one of the common approaches for bandwidth selection (Silverman, 1986). Even though Silverman's rule of thumb is easy to compute, it can yield widely inaccurate estimates when the density is not close to being normal. In this study, we apply Botev's approach for density estimation (Botev, Grotowski and Kroese, 2010). This adaptive kernel density estimation method is based on the smoothing properties of linear diffusion processes. Because it improved local adaptivity and reduced boundary bias, it increased the accuracy and reliability by reducing the sensitivity to outliers, asymptotic bias and mean square error. This non-parametric bandwidth selection method does not require a preliminary normal model for data. Hence, this approach has high flexibility and accessibility.

4.2 Least-squares fitting

In this study, we generated the border of high-density areas of T+ or T- cells based on our experimental data by the least-squares fitting. High-density areas will be used to describe the observed pattern formation as well as to evaluate our models' performance.

Least-squares fitting is one of the mathematical procedures for finding the best-fitting curve to a given set of points by minimizing the sum of the squares

of the offsets made in each point to the curve. To illustrate the concept, we take linear least-squares fitting of a group of sample points as an example here. As shown in Figure 4-1, we have N sample points and we are looking for a line that best fits them. For each point, there is an error e_N shows the offset (vertical offset) of the point from the line. The best fit line is the line with the smallest value of the sum of the square of the errors, as $min \sum_{i=1}^{N} {e_i}^2$. The fit we used in this study was proposed by Pratt (Pratt, 1987). His spherical fit takes advantage of a special property of circles and spheres that permits robust fitting (to obtain the true best fit).

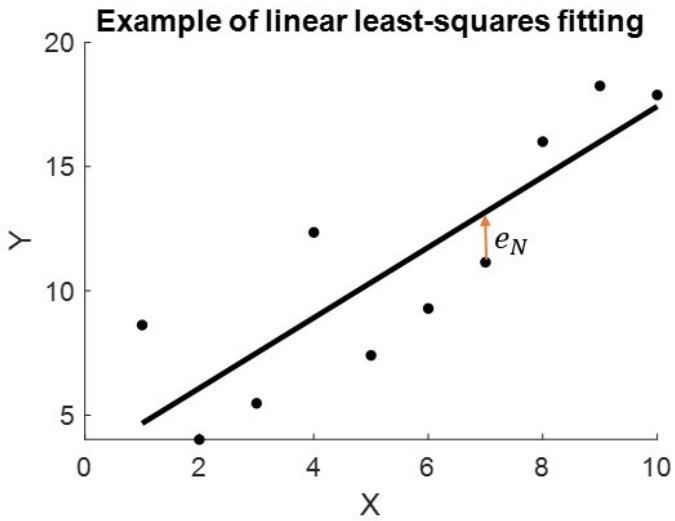


Figure 4-1: An example of linear least-squares fitting

4.3 Proximity measurements

In this study, we observed symmetry breaking pattern formation in geometry confined ESCs on 2 dimensions. Hence, we focus on the 2D pattern observed instead of the 3D dome shape caused by cells overlapping. 2D patterns were produced by projecting 3D data on 2D. We collected the coordinates of cell locations by selecting the x and y-axis and omitting the z-axis.

Blin and colleagues demonstrated that T+ and T- cells prefer different numbers of neighbouring cells, with T- cells preferring more neighbours than T+ cells (Blin *et al.*, 2018). In this study, we extend on these findings and look into the preferred proximity (closeness) of neighbours between the two types of cells. We applied two new measurements to quantify the proximity of T+ and T- cells in mESC colonies and assess the difference of proximity within different patterning constraints. The results from quantifying proximity between different types of cells assist us to propose cell behavioural rules for modelling.

4.3.1 Minimum spanning tree

To illustrate the concept minimum spanning tree, we have a number of sample points localised in a 2D coordinate system without any linkages (shown as nodes in Figure 4-2). As an example, we apply a minimum spanning tree to obtain the shortest path (with Euclidean distance) to connect all sample points.

In graph theory, graphs are formed by taking *nodes* (*vertices*) as fundamental units and each of the related pairs of nodes is called an *edge* (*link*) (Trudeau, 1993). An undirected graph consists of a set of nodes and a set of edges connecting unordered pairs of nodes, while a directed graph consists of a set of nodes and a set of edges connecting ordered pairs of nodes (Bender and Willianson, 2010). A weighted graph is a special type of labelled graph in which each edge is given a numerical weight. A spanning tree of an undirected graph is a sub-graph that connects every node without any cycles (Kruskal, 1956). As there are different spanning trees for a graph, the minimum spanning tree is the spanning tree connecting the nodes through edges and has the smallest total weight (Prim, 1957). That is, it is a spanning tree whose sum of edge weights is minimised.

Hence, in our example, we take each sample point as a node and build a weighted graph by taking the Euclidean distance between any two nodes as the weight of this edge. Figure 4-2 shows the graph of our sample points and

their minimum spanning tree. Therefore, the minimum spanning tree can be used to find the shortest path to connect every node in the graph. There may be more than one minimum spanning tree in a graph. In this study, we used Prim's approach, adding edges to the tree while traversing the graph from the root node, to find the minimum spanning tree of the connected graph.

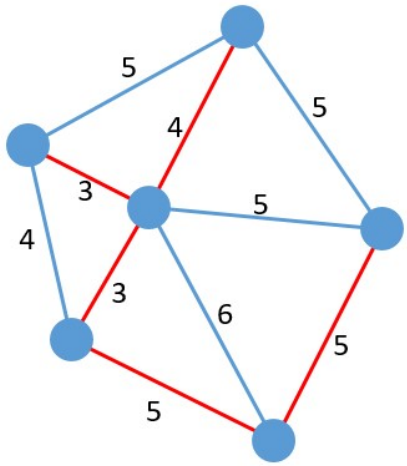


Figure 4-2: An example graph and its minimum spanning tree. Blue circles are nodes connected by edges with labels showing their weights. Red lines show the minimum spanning tree for this graph.

We built a connected graph (Wilson, 1996) for T+ or T- cells in each cell colony by taking each cell as a node and connecting any two nodes with a weighted edge. The weight of each edge was defined as the Euclidean distance between the two nodes. The minimum spanning tree calculates the shortest path that connects all T+ and T- cells within the colony. Therefore, we got the minimum spanning tree of connecting all T- and T+ cells respectively. We calculated the average path distance between two nodes (cells) by dividing the path distance of the minimum spanning tree by the number of T+ or T- cells respectively. A smaller average path distance indicates higher proximity. Following this, we applied kernel density estimation (Botev, Grotowski and Kroese, 2010)

(Section 4.1) of the average path distance we received from T+ and T- cells from each cell colony. The number of mesh points used in the kernel density estimation was 256. We also calculated the minimum spanning tree for T+ and T- cells in each pattern group. The results of applying the minimum spanning tree to measure the proximity among different cell types will be provided in Section 6.2.1.

4.3.2 Quantifying average distance for each query object to five nearest targets

In addition to the minimum spanning tree, we also applied a new measurement to assess the proximity between different types of cells. For each T+ or T- cell (i.e., object) within the disc and ellipse micropatterns, we found the five closest T+ and T- cells (i.e., targets). We calculated the average distance from each object to its targets (referred to as D). The data presented four different cases of calculating D: 1) the object is a T- cell with T- cell targets; 2) the object is a T- cell with T+ cell targets; 3) the object is a T+ cell with T- cell targets; 4) the object is a T+ cell with T+ cell targets. For these four cases, we applied KDE (Botev, Grotowski and Kroese, 2010) (Section 4.1) of D from all cells within each pattern group. Based on the borders of the HDA, we applied kernel density estimation of the cells in the HDA and cells outside the HDA separately to investigate the difference in proximity between the different cell types in both regions. The results will be provided in Section 6.2.2.

4.4 Evaluation metrics: comparing probability distributions to assess probabilistic estimates against some known ground truth

In this study, we use the experimental data (as described in Chapter 3) as ground truth. Since we have a series of images of cells growing in confined

areas, our data is probabilistic due to the nature of the data. Similarly, we constructed probabilistic models to test cell behaviours. Therefore, we are looking for an evaluation metric to assess probabilistic estimates by comparing probability distributions from model outputs against the probability distributions from the experimental data.

In the following sections, we summarise a range of established evaluation approaches in similar settings and also explain the novel performance metric we proposed. To explain the motivation of developing a new performance metric for the needs of this particular application, we applied these existing evaluation approaches to the artificial data to demonstrate the problem.

4.4.1 Known evaluation metrics

In this study, we have multivariant distributions from experimental data and model outputs describing the allocations of cells. The performance of models was evaluated by comparing probabilistic distributions from model output and experimental data. To assess the performance of our models, a series of known evaluation metrics were tested.

4.4.1.1 Kullback-Leibler divergence

The Kullback-Leibler divergence (KL divergence) (Kullback and Leibler, 1951) is commonly used for comparing two probability distributions. *Divergence* is a function that quantifies the differences of one probability distribution to the other in statistics. Different from the notion of *distance*, the divergences need not be symmetric. That is, assume we have two probability distributions P and Q, the divergence from P to Q is not equal to the divergence from Q to P, and need not satisfy the triangle inequality. The divergence between p and q is represented as $D(P \parallel Q)$.

The KL divergence is the most commonly used divergence. Since it is based on Shannon's foundations on information theory and the definition of entropy

(Shannon, 1948), it is also called *relative entropy.* For probability distributions P and Q defined on the same probability space, Ω, Equation (4.3) shows the definition of the KL divergence from Q to P.

$$D_{KL}(P \parallel Q) = \int P(x) \cdot log(P(x)/Q(x))dx \tag{4.3}$$

Hence, a smaller value of KL divergence indicates that the difference from one distribution to another is smaller. A KL divergence of 0 indicates that one distribution in question is identical to another distribution. In this study, we take the sum of the divergence from P to Q and the divergence from Q to P as the final result based on KL divergence.

4.4.1.2 Earth mover's distance

In statistics, the earth mover's distance (EMD) is a method to evaluate dissimilarity between two multi-dimensional probability distributions. Intuitively, given two distributions, which can be seen as two different ways of a certain mass of earth spreading in the space. The EMD measures the minimum amount of work needed to turn one pile of earth into the other.

EMD calculates the distance between two distributions represented by a set of grouped samples, called the *signature*. For a set of points in $\mathbb{R}^d$ (dimension d), each grouped sample is a single point in $\mathbb{R}^d$ and has its own weight. For calculating EMD, two signatures can have different sizes.

Computing the EMD is based on a solution to the well-known transportation problem (Hitchcock, 1941). Formally, let P be the first signature with m grouped samples, where x_i is the grouped sample representative and w_{x_i} is the weight of the cluster as shown in Equation (4.4); Q be the second signature with n clusters as shown in Equation (4.5).

$$P = \{(x_1, w_{x_1}), \dots, (x_m, w_{x_m})\} \tag{4.4}$$

$$Q = \{(y_1, w_{y_1}), \dots, (y_n, w_{y_n})\} \tag{4.5}$$

Let $D = [d_{ij}]$ be the ground distance matrix where d_{ij} is the ground distance between samples x_i and y_j; $F = [f_{ij}]$ be the flow matrix where f_{ij} stands for the flow between x_i and y_j. The overall cost can be calculated according to Equation (4.6) with the constrains shown in Equations (4.7) to (4.10). We want to find a flow F that minimizes the overall cost.

$$WORK(P, Q, F) = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} d_{ij} \tag{4.6}$$

$$f_{ij} \geq 0, 1 \leq i \leq m, 1 \leq j \leq n \tag{4.7}$$

$$\sum_{j=1}^{n} f_{ij} \leq w_{x_i}, 1 \leq i \leq m \tag{4.8}$$

$$\sum_{i=1}^{m} f_{ij} \leq w_{y_j}, 1 \leq j \leq m \tag{4.9}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min\left(\sum_{i=1}^{m} w_{x_i}, \sum_{j=1}^{n} w_{y_j}\right) \tag{4.10}$$

(4.7) allows moving earth from P to Q and not vice versa. (4.8) and (4.9) limits the amount of earth they can move from P to their weights, and the amount of earth they can receive in Q. (4.10) forces to move the maximum amount of earth possible, called the *total flow*. Once the transportation problem is solved with finding the optimal flow F, the earth mover's distance as the work normalised by the total flow as shown in Equation (4.11).

$$EMD(P, Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} d_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} \tag{4.11}$$

4.4.1.3 Bhattacharyya distance

Besides EMD, we also calculated the Bhattacharyya distance between our experimental data and model outputs. Bhattacharyya distance is a measurement of the similarity of two probability distributions. For probability distributions A and B over the same domain X, Equation (4.12) defines the Bhattacharyya distance. $BC(A,B)$ as shown in Equation (4.13) is the Bhattacharyya coefficient for discrete probability distributions.

$$D_B(A,B) = -\ln\big(BC(A,B)\big) \tag{4.12}$$

$$BC(A,B) = \sum_{x \in X} \sqrt{A(x)B(x)} \tag{4.13}$$

4.4.1.4 Continuous ranked probability score

The continuous ranked probability score (CRPS) is a widely used measure of performance for probabilistic forecasts of a scalar observation. The CRPS can be calculated to assess the respective accuracy of two probabilistic forecasting models based on generalizing the mean absolute error. Assume we have a random variable X, F stands for the cumulative distribution function (CDF) of X. Hence, the measured CDF (the CDF associated with the empirical data) is as shown in Equation (4.14).

$$F(y) = \mathrm{P}[X \leq y] \tag{4.14}$$

Let x be the observation. The CRPS between x and F is defined as in Equation (4.15), where $\mathbb{1}$ is the Heaviside step function and denotes a step function along the real line that attains. If the real argument is positive or zero, then the value would be 1, otherwise, the value would be 0.

$$CRPS(y,x) = \int_{-\infty}^{\infty} \big(F(y) - \mathbb{1}(y-x)\big)^2 dy \tag{4.15}$$

4.4.2 Comparing the conceptual basis of the performance metrics

To illustrate the motivation of proposing a novel evaluation metric, we carried out CRPS, EMD, Bhattacharyya distance and KL divergence on simple artificial data. The artificial data we generated shows the natural problem of our experimental data which contains a high variability.

Suppose we have 10 datasets for ground truth, each holding 100 sample numerical data within a one-dimensional space. The sample data in ground truth were random numbers generated from a normal distribution with a mean of 50 and a standard deviation of 15. We choose a big value of standard deviation to represent the high variety in our experimental data. Figure 4-3 (A) shows the histogram of aggregating all sample points from 10 datasets. Figure 4-3 (B) shows the density plot after applying the KDE of aggregated sample points.

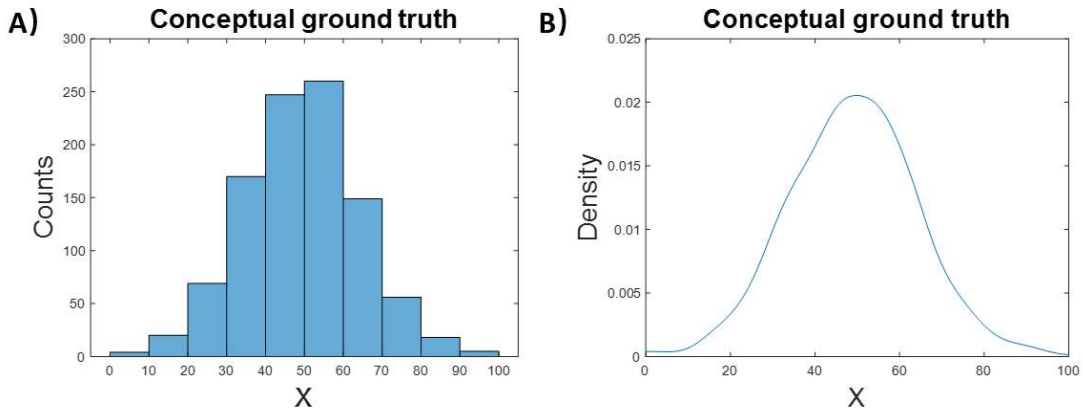


Figure 4-3: A) histogram and B) density plot of the aggregated sample points taking as ground truth in artificial data.

Subsequently, we generated 10 datasets for the random sample and 10 datasets for the comparison sample. Same as the conceptual data for ground truth, each dataset holds 100 numerical data. In a random sample, data were randomly generated from 0 to 100. In comparison sample, data were generated randomly from a normal distribution with a mean of 50 and a

standard deviation of 1. The histograms and density plots (after KDE) of the random sample and comparison sample are shown in Figure 4-4 and Figure 4-5.

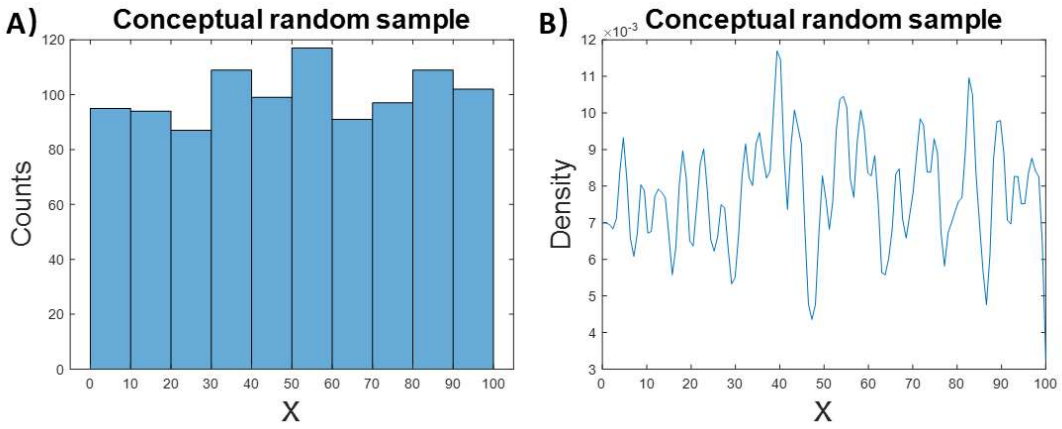


Figure 4-4: A) histogram and B) density plot of the aggregated sample data in the random sample.

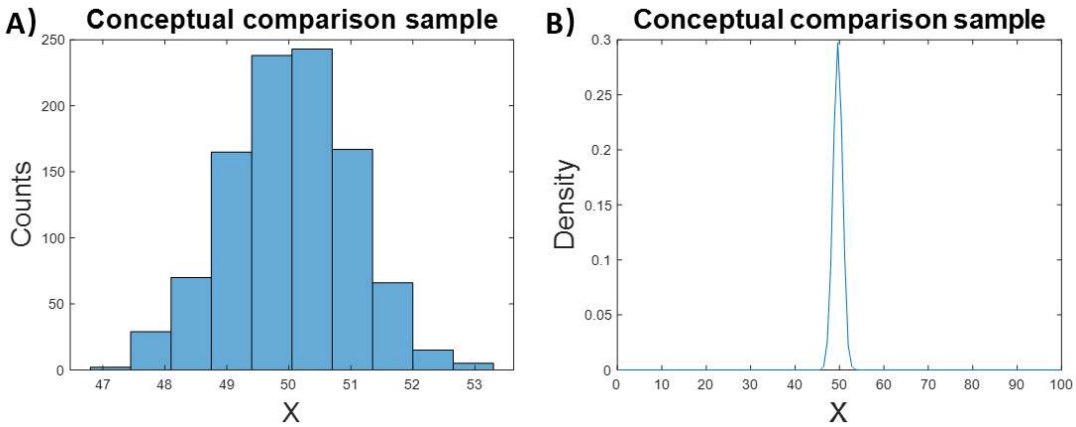


Figure 4-5: A) histogram and B) density plot of the aggregated sample data in the comparison sample.

We calculated KL divergence, EMD, Bhattacharyya distance and CRPS of random sample and comparison sample against ground truth separately. In addition, we tested these measures by comparing two random samples. Table

4-1 shows the results of these existing evaluation metrics. Based on these results, the random sample is closer to the ground truth than the comparison sample.

Table 4-1: Evaluation results from existing evaluation metrics carried out on artificial data.

	Random sample versus ground truth	Comparison sample versus ground truth	Between two random sample
KL divergence sum	0.0092	0.1413	4.6757e-05
EMD	12.8537	11.0677	0.1806
Bhattacharyya distance	0.1434	0.9879	7.4809e-04
CRPS	0.0060	0.0128	0.0022

It is understandable that random sample achieves better performance with these metrics because of the high standard deviation in the ground truth and the slight differences between the mean value in ground truth and comparison sample. To provide better illustrations, we carried out these measures on a series of ground truth data with different standard deviations (as shown in Figure 4-6). With big standard deviations, these metrics would suggest that the random sample is closer to the ground truth data compared to the comparison sample. However, the comparison sample holds a similar pattern as in ground truth. Furthermore, we tested these metrics on a series of ground truth datasets with different standard deviation (from 1 to 20). Figure 4-6 shows the results based on these metrics for both random and comparison samples. The performance of these metrics decreases while the standard deviation in ground truth is increasing.

By taking these conceptual data as an example, we illustrated that these performance metrics are not compatible with this study. The main reason is that even though we do observe a specific pattern formation in cells on an aggregated level, the variety in our experimental data is extremely high. More explanation of the motivation of proposing a novel metric will be provided in Section 7.1.

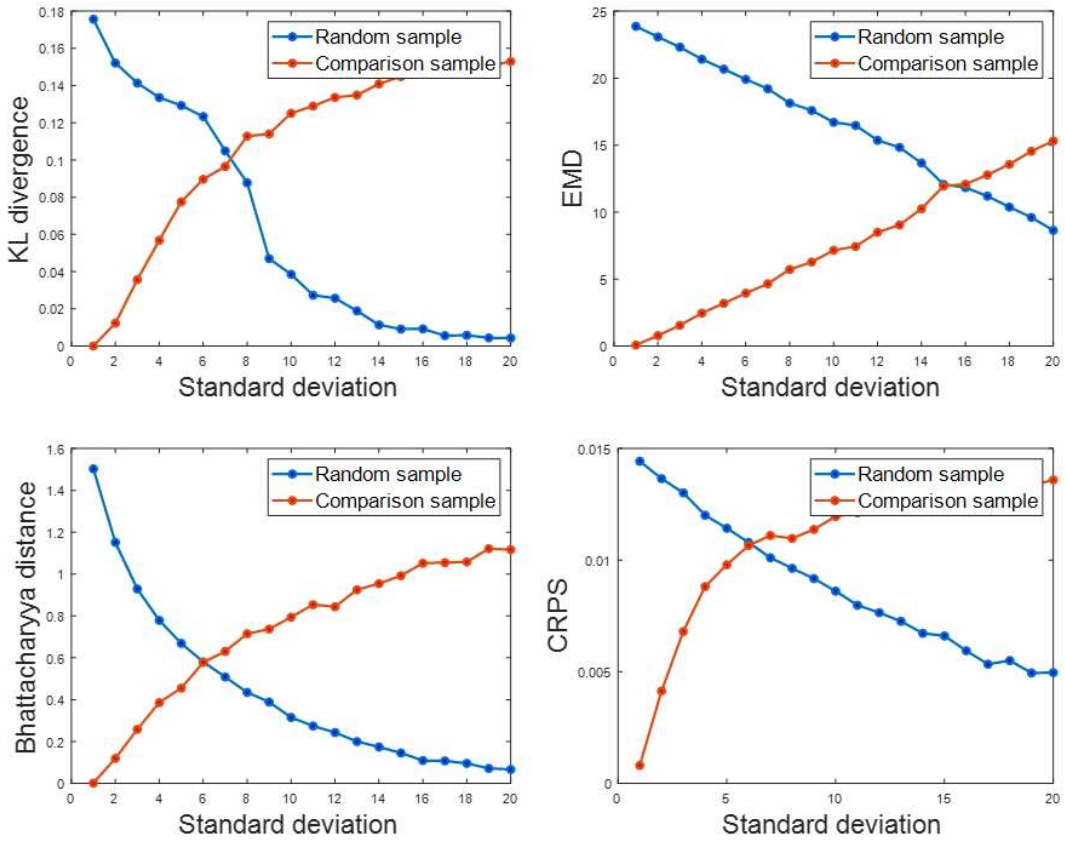


Figure 4-6: Evaluation results based on KL divergence, EMD, Bhattacharyya distance, and CRPS for random and comparison sample against the ground truth with different standard deviation.

4.4.3 Novel performance metric comparing probabilistic density estimates and probabilistic ground truth: the stem cell aggregate pattern distance (SCAPD)

The existing evaluation metrics are not providing intuitively appealing results in the particular problem investigated in the study. The underlying cause is that the ground truth is stochastic, and hence we need to compare probabilistic estimates of models and the ground truth which is in the form of a probability distribution (this will become clear when describing the underlying data in Section 6.1). Due to the high variety in our experimental data, the results of model evaluation based on existing metrics show that the random model is a very competitive model which is not consistent with our visual impression (as shown in Section 7.1).

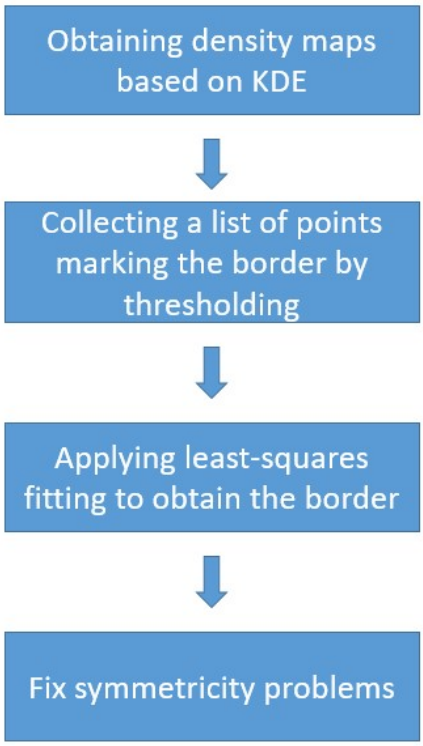


Figure 4-7: The procedures of obtaining High-Density Areas (HDA) borders in SCAPD.

We developed a new metric to quantify the difference in the constellation of the cell patterns between the model results and the empirical data, which we can the *stem cell aggregate pattern distance* (SCAPD) (Wang *et al.*, 2020). The metric takes both types of cells into account (T+ and T- cells in this study). Based on the empirical data, we generated the border of high-density areas (HDA) of T+ and T- separately, which is the defined areas for evaluation. Figure 4-7 shows the procedures for generating HDA. We firstly apply KDE on aggregated cell location data for T+/T- cells. Based on the density maps, we thresholded the grid space to collect a list of points marking the border of the areas. The threshold we used is the mean of the maximum and minimum value of the density of each grid point. Subsequently, we generated the border by applying the least-squares fitting and fixed the symmetricity problem. We generated circular borders for disc micropatterns and elliptical borders for ellipse micropatterns. In our case, we believe the HDA in both disc and ellipse micropatterns should follow the symmetrical properties of disc and ellipse. Because all cells were seeded randomly and there is no fundamental reason to suggest the micropatterns should exhibit particular preference in any direction. The steps of getting the borders of HDA will be provided in Section 6.1.2.

Based on the borders of HDA in experimental data, we calculated the total density within these areas for T+ and T- cells separately in our experimental data on both disc and ellipse micropatterns. The total density is the sum of the density of the grid points from the kernel density estimation results within the defined borders. Hence, for experimental data, we got the total density within the area for both T+ and T- cells noted as t_{e_T+} and t_{e_T-}.

In this study, we take the experimental data collected from the wet lab as ground truth. For the evaluation of the models, we compare the simulation results against the experimental data. To do so, we firstly applied kernel density estimation using the same parameters and approaches we applied in experimental data. Subsequently, with the borders we generated from the experimental data, we calculated the total density within the defined borders

of HDA on our simulation results as well. We run each model 100 times to take into stochasticity effects, and calculated the total density of the aggregated 100 runs results from both T+ and T- cells separately noted as t_{m_T+} and t_{m_T-}. We calculate SCAPD to express the difference observed between the model results and the experimental data (used as ground truth), as follows:

$$SCAPD = \left|t_{e\_T+} - t_{m\_T+}\right| + \left|t_{e\_T-} - t_{m\_T-}\right| \tag{4.16}$$

In the next step, we got the sum of the absolute difference of T+ and T- cells total density within the area for both disc and ellipse colonies. A small sum of absolute total density difference indicates the model is close to the experimental data (if completely matching the empirical data, then the absolute difference would be exactly zero).

Table 4-2: Summary table of multiple existing and novel metrics of assessing probabilistic estimates against some known ground truth

Metrics	Descriptions
KL divergence sum	Indicates the difference between two probability distributions by comparing each grid point on the same probability space.
EMD	Measures the minimum amount of work needed to turn one pile of earth (one probability distribution) into the other.
Bhattacharyya distance	Measures the similarity of two probability distributions by measuring the amount of overlap between two distributions.
CRPS	Assesses the respective accuracy of two probabilistic forecasting models based on generalizing the mean absolute error.
SCAPD	Calculates the distance between two aggregated density maps according to defined areas. By defining areas, this metric emphasizes the desired pattern on an aggregated level.

Here, we provide a summary table of the descriptions of the known evaluation metrics we tested in this study as well as the novel performance metric SCAPD in Table 4-2.

In this chapter, we explained the mathematical algorithms we applied in this study. We described the existing metrics we tried for model evaluation in this study. We illustrated the necessity of the new evaluation matric in this study by taking simple artificial data as an example. Afterwards, we explained the novel performance metric we proposed in this study. The details of applying these evaluation metrics in this study including the parameter settings are described in Section 5.5. The evaluation results based on these metrics will be provided in Chapter 7 including a further explanation of the motivation for proposing a novel metric in Section 7.1.

Chapter 5

5 Model description and optimisation

We had previously described the data available in this study (Chapter 3) and outlined the key methodological approaches towards quantitative analysis (Chapter 4). This chapter describes the detailed setup we used aiming to develop biologically plausible rules to model the behaviour of stem cell pattern formation.

We used agent-based techniques to develop models of stem cell behaviour and study the specific pattern formation we observed in ESCs. Following the basic concepts in agent-based modelling, we constructed our models from scratch. In this chapter, we firstly provide detailed explanations of the assumptions of modelling. Subsequently, we describe our models in natural language (for understandability) and as a formal specification in logic (for precision). We illustrate the biologically plausible rules we tested in our models including the reasons for proposing the rules. The descriptions of the user interface, the list of models, and models optimisation methods are also provided in this chapter.

5.1 Assumptions of modelling

The interesting pattern formation in empirical data is mainly on 2 dimensions as we are interested in the potential power that drives symmetry breaking in geometry confined ESCs. We focus on the 2D pattern observed in empirical data instead of the 3D dome shape caused by cells overlapping. 2D patterns were produced by projecting 3D data on 2D. We collected the coordination of cell locations by selecting the x and y-axis and omitting the z-axis. Therefore,

we constructed 2D models to reproduce the observed pattern formation in ESCs.

In our models, we assume cells only migrate within the confined areas because we take the cells migrating away from the confined areas is due to some uncontrollable factors in the wet lab. We assume cells can move freely in the unoccupied environment with a consistent speed and do not model the adhesion forces between cells and their environment.

The project aims to construct a minimal model to investigate the contribution from cell motility to pattern formation, therefore, we omit the contribution from different shapes of cells, cell differentiation, division and apoptosis.

5.2 Model construction

To capture the social behaviours that lead cells to the observed pattern formation, a set of agent-based models was constructed to access the different hypothesized behaviours of embryonic stem cells and their combinations. Even though we can construct agent-based models in any programming language from scratch, there are many existing software packages. Most of the commonly used ABM platforms follow the “framework and library” paradigm, providing a framework, a set of standard concepts for designing and describing ABMs, along with a library of software implementing the framework and providing simulation tools. (Allan, 2010) provided a report outlining over 30 packages of agent-based modelling. Among these packages, MASON, NetLogo, Repast, and Swarm are popular choices. In this study, models were constructed in NetLogo 6.0.4, an open-source environment specifically designed for agent-based modelling. The source code is available online at https://github.com/MinhongW/ESCs_models.

We constructed our models in NetLogo in this study because it is a simple yet powerful programming language, and provides built-in graphical interfaces and comprehensive documentation. It is particularly well suited for modelling

complex systems developing over time. As it is one of the popular choices of agent-based modelling, NetLogo has extensive documentation and tutorials and also comes with a models library. NetLogo come with a user-friendly interface that we can customise with lots of buttons, switches, sliders and monitors. These interface elements allow us to interact with the model, including setting up parameters of the model. We also used an existing toolbox that allows linking NetLogo with R and Matlab. Besides, NetLogo allows clustered computing by providing an integrated software tool *BehaviorSpace*.

Each model consists of a set of cells and their environment, which is a two-dimensional area constrained to a disc or an ellipse (representing disc or ellipse micropatterns). Hence, there are two types of agents in our models: cells and the environment. Each type is characterized by its own parameters. The environment is divided into a grid of square patches. Each patch represents a 1 μm^2 spatial domain that a cell can adhere to and migrate in. The size of disc and ellipse environments were described in Section 3.1. In addition, cells only can immigrate to the patches within the confined area that are not occupied by other cells yet. Besides migrating on the environment and interacting with the environment, cells are agents that can interact with each other as well. Cells have multiple parameters such as location, cell type, and direction etc. The parameters of the environment and cells and their explanations are listed in Table 5-1.

To provide a precise specification of our models, we applied Horn clauses to support our model description. Horn clauses are named after the logician Alfred Horn, who first pointed out their significance in 1951 (Horn, 1951). In logic programming, a Horn clause is a logical formula of a particular rule-like form which gives it useful properties for use in logic programming, formal specification, and model theory.

Table 5-1: List of terms and corresponding descriptions of different types of agents.

Agent types	Terms	Descriptions
Environment	Inside/outside	Inside or outside of the micropattern
	Occupied/Not occupied	Stands for the patch is occupied by a cell or not
Cell	Location	The location of the cells' centre point (consist of x and y)
	Closest distance to neighbouring cells	The closest distance cells can be, same value for T+ and T- cells (set as 2 μm due to projecting 3D in real-world to 2D in models)
	Cell type	T+ or T- cells
	Mean velocity (v)	The mean velocity of cells (As the value of velocity have been measured experimentally, we set 100 $\mu m/h$ for T+ cells and 40 $\mu m/h$ for T- cells (Turner, Rue, *et al.*, 2014; Phadnis *et al.*, 2015))
	Velocity ratio	The ratio of mean velocity to obtain the actual velocity, more details in Equation (5.9)
	Directional persistence time	The time that migratory cells spend without changing direction (105 minutes for T+ cells and 15 minutes for T- cells), inspired by (Mori *et al.*, 2009)
	Direction	The heading direction of cells
	Sensing radius (R)	How far away the neighbouring cells they can sense for chemical signals
	Standard deviation (σ)	For generating the variance of receiving mechanical forces in real life (e.g. from the different shape of the cells or variance caused by some diffusible signals we do not understand yet)
	Angle change (α)	The angle cells change when they reach the border of the micropattern (we tested 10, 20, 30, and 40 degrees in this study)

Horn clauses are constructed by literals. A literal is an atomic formula or its negation. We will provide a detailed explanation of the literals we defined in the formal specification of models. In Horn clauses, the normal logical operators are being used. For example, the operators ←, ⋀, and ⋁ are the connectives for implication, conjunction and disjunction. In addition, we also use [ ] to represent an empty list and {} to represent an empty set.

Our models are state-based and rule-based. Specifically, the model iterates through time with state transitions. There are 192 states for each model, equal to 48 hours in real life, which means the model iterates through time with one state transition that stands for 15 minutes. For each state transition, cells move around autonomously based on the rules we set for testing. Equation (5.1)-(5.6) shows the formal specification of the state transitions based on horn clauses.

$$state(1, S) \leftarrow initial\_cells(S) \tag{5.1}$$

$$state(N, S_n) \leftarrow N_p = N - 1 \; and \; state(N_p, S_p) \; and \; transitions(S_p, S_p, S_n) \tag{5.2}$$

$$transitions(S, [\,], [\,]). \tag{5.3}$$

$$\begin{gathered} transitions(S, S_p, S_n) \leftarrow \\ select(S_p, C_p, R_p) \; and \; transition(S, C_p, C_n) \\ and \; transitions(S, R_p, R_n) \; and \; add(R_n, C_n, S_n) \end{gathered} \tag{5.4}$$

$$select(S, X, R) \leftarrow S = (R \cup \{X\}) \tag{5.5}$$

$$add(S, X, S_n) \leftarrow S_n = (S \cup \{X\}) \tag{5.6}$$

A $state$ is specified by variables N and S. N is a positive integer indexing the state and S is the set of cells as $[cell(Id, Atts), \dots]$, where Id is the cell identifier and $Atts$ is its list of attributes in the state.

Equation (5.1) specifies the initial state with $initial\_cells(S)$ defines the initial set of cells $[cell(Id, Atts), \dots]$. The pseudocode of seeding cells is provided in Appendix B. In state 1, cells are seeded randomly within the constrained areas to test the effect of cell motility on pattern formation. 32 T+ cells and 354 T- cells are seeded on disc micropatterns, and 42 T+ cells and 367 T- cells are seeded on ellipse micropatterns. The number of cells is equal to the mean cell number in empirical data. We seed cells randomly distributed at the initial state to test the effect of cell motility on pattern formation.

Equation (5.2) illustrates the iterations of the state transitions. $transitions(S, S_p, S_n)$ is specified in Equation (5.4), where S is the full current state, S_p is the set of cells yet to be transitioned and S_n is the new set of transitioned cells. $transition(S, C_p, C_n)$ shows the iterations of the list of cells to ask each cell move based on the predefined rules, which will be described in the following section. S is the full current state again, C_p is the current cell definition, and C_n is the new definition when transitioned. Equation (5.3) terminates the looping. In summary, all models we tested has the same structure for state transitions and we loop through the set of cells in each state. Each model has its own $transition$ function contains the predefined rules for cell immigration to test. The pseudocode of $transition$ function is provided in Appendix B.

5.3 Biologically plausible rules

The agent-based models are the structural framework that we use for modelling ESCs. The aim is to develop models which at some level approximate the underlying biological processes of stem cell pattern formation, and hence provide some kind of mechanistic insights. Therefore, we need to develop some empirical rules that we believe might explain some of the properties we observe in the data. Specifically, we proposed four new biologically plausible rules of cell motility that might be the potential underlying

movement stimulus on cell level that lead to ESCs pattern formation. Based on former studies and experimental observations, the four proposed biologically plausible rules are: 1) velocity, 2) directional persistence time, 3) directional movements based on neighbouring cells, and 4) border effects. Figure 5-1 shows the diagram illustrating these rules.

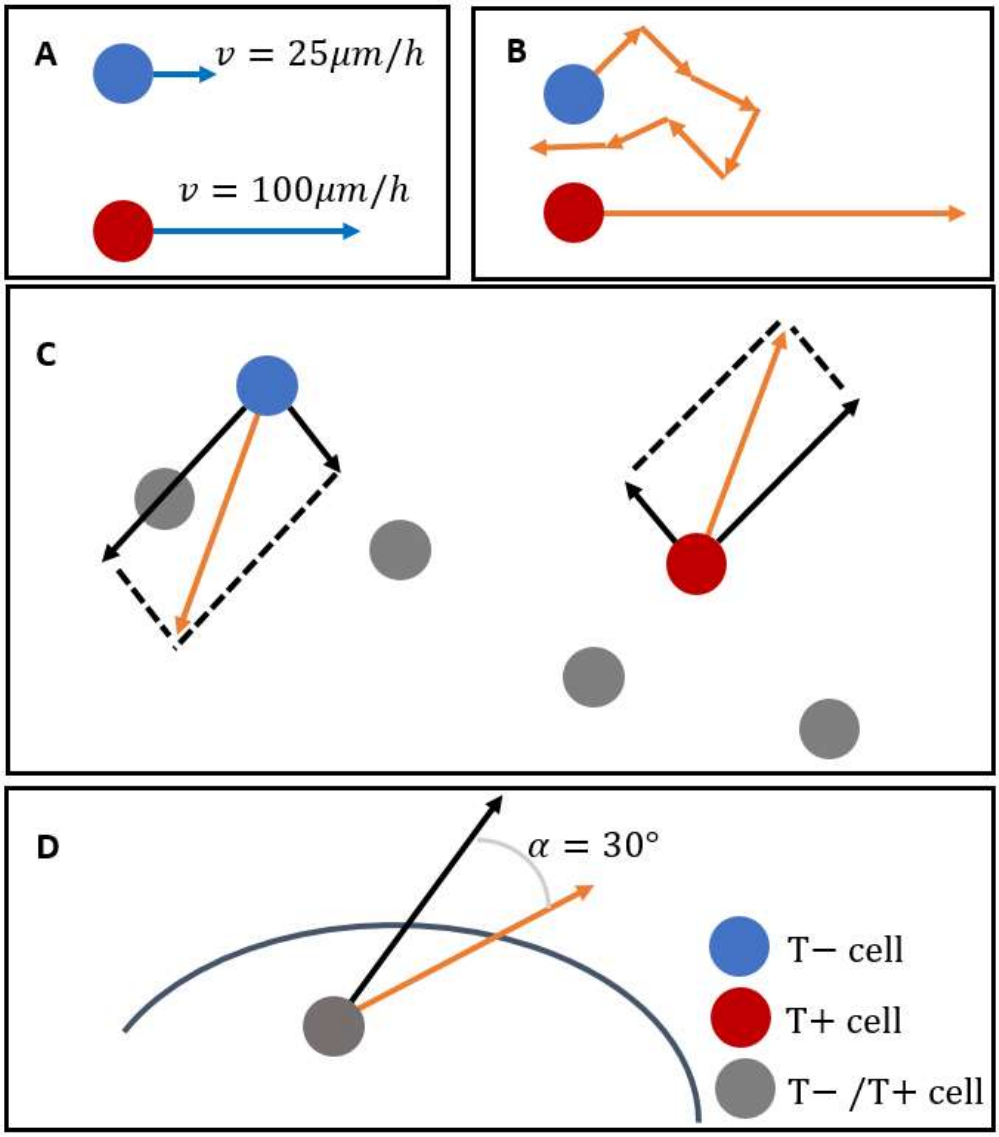


Figure 5-1: Illustration of rules: A) different velocity of T+/T- cells; B) T+ cells have higher directional persistence time; C) directional movements decided by neighbouring cells for T+/T- cells; D) border effect of cells. Blue circles stand for T- cells, red circles stand for T+ cells, and grey circles stand for both T+ and T- cells. Blue arrows stand for velocity (without direction), orange arrows stand for actual direction, and black arrows stand for forces received from neighbouring cells within a distance.

1) Different velocity of T+/T- cells:

   Earlier work (Turner, Hayward, *et al.*, 2014) reported that differentiated cells have a higher velocity than naïve cells. Following the studies of (Turner, Hayward, *et al.*, 2014) and (Phadnis *et al.*, 2015), the mean velocities of T+ and T- cells in our models were set as 100 $\mu m/h$ and 40 $\mu m/h$ repectively.

2) T+ cells have higher directional persistence time:

   Mori et al. demonstrated that directional persistence time plays an important role in sorting MMP14-expressing cells to the two ends of engineered mammary ducts (Mori *et al.*, 2009). Inspired by their study, one of our hypothesized rules is that T+ cells have a higher directional persistence time that T- cells. Persistence time is the threshold of time that the cell will continue migrating in its current direction. This means that T+ cells change their migration direction less often than T- cells. Mori et al (Mori *et al.*, 2009) give the indication of the mean persistence time. We set the persistence time as 7 states (7×15 = 105 minutes) for testing.

3) Directional movements:

   Based on the information from Blin's observation in wet labs as well as the results from quantifying the proximity between different types of cells (Section 6.2), we hypothesize that cells decide directional movements based on local relative densities of cell populations. Hence, instead of random movements, cells decide their directions based on neighbouring cells within a distance. T-cells preferentially move towards neighbouring cells, while T+ cells preferentially move away from the neighbouring cells. Cells get forces from the neighbouring cells within a distance. We calculate the force according to the standard equation that is generically applicable in nature where the exerted force is inversely proportional to the square of the distance (e.g. this concept finds wide

applicability across fields, from gravitational laws in physics (Feynman, 1965) to electrical charge (Malvino and Bates, 2016)):

$$F_{x,y} \sim \frac{1}{d^2} \tag{5.7}$$

where $F_{x,y}$ is the force with x and y representing the direction of the force. d is the Euclidean distance between cells. The forces of all neighbouring cells are summed as vectors, which means having computed all forces from neighbouring cells, we summed up the forces by orthogonal decomposition (sum of forces on x-axis and y-axis). Then we computed the final force with a summarized direction. Since there may exist a level of randomness in the experimentally collected data, the final direction is generated by the normal distribution of the final sum force on the x-axis and y-axis with a specific standard deviation (σ). We applied the same equation to T- cells and T+ cells but with a different direction for vectors. For T+ cells, each neighbouring cell exerts a force pushing it away, whilst T- cells experience attractive forces from neighbouring cells. In this way, T+ cells move toward lower density and T- cells move toward higher density.

4) Border effects:

We hypothesize that cells can sense their environment and once they sense they are on the border of the constrained area, they prefer to stay within the area rather than escaping the area that they can survive on. Hence, with this rule, cells change direction to a small degree once they reach the border. Especially, if the cells' next location, based on the current direction and speed, is outside of the constrained area, cells will change their heading direction with a small angle (α). We tested multiple values and demonstrated that our models are not sensitive to this parameter (as described in Section 7.2). By comparing the current direction and the direction to the closest patch outside of the constrained area, the cell will decide the angle change is clockwise or counterclockwise.

5.4 Model descriptions

Following the definitions of the physiologically plausible rules, we subsequently tested all possible combinations of these four rules with 16 agent-based models. The list of the model numbers with corresponding rules is shown in Table 5-2 and Table 5-3.

Table 5-2: Four proposed biologically plausible rules.

Rules	Descriptions
i	Different velocity
ii	Different persistence time
iii	Directional movements
iv	Border effect

Table 5-3: Models and their corresponding rule combinations.

Model	Rules	Model	Rules
1	None	**9**	ii+iii
2	i	**10**	ii+iv
3	ii	**11**	iii+iv
4	iii	**12**	i+ii+iv
5	iv	**13**	i+ii+iv
6	i+ii	**14**	i+iii+iv
7	i+iii	**15**	ii+iii+iv
8	i+iv	**16**	i+ii+iii+iv

$$transition(S, C_p, C_n) \leftarrow speed(C_p, T, V)\ and\ direction(S, C_p, T, D)\ and\ move(C_p, V, D, C_n) \quad (5.8)$$

Following Equation (5.1)-(5.6) in Model construction, Equation (5.8) illustrates the function $transition$ (Appendix B), which is different for 16 models. Function $speed$ depends on whether rule i is selected or not. If rule i is selected, cells get different speed V based on their cell types T. More specifically, as stated in Table 5-1, T+ cells have a higher mean velocity as 100 $\mu m/h$, while T- cells have lower mean velocity as 40 $\mu m/s$. Function $direction$ based on the combination of the selection of rule ii, rule iii and rule iv with a specific order. We firstly check if rule ii is selected, and then rule iii and rule iv. If rule ii is not selected, then the cell can be updated with a new direction based on the following rule iii and rule iv. If rule ii is selected, then cells can only update their directions once if their persistence time is bigger or equal to the threshold. For getting a new direction, we firstly check rule iii. If rule iii is not selected, then each cell gets a random direction as their new direction; while rule iii is selected, each cell gets a new direction based on the neighbouring cells according to the equation (5.7) described in Section 5.3. Subsequently, we get the final direction based on rule iv. If rule iv is not selected, then the former direction we got is the final direction; however, if rule iv is selected, then a small angle change to obtain the final direction will be applied if the cell is close to the border according to the former description of rule iv.

Consequently, based on the previous rule that cells' direction may be affected by neighbouring cells and move with speed, we improved the selected best models by adjusting cell velocity by neighbouring cells, hence, cells do not always migrate with the same velocity. In addition to the previous directional movements, we calculate the ratio of actual velocity and maximum velocity based on the sum of forces divided by the sum of the magnitude of the forces. s_{max} is the maximum velocity of this type of cell. s is the actual velocity of cell migration.

$$s = s_{max} \times \frac{\sum F_{x,y}}{\sum |F_{x,y}|} \tag{5.9}$$

5.5 Assessing the model performance

To assess the performance of the 16 models we constructed, we firstly applied several existing metrics, including KL divergence, EMD, Bhattacharyya distance and CRPS (Section 4.4.1), to compare the model outputs against the experimental data. Since the evaluation results based on existing metrics do not follow our visual impression of the model outputs, we applied a novel metric, SCAPD (Section 4.4.3), to assess the model performance. The evaluation results based on existing metrics will be provided in Section 7.1 following the results based on SCAPD in Section 7.2. The motivation towards the novel metric was explained in Section 4.4.2 along with providing calculations on artificial data as an example.

5.5.1 Assessing the model performance with known metrics

- Assessing with KL divergence

To assess our model performance, one of the metrics we applied was KL divergence (described in Section 4.4.1.1). We firstly applied KDE on both aggregated cell locations in experimental data and model outputs with the same parameter settings. Aggregating means we analyse all cells from all images together. The number of mesh points used in the kernel density estimation was 256. Hence, we got the probability density over the 256×256 grid space. After the normalisation of probability density from both experimental data and model output, we calculate the sum of the KL divergence from experimental data to model output and the KL divergence from model output to experimental data. Figure 5-2 provides the pseudocode of the algorithm we applied.

Algorithm: Calculate the similarity based on KL divergence

```
% input data for T+ cells
Input Experimental_data, Model_data

% density estimation and normalisation
Experimental_density = apply kernel density estimation on Experimental_data
Experimental_density = apply normalisation on Experimental_density
Model_density = apply kernel density estimation on Model_data
Model_density = apply normalisation on Model _density

% calculate KL divergence
KL_divergence1 = KL divergence from Experimental_density to Model_density
KL_divergence2 = KL divergence from Model_density to Experimental_density

% output
Result_KL = KL_divergence1 + KL_divergence2
Output Result_KL
```

Figure 5-2: Pseudocode of the algorithm for calculating the similarity based on KL divergence.

- Assessing with EMD

We also applied EMD (described in Section 4.4.1.2) to evaluate our model outputs on an aggregated level. Two signatures were generated from experimental data and model output separately by taking cell 2D locations as points and all points have the same weight. The total weight of the points is 1. Subsequently, we calculated the EMD between the two signatures to get the evaluation results of 16 models based on EMD. We used the function from https://github.com/garydoranjr/pyemd to calculate EMD by feeding in sample data.

Algorithm: Calculate the similarity based on EMD

```
% input data for T+ cells
Input Experimental_data, Model_data

% calculate EMD
Result_EMD = calculate EMD between Experimental_data and Model_data

% output
Output Result_EMD
```

Figure 5-3: Pseudocode of the algorithm for calculating the similarity based on EMD.

- Assessing with CRPS

Instead of only analysing on an aggregated level, we applied CRPS to analyse the distribution of the density in each grid point. Before calculating CRPS between experimental data and each model output, we applied KDE on each image in experimental data and obtained a 2D matrix of the density distribution of each image. By aggregating all these density distributions, we obtained a 3D matrix with the third dimension stands for different images. Afterwards, we carried out the same approach on each model output and obtained a 3D matrix for results from each model. Based on these aggregated density distributions, we calculated CRPS between experimental data and each model output.

We took our experimental data, 186 images of cells on disc micropatterns and 152 images of cells on ellipse micropatterns, as ground truth for model evaluation. Since we are interested in the pattern formation of T+ cells, we applied the KDE of T+ cells in each cell colony (image). We omitted images with a too-small number of T+ cells (we take 20 cells as the threshold here). Similarly, we applied KDE with the same parameters for T+ cells locations of each run output from our 16 original models separately. Hence, we got the distribution density of each grid point (256×256 grid space here) for both

experimental data and model outputs. We calculated the CRPS of the distribution density of each grid point from experimental data to model outputs. Finally, we take the mean CRPS value of all CRPS value on each grid space as the final result of each model evaluation.

Algorithm: Calculate the similarity based on CRPS

```
% input data for T+ cells
Input Experimental_data, Model_data

% apply KDE on each image in both experimental data and model data
Experimental_data_kde_results = an empty 3D matrix
For each image in Experimental_data do
        Single_image_kde_result = Apply KDE to image data to obtain a 2D matrix
        Single_image_kde_result = Apply normalisation on Single_image_kde_result
        Append Single_image_kde_result to Experimental_data_kde_results on the
third dimension
End
Model_data = select images with over 20 cells from Model_data
Model_data_kde_results = an empty 3D matrix
For each image in Model_data do
        Single_image_kde_result = Apply KDE to image data to obtain a 2D matrix
        Single_image_kde_result = Apply normalisation on Single_image_kde_result
        Append Single_image_kde_result to Model_data_kde_results on the third
dimension
End

% calculate CRPS
% Experimental_data_kde_results and Model_data_kde_results have same size on the first and
second dimensions
For i  = 1 to the size of the first dimension of Experimental_data_kde_results
        For j  = 1 to the size of the second dimension of Experimental_data_kde_results
                Result_CRPS_all = calculating CRPS on the third dimension of
Experimental_data_kde_results and Model_data_kde_results on current i and j value
for first and second dimensions, and save this value to the matrix Result_CRPS_all
        End
End
Result_CRPS = the mean of Result_CRPS_all

% output
Output Result_CRPS
```

Figure 5-4: Pseudocode of the algorithm for calculating the similarity based on CRPS.

Algorithm: Assessing model performance based on SCAPD

% input data for T+ and T- cells
Input `Experimental_data_tpos`, `Experimental_data_tneg`, `Model_data_tpos`, `Model_data_tneg`

% density estimation and normalisation

`Experimental_density_tpos` = apply kernel density estimation on `Experimental_data`

`Experimental_density_tpos` = apply normalisation on `Experimental_density_tpos`

Apply same procedures on `Experimental_data_tneg`, `Model_data_tpos`, and `Model_data_tneg` to obtain `Experimental_density_tneg`, `Model_density_tpos`, and `Model_density_tneg`

% obtain borders for HDA by taking Experimental_density_tpos as an example
`Points_based_on_contour_plot` = extract the points marking the HDA from the contour plot of `Experimental_density_tpos` (2 levels, threshold is the mean of the max and min density value in the grids)
`Best_fit_circle_or_ellipse` = get the parameters of the best fit circle or ellipse from `Points_based_on_contour_plot`
`HDA_parameters_tpos` = fix the asymmetricity of the parameters of `Best_fit_circle_or_ellipse`
Apply same procedures on `Experimental_data_tneg` to obtain `HDA_parameters_tneg`

% calculate the total density in HDA for both experimental data and model data
`Total_density_experimental_tpos` = calculate the total density based on the `HDA_parameters_tpos` and `Experimental_density_tpos`

`Total_density_experimental_tneg` = calculate the total density based on the `HDA_parameters_tneg` and `Experimental_density_tneg`

Apply same procedures to calculate `Total_density_model_tpos` and `Total_density_model_tneg`

% calculate SCAPD
`Result_SCAPD` = absolute value of (`Total_density_experimental_tpos` – `Total_density_model_tpos`) + absolute value of (`Total_density_experimental_tneg` – `Total_density_model_tneg`)

% output
Output `Result_SCAPD`

Figure 5-5: Pseudocode of assessing model performance based on SCAPD

5.5.2 Assessing the model performance with SCAPD

Figure 5-5 shows the pseudocode of assessing model performance based on SCAPD. We used 256 as the number of mesh points for KDE again in both experimental data and model outputs. The borders of HDA generated from the experimental will be illustrated in Section 6.1.2 following with the results of total density within HDA for both T+ and T- cells on disc and ellipse micropatterns

in Section 6.1.3. Based on the borders of HDA from experimental data, we calculated the total density within the areas of T+ and T- cells in each model output. The detailed methodology of SCAPD was described in Section 4.4.3. The evaluation results based on SCAPD will be provided in Chapter 7, from basic models to best performance models.

5.6 Parameter optimisation through grid search

Based on the model with improved rules, we tested a set of parameters. Because the number of free parameters is not big, we applied the grid search for parameter optimisation to test all combinations of the values of these free parameters. The free parameters we tested are different sensing radius (R) for getting neighbouring cells and standard deviation (σ) for generating the final direction (as listed in Table 5-4). According to the results illustrated in former experiments of growing T+/T- cells in confined areas (Blin *et al.*, 2018), T+ and T- cells shows different preferential localisation by taking 50 μm and above as radius for neighbourhood on disc and ellipse micropatterns. Hence, we are interested to test 50 μm and above for sensing radius. Considering the size of the micropatterns, we tested 25, 50, 75, and 100 μm for sensing radius in our models. We remark that we used the same parameter values for the disc and the ellipse because we wanted to assess how the plausible underlying mechanisms might generalize across these different experimental settings.

Following parameter optimization, we tested applying different sensing radius (R) and different standard deviation (σ) for T+ and T- cells separately. Hence, T+ and T- cells can have different optimised value for these parameters. Again, we applied the grid search for parameter optimisation by testing all combinations of the potential values. We get the model with the best performance by applying the metric explained in Section 7.4.

Table 5-4: The list of possible values of the parameters used in a grid search setting.

Parameters	Values for disc	Values for ellipse
Sensing radius (R)	25, 50, 75, 100	25, 50, 75, 100
Standard deviation (σ)	1, 3, 5	1, 3, 5

5.7 Model realisation and user interface

Our models include an accessible user interface (an example is shown in Figure 5-6) with providing the visualisation of running the model. Blue background stands for the micropattern that cells can adhere to; green circles stand for T- cells and red circles stand for T+ cells. Models can be initialised by clicking the setup button and can be run by clicking the go button. The initial cell numbers can be set up by typing in the specific number in the boxes. The biologically plausible rules can be turned on through the switches. Moreover, the parameters can be adjusted through sliders.

The multiple runs of the model can be achieved through *BehaviorSpace*, which is a software tool integrated with Netlogo. It allows running a model many times with systematically varying the model's settings and recording the results of each model run. BehaviorSpace also allows us to run NetLogo on a cluster of computers. We achieved the grid search of parameters using BehaviorSpace. Due to the high computation required from running each model, we run each model 100 times by submitting jobs to clusters to obtain the probability output for model evaluation and optimisation.

In this chapter, we described the model construction in detailed, including the assumptions of modelling and the cell motility rules we tested in this study. We explained the application of existing or novel metrics for model evaluation as well as the method of model optimisation. The evaluation results and optimisation results are provided in Chapter 7.

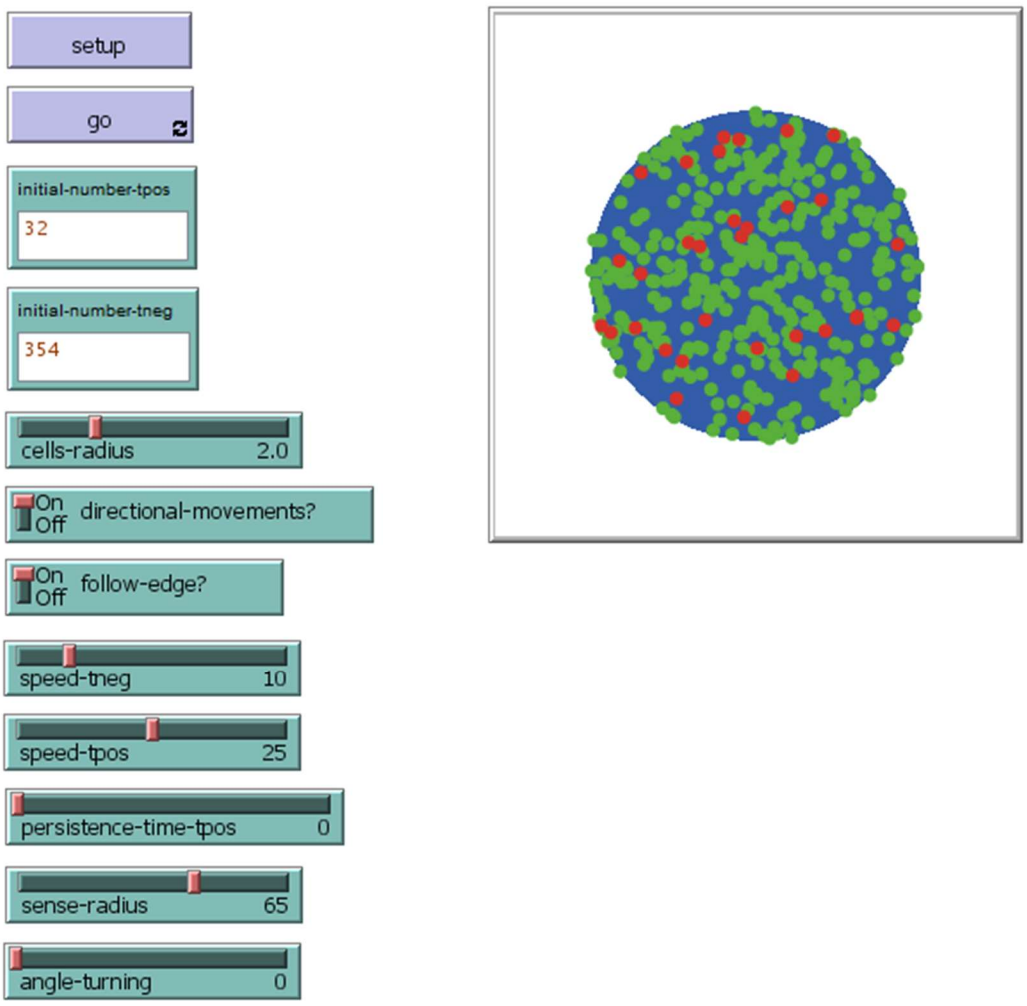


Figure 5-6: An example of the user interface of our models

Chapter 6

6 Analysis results from experimental data

In the former chapters, we described the experimental data in Chapter 3, following by the explanation of the methodology (Chapter 4) and illustrated model construction (Chapter 5). In this study, we take the experimental data as ground truth for our modelling. Intending to gain more insights into the experimental data, as well as looking for a good way to describe the ground truth, we carried out a series of analysis and measurements (described in Chapter 4) on the experimental data.

In this chapter, we provide the analysis results from the experimental data. In the first part of this chapter, we describe the pattern formation we observed on an aggregated level along with defining high-density areas (HDA). We provide the results of the total density within HDA and also the results showing the variety in experimental data. In the second part, we measure the proximity of T+/T- cells and provide the results proving different proximity preference in T+ and T- cells.

6.1 Pattern observation

6.1.1 Pattern observation on aggregated images

For visualising the pattern in ESCs on an aggregated level, we generated the density maps of T+ and T- cells on disc and ellipse micropatterns separately. We applied 2D kernel density estimation (Botev, Grotowski and Kroese, 2010) with Gaussian kernels to aggregate cell colonies for disc and ellipse micropatterns. Specifically, we applied Botev's approach for density estimation,

which increased accuracy and reliability. The density maps are shown in Figure 6-1. Figure 6-2 illustrates the contour plots of the density maps of binned cell cultures for disc and ellipse micropatterns for T- and T+ cells. In colonies that are geometrically constrained within a disc, T- cells preferentially localise on the centre of the disc, whilst T+ cells preferentially localise on the edge of the disc. Interestingly, in colonies that are geometrically constrained within an ellipse, while T- cells preferentially localise on the centre of the ellipse as well, T+ cells preferentially localise on the tips of the major axis of the ellipse instead of localising on the whole border of the ellipse.

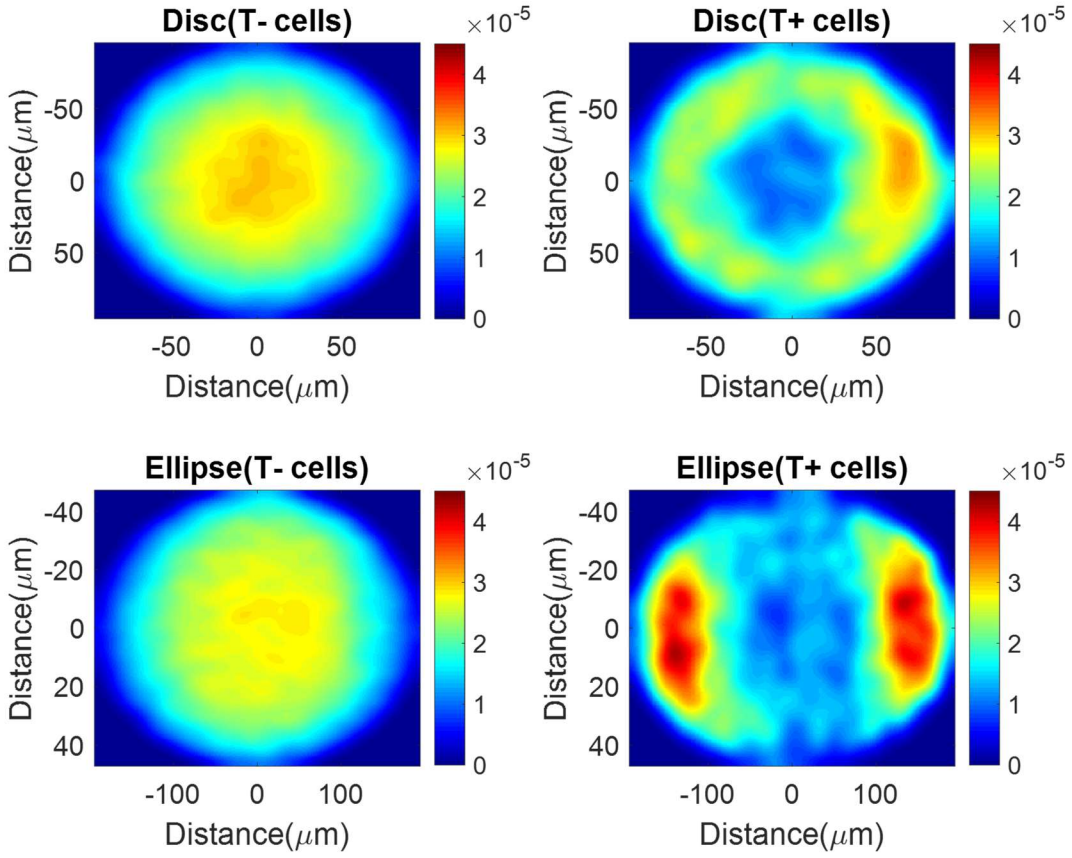


Figure 6-1: Density maps of T- and T+ cells on disc and ellipse micropatterns from experimental data.

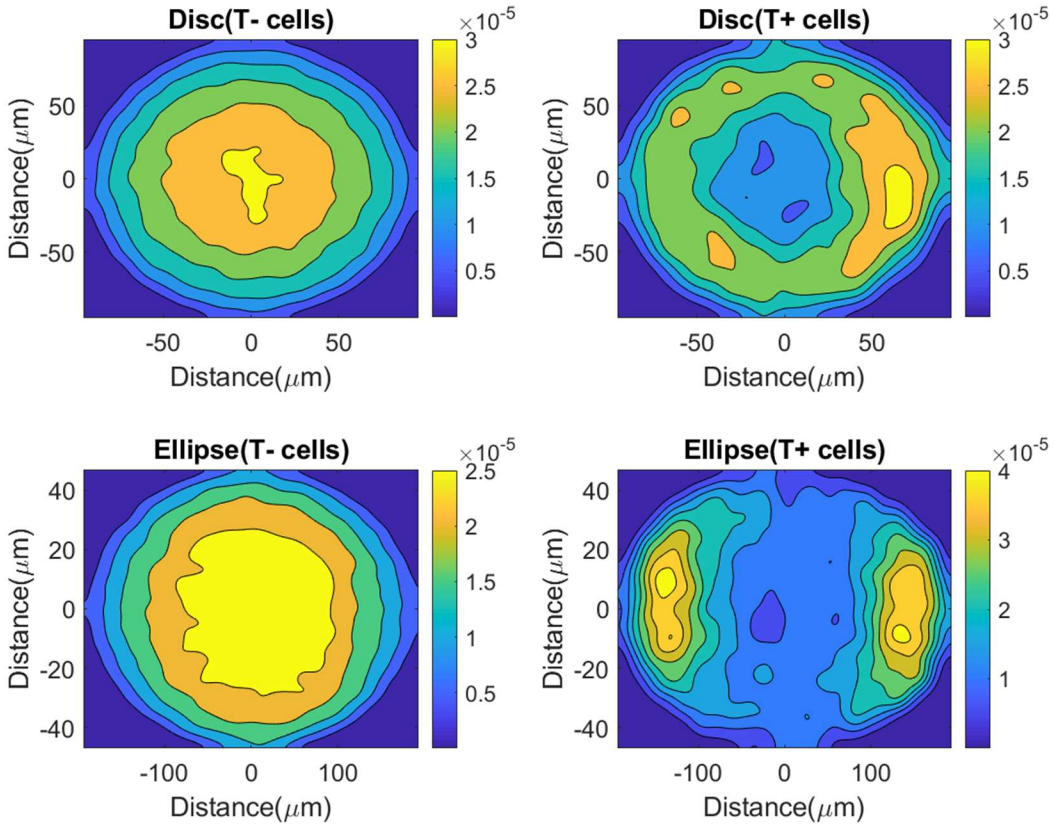


Figure 6-2: Contour plots of density maps of T- and T+ cells on disc and ellipse micropatterns from empirical data

The finding from our experimental data, that T+ cells prefer to stay at the border of disc-shaped micropatterns, are consistent with the previous studies that worked on the relationship between geometric confinement and pattern formation as described in Section 2.1.2 (even there are differences of cell types and biomarkers in these studies). Compared to the existing findings as described before, our experimental data includes the patterns from ellipse-shaped micropatterns. Different from disc-shaped micropatterns, which are symmetric according to any axis, the patterns observed on ellipse-shaped micropatterns might provide more insights into the symmetric breaking behaviours of cells.

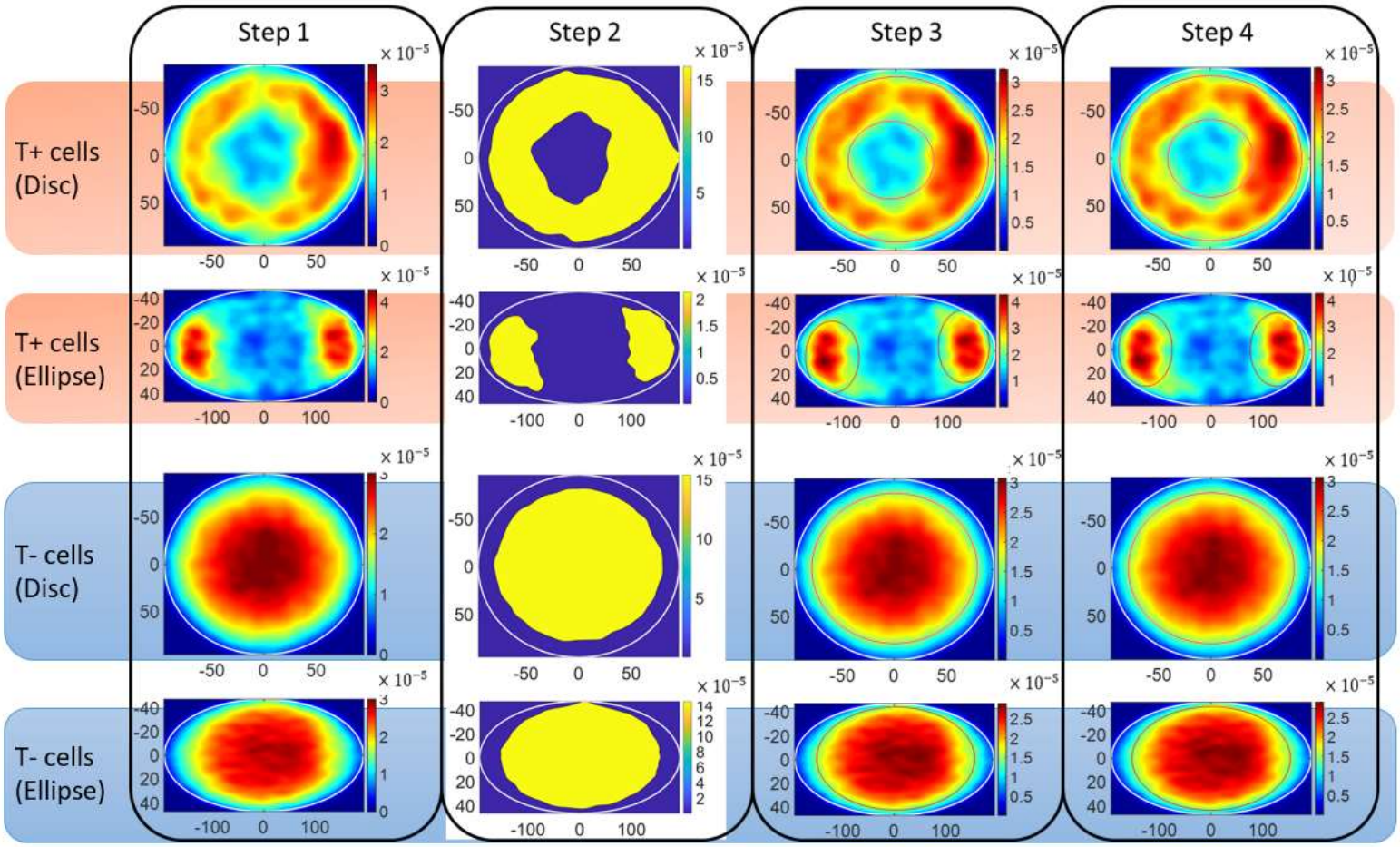


Figure 6-3: The process of getting the borders of HDA for T+ and T- cells separately for disc and ellipse micropatterns. White lines stand for the border of the micropatterns; red lines show the border used for evaluation we found in step 3 and step 4.

6.1.2 High-density area

For both disc and ellipse experiments, we obtained the borders of high-density areas (HDA) for describing the pattern and evaluation by applying SCAPD (Section 5.5.2). Figure 6-3 illustrates the process of getting the borders of HDA. Pseudocode was provided in Figure 5-5. The steps of getting the borders are:

- **Step 1** Computing density maps: We applied 2D KDE with Gaussian kernels to generate aggregate density maps of T+ and T- cell colonies separately. We applied Botev's approach for bandwidth selection. Because it improved local adaptivity and reduced boundary bias, it increased the accuracy and reliability by reducing the sensitivity to outliers, asymptotic bias and mean square error. We calculated the density over a 256×256 grid to obtain smooth results hence the results contain 256×256 pixels.

- **Step 2** Computing points by thresholding: Based on empirical data, we collected a list of points marking the border of the area by thresholding. The threshold t was calculated as the mean of maximum and minimum value of density of each grid point (g_i), as shown below.

$$t = \frac{\max(g_i) + \min(g_i)}{2} \tag{6.1}$$

- **Step 3** Best fit circles/ellipses: Based on the points we marked, we generated the best fit circles or ellipses based on least-squares fitting (Hastie, Tibshirani and Friedman, 2009).

- **Step 4** Fixing asymmetric: Because of the symmetrical properties of disc and ellipse, we believe that the observed pattern on disc micropattern should be symmetric according to any arbitrary angle, and the pattern on ellipse micropattern should be symmetric according to the x and y-axis. For disc micropattern, we keep the radius of the best-fit circles and moved the centres to the centre of the disc micropattern (point(0,0)). For ellipse micropatterns, we use the maximum value for

both semi-minor and semi-major axis of the best fit ellipses, and then we keep the absolute value of x of the centre of the best-fit ellipse as the mean of the absolute values of two x values, and y of the centre of the best-fit ellipses as 0.

6.1.3 Total density within HDA from experimental data

The explanation of our novel evaluation metric SCAPD was provided in Section 4.4.3. The borders we used in calculating SCAPD is the borders of HDA as described in the former Section 6.1.2 (as shown in Figure 6-3). We calculated the total density within the areas for T+ and T- cells in the empirical data on both disc and ellipse micropatterns. The results of total density from the empirical data are shown in Table 6-1.

Table 6-1: Total density within HDA from experimental data.

	T- cells	T+ cells
Disc	0.8233	0.7407
Ellipse	0.8471	0.5064

6.1.4 Pattern grouping

We compared the ratio of HDA/non-HDA size to the ratio of HDA/non-HDA cell number to get an indication of cell spacing. By comparing the ratio of size and the ratio of cell number of different areas, we can be informed whether the cell density is higher on HDA or not. Subsequently, we obtained three distinct groups of different cell patterns based on the grouping of T+ cells: 1) cell colonies with a relatively higher density of T+ cells on HDA; 2) cell colonies with a relatively low density of T+ cells on HDA; 3) cell colonies with no T+ cells.

Figure 6-4 shows the percentage of different patterns observed in the disc and ellipse micropatterns. We defined pattern 1 as a relatively higher density of T+ cells within the HDA. In pattern 2, the density of T+ cells within the HDA was lower than the density of T+ cells within the non-HDA. On pattern 3, there were no T+ cells.

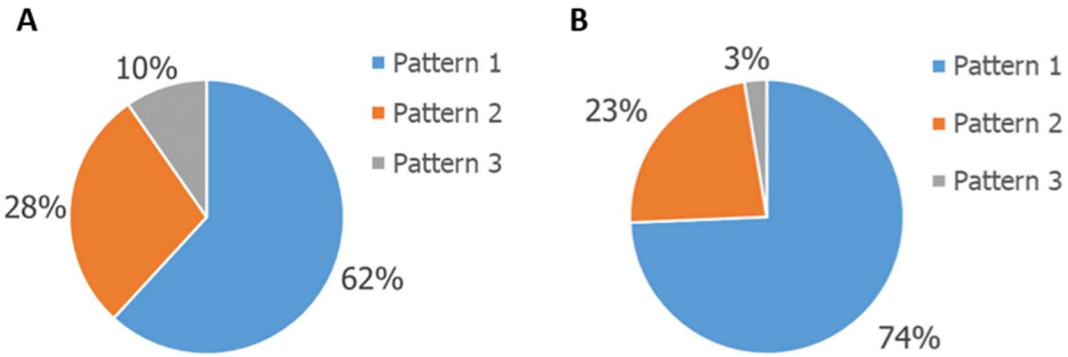


Figure 6-4: The percentage of the 3 different patterns observed within the A) disc and B) ellipse micropatterns. Pattern 1: high density of T+ cells within the HDA. Pattern 2: density of T+ cells within the HDA was lower than the density of T+ cells within the non-HDA. Pattern 3: no T+ cells.

6.1.5 Investigating the variation in empirical data

Even though we observed the pattern by aggregating all images, it is noteworthy that there is a high variety in our empirical data. Figure 6-4 shows that over 20% of colonies on disc and ellipse micropatterns have a higher density of T+ cells on non-HDA. In the next step, we calculated the ratio of T+ cells on HDA through equation (6.2). r is the ratio of T+ cells on HDA. n_1 is the number of T+ cells on HDA, and n_{all} is the number of T+ cells on this colony.

$$r = \frac{n_1}{n_{all}} \tag{6.2}$$

Figure 6-5 shows the distribution of the ratio of T+ cells on HDA on disc and ellipse micropatterns (excluded the images with 0 T+ cells). It is noteworthy that there is a high variety among the empirical data as the ratio of T+ cells on HDA spreads out from 0 to 1 for both disc and ellipse micropatterns. The ratio of T+ cells on HDA of the colonies on disc micropatterns is higher than the ratio of ellipse micropatterns.

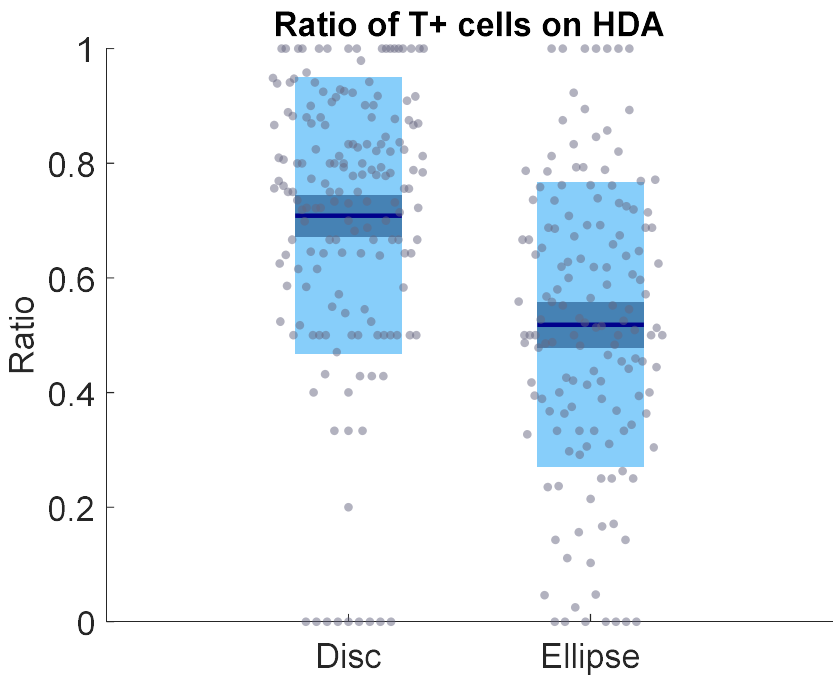


Figure 6-5: The plot of the ratio of T+ cells on HDA on disc and ellipse micropatterns. Scattered points on the plot represent the raw data. The dark blue lines stand for the mean of the grouped data, light blue shows 95% confidence interval, 1 standard deviation is also shown within grey-blue. Images were generated by Matlab function notBoxPlot.

6.2 Proximity measurements

In this study, we investigated the proximity between different types of cells by carrying out different measurements. The results from proximity measurements provided us with clues for proposing biologically plausible rules of modelling. We projected 3D cell colonies to the 2D surface (more discussion in Section 3.2). For getting a sense of how close cells are to each other, we calculated the percentage of cells with at least one neighbouring cell within a certain radius (as shown in Figure 6-6). The average diameter of mESC is about 15 μm (Zhou *et al.*, 2016). The cell size various depends on the age of the cell and the different differentiation stage the cell on. However, the Euclidean distance between two cells on 2D space can be much smaller than the average size of ESC. This is due to we projected the cells from 3D space to 2D space.

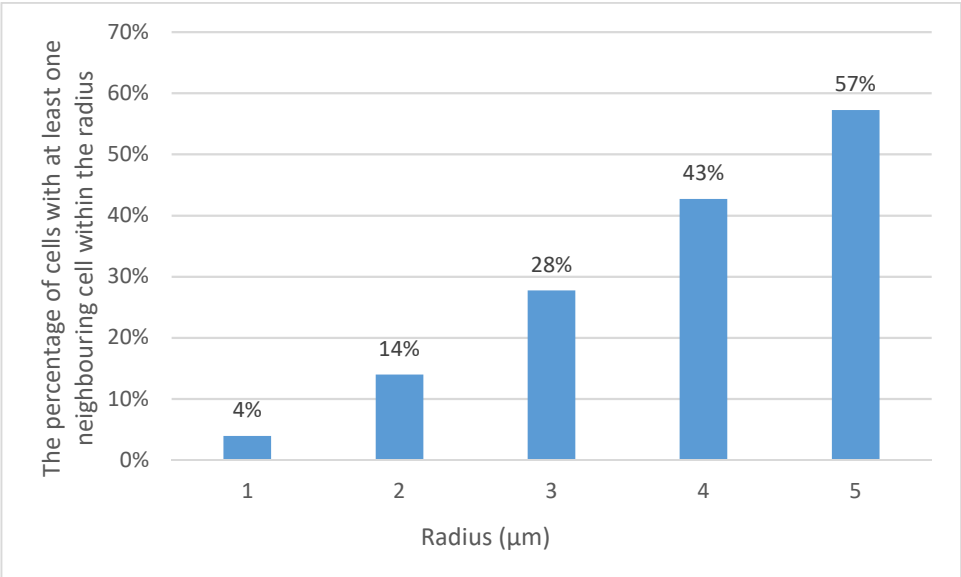


Figure 6-6: The percentage of cells with at least one neighbouring cell within a certain radius.

To measure the proximity (closeness) preference between different types of cells, we applied two measurements (minimum spanning tree and the average distance for each query object to five nearest targets) to the experimental data. The explanations of the measurements were provided in Section 4.3.

6.2.1 Results of applying minimum spanning tree to quantify the cell proximity

We first applied the minimum spanning tree of all T+ and T- cells separately. The results of the overall distribution of average path distance based on the minimum spanning tree are shown in Figure 6-7. The overall aggregate results for the disc and the ellipse were consistent. It is noteworthy that T- cells have a much smaller average path distance than T+ cells. The results indicate that T- cells have closer proximity than T+ cells, and the variation in the proximity of T+ cells was greater than in T- cells.

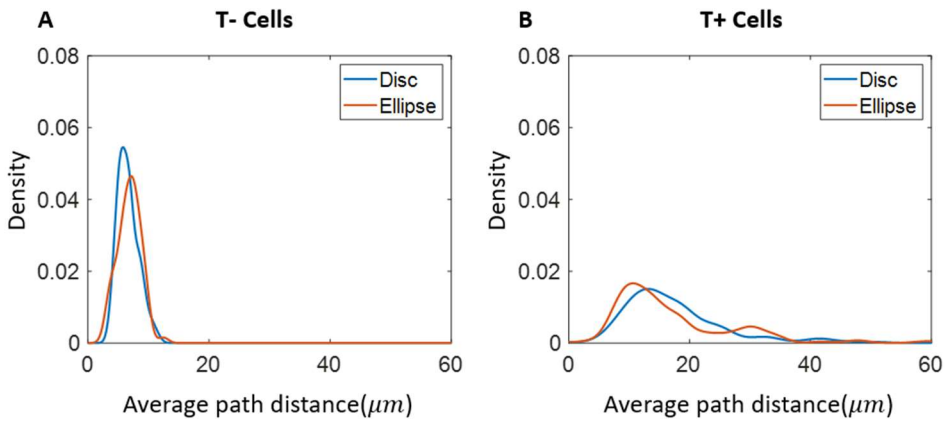


Figure 6-7: Kernel density estimation of average path distance (μm) of A) T- cells and B) T+ cells within disc and ellipse micropatterns.

Subsequently, we calculated the minimum spanning tree of T+/T- cells on different pattern groups. The pattern groups were described in Section 6.1.4.

We applied KDE to the results separately (Figure 6-8). The difference found in T- cells between the pattern groups were relatively small. For Pattern 3, in which there are no T+ cells, T- cells have the highest proximity. T+ cells in Pattern 1 (more T+ cells on HDA) are slightly more compact than T+ cells in Pattern 2. Again, the results from disc and ellipse micropatterns are consistent. The double peaks of T+ cells on ellipse micropatterns might due to the fact that in some colonies T+ cells were denser at one tip of the ellipse and in some colonies T+ cells were denser at both tips.

The results of the average path distance based on the minimum spanning tree indicate that T- cells prefer closer proximity than T+ cells. The T+ cells in colonies that formed the pattern with more cells in the HDA (Pattern 1) have higher proximity than the T+ cells in the other pattern groups.

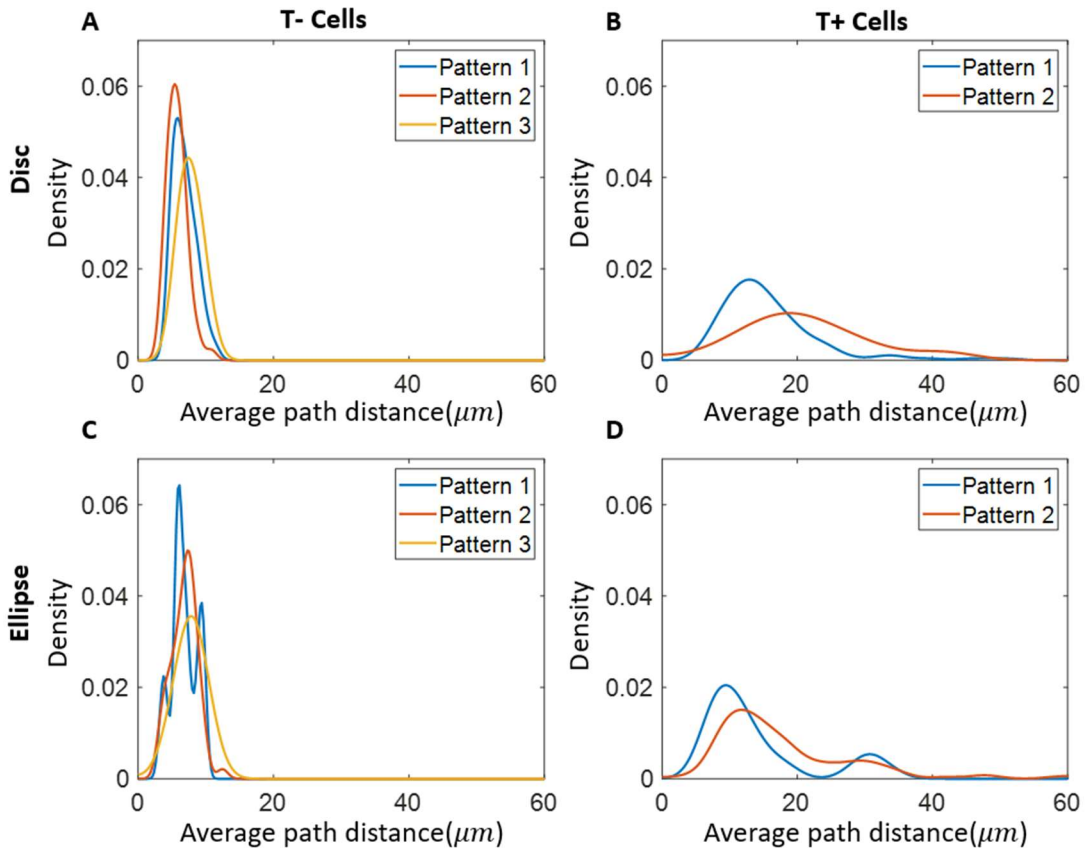


Figure 6-8: Kernel density smoothing of average path distance of A) T- cells on disc micropatterns; B) T+ cells on disc micropatterns; C) T- cells on ellipse micropatterns; D) T+ cells on ellipse micropatterns.

6.2.2 Results of quantifying average distance for each query object to five nearest targets

The measurement was described in Section 4.3.2. Figure 6-9 shows the kernel density estimation results of the average distance from the object to the five closest targets (D) from the aggregated cells on disc and ellipse micropatterns. The case of T- cells as the objects and T- cells as the targets results in the highest peak density in both disc and ellipse micropatterns. This is consistent with the results from the minimum spanning tree which showed that T- cells have high proximity. For the case in which T- cells are taken as objects and T+ cells as targets the distribution of D has the high mean value and a wide spread. This result could be due to the fact that there are more T- cells than T+ cells in the colonies. Hence, for T- cells it takes a longer distance to retrieve for the closest T+ cells on average. For T+ cells, the distribution of D to its five closest T+ or T- cells are relatively similar.

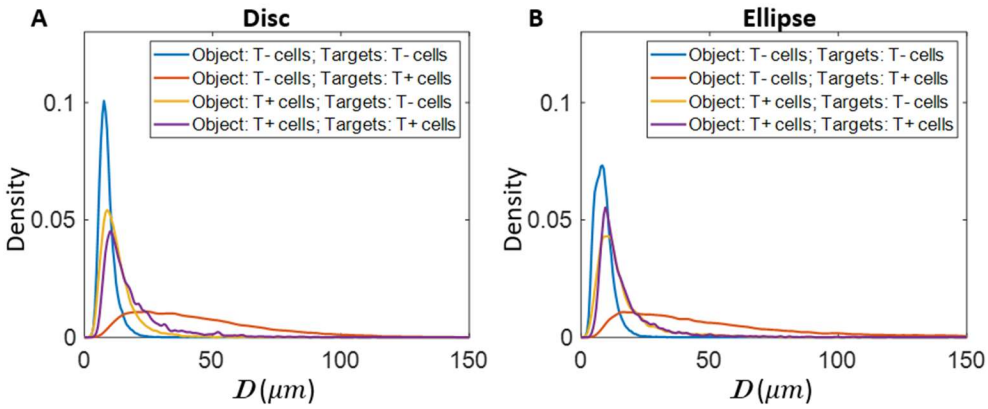


Figure 6-9: Kernel density smoothing results of D from the aggregated cells on A) disc and B) ellipse micropatterns.

With this new measurement, we also get insights into different proximities of cells in the HDA and outside the HDA. The steps of generating the borders of HDA will be provided in Section 4.4.3, and the results of the border of HDA will be provided in Section 6.1.2. Based on the borders of the HDA, we applied

kernel density estimation of the cells in the HDA and cells outside the HDA separately to investigate the difference in proximity between the different cell types in both regions.

Figure 6-10 shows the kernel density estimation results of D which is grouped by cell locations. It shows for all object T- cells the average D of the five closest T- cells showed relatively low variability whether contained within the HDA or not. For T- cells not in the HDA, the average distance to the five closest T+ cells is higher than for T- cells in the HDA. Interestingly, for the cases in which T+ cells are taken as objects, and either T+ or T- cells as targets, the T+ cells in the HDA have a higher average D than T+ cells not in the HDA and the distribution of D is slightly more spread out.

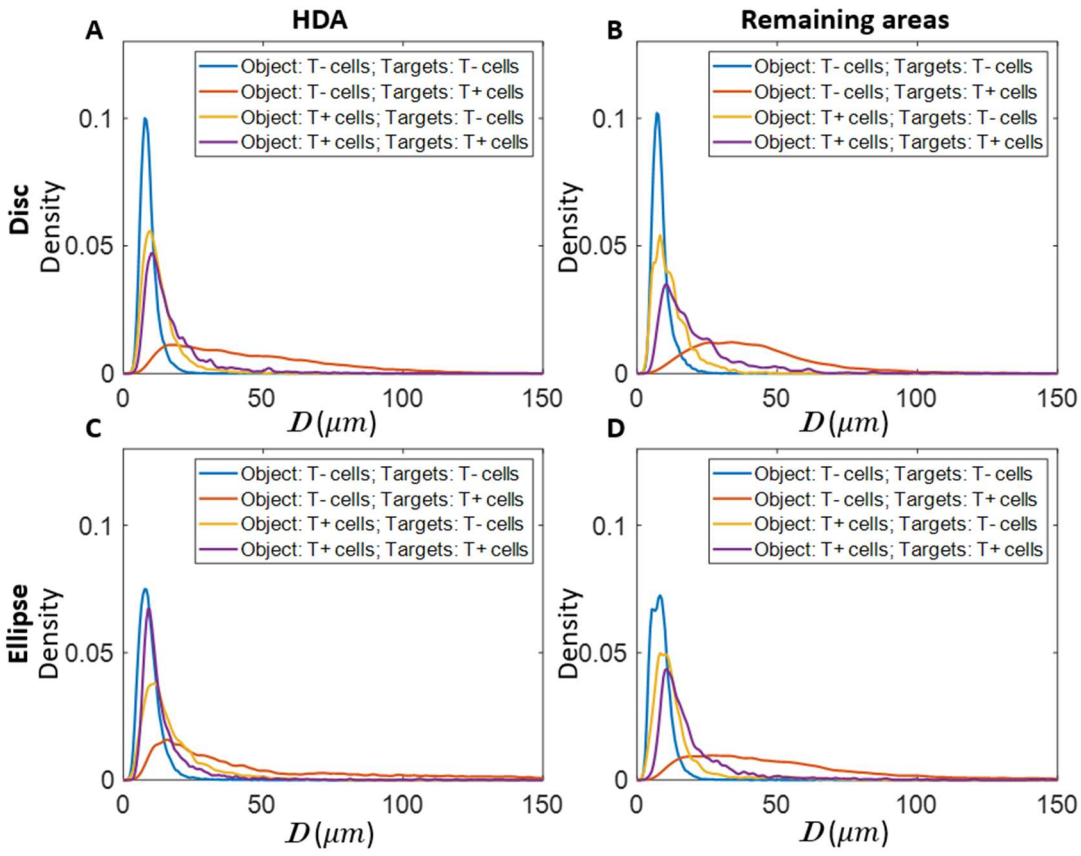


Figure 6-10: Kernel density smoothing results of D from the aggregated A) object cells on HDA on disc; B) object cells on non-HDA on disc; C) object cells on HDA on ellipse; D) object cells on non-HDA on ellipse.

The results from D are consistent with the results from the minimum spanning tree that T- cells have closer proximity than T+ cells. It also suggests that T+ cells tend to stay away from other cells as they have relatively low proximity to both T- and T+ cells, compared to the high proximity we observed in T- cells.

In this chapter, we first described the observed patterns in our experimental data. The high-density areas based on aggregated experimental data were defined, which will be used for calculating SCAPD for model evaluation. Secondly, we carried out two measurements for quantifying different proximity between different types of cells. These results also support the cell motility rules we test in this study.

Chapter 7

7 Simulation results

In Chapter 5, we described the methodologies we used for model construction, evaluation and optimisation. In Chapter 6, we illustrated the analysis results from experimental data (ground truth) and also described the desired pattern for reproducing.

In this chapter, we visualise the results from our models and also demonstrate the evaluation results of our models by applying both known metrics and the novel metric SCAPD. We deliver the results of the sensitivity test of a specific parameter and explained the parameter selection for optimisation. We provide the evaluation results for a series of models, from basic models to our best performance models.

7.1 Evaluation results based on existing metrics

Before we carried out our novel evaluation metric SCAPD, we tested multiple related existing approaches (explained in Section 4.4.1) to evaluate pattern formation in disc and ellipse models against the empirical data. We considered Kullback-Leibler divergence (KL divergence) (Cover and Thomas, 2006), earth mover's distance (EMD) (Rubner, Tomasi and Guibas, 2000), Bhattacharyya distance (Kailath, 1967) and continuous rank probability score (CRPS) (Gneiting and Raftery, 2007) metrics.

- Evaluation results based on KL divergence

As described in Section 5.5.1, we calculated the sum of the KL divergence from experimental data to model output (Table 5-3) and the KL divergence

from model output to experimental data. Figure 7-1 shows the results of 16 models based on KL divergence. Based on the results from KL divergence, the random model (Model 1) is very competitive and almost is the best model. However, we do not observe anything close to our desired pattern in random model outputs (more details will be provided later in this section).

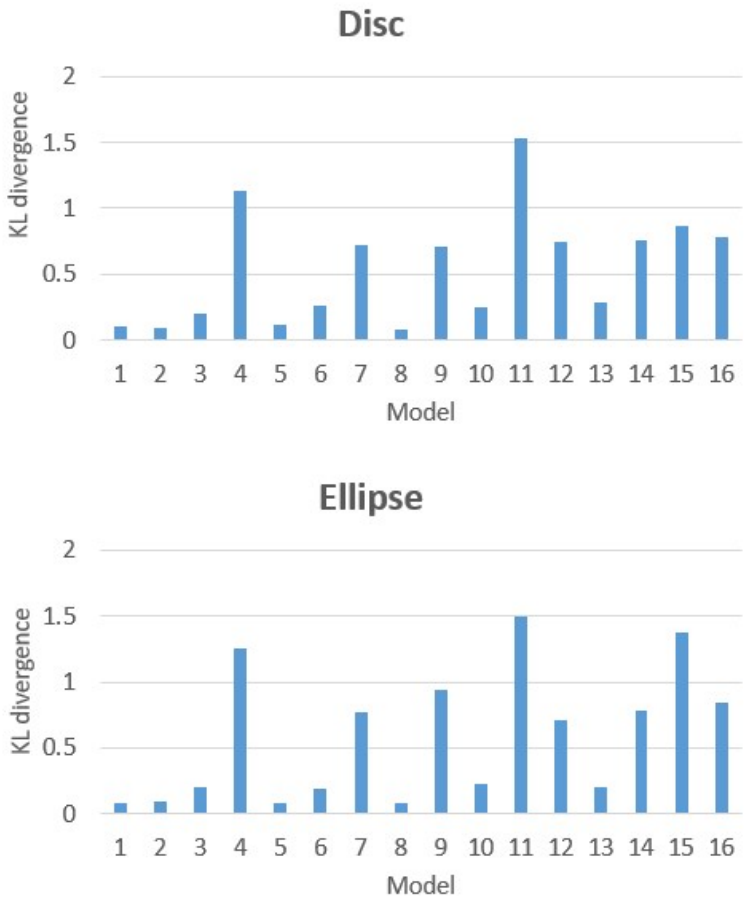


Figure 7-1: KL divergence results of model outputs compared to experimental data.

- Evaluation results based on EMD

We calculated EMD based on aggregating all experimental images and model outputs separately (Section 5.5.1). The results of EMD is shown in Figure 7-2.

Similar to the results from KL divergence, the random model (Model 1) is very competitive based on the results from EMD.

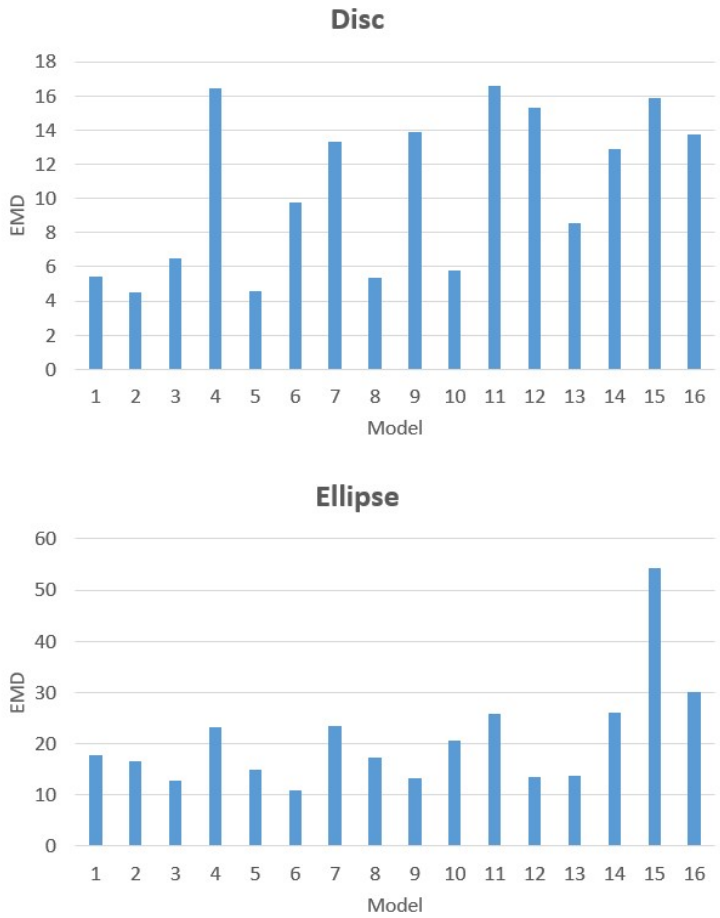


Figure 7-2: EMD results of model outputs compared to experimental data.

- Evaluation results based on Bhattacharyya distance

Subsequently, we calculate the Bhattacharyya distance between the probability distributions from our experimental data and model outputs (as described in Section 5.5.1). The results again suggest that the random model is a competitive model, which is not consistent with our observation (more details will be provided later in this section).

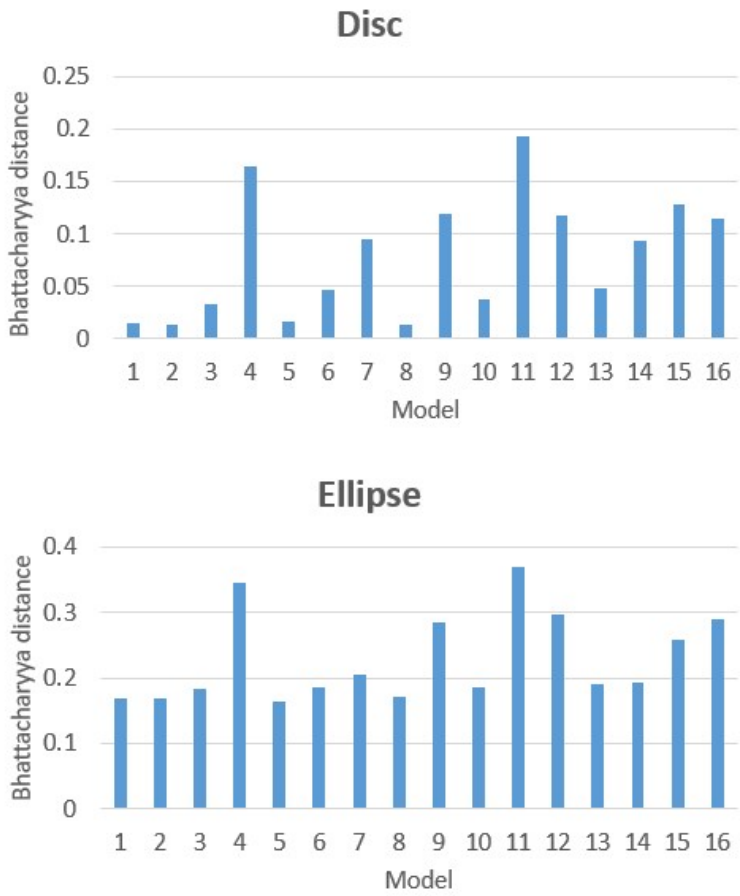


Figure 7-3: Bhattacharyya distance results of model outputs compared to experimental data.

- Evaluation results based on CRPS

Figure 7-4 shows the results of calculating the CRPS of 16 models. Different from KL divergence and EMD that compare two distribution on an aggregated level, CRPS takes the distribution of the density in each grid point into account. However, the results from CRPS does not show the different performance between models as well.

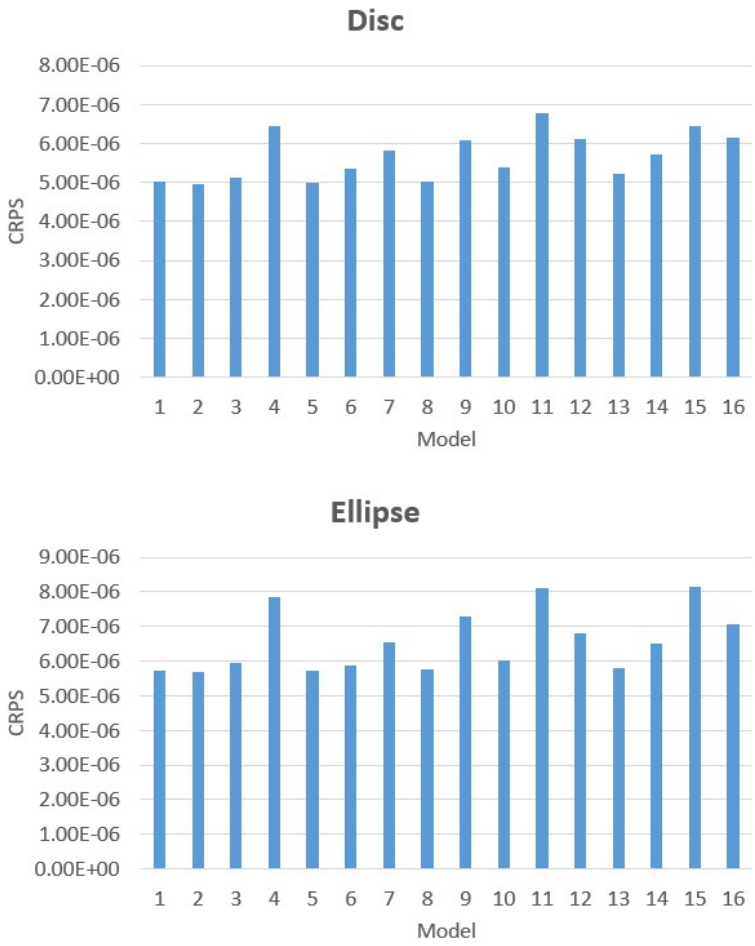


Figure 7-4: CRPS results of model outputs compared to experimental data.

To illustrate the motivation of proposing a novel evaluation metric in more details, we take the output from Model 1 and Model 7 as examples to show the limitation of the existing metrics. Model 1 is the random model and model 7 is the model that have different speed and directional movements for T+ and T- cells. Figure 7-5 shows the density plots of T+ cells distribution after running models 100 times. The results from Model 7 is closer to our desired pattern as T+ cells prefer to localise at the border on the disc. Table 7-1 shows the results of KL divergence sum, EMD, Bhattacharyya distance and CRPS. Hence, for all these established approaches, the resulting outcome did not follow the visual impression we had from observing the resulting patterns, which

motivated the development of a new approach (SCAPD) that was tailored specifically to this application.

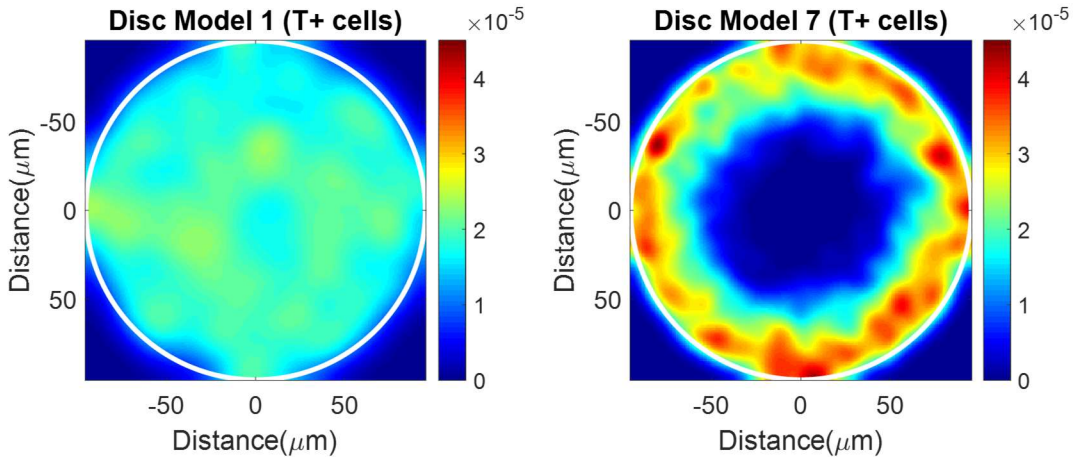


Figure 7-5: Density plot showing the aggregated results from A) model 1 and B) model 7 on disc micropatterns for running models 100 times.

Table 7-1: The evaluation results of Model 1 and Model 7 based on existing metrics.

	Model 1	Model 7
KL divergence sum	0.1011	0.4695
EMD	5.4559	13.306
Bhattacharyya distance	0.0145	0.0948
CRPS	5.00e-6	5.77e-6

7.2 Basic models

In Figure 7-6, we show the SCAPD results from 16 models for both disc and ellipse micropatterns. As shown in Figure 7-6, for both disc and ellipse micropatterns, Model 7 (with different velocity and directional movements) and Model 14 (with different velocity, directional movements, and border effect) have the best performance. For disc models, model 14 is slightly better than

Model 7, while for ellipse models, Model 7 is slightly better than Model 14. However, the difference between Model 7 and Model 14 is minimal. So in the next step, we focus on Model 7 and Model 14.

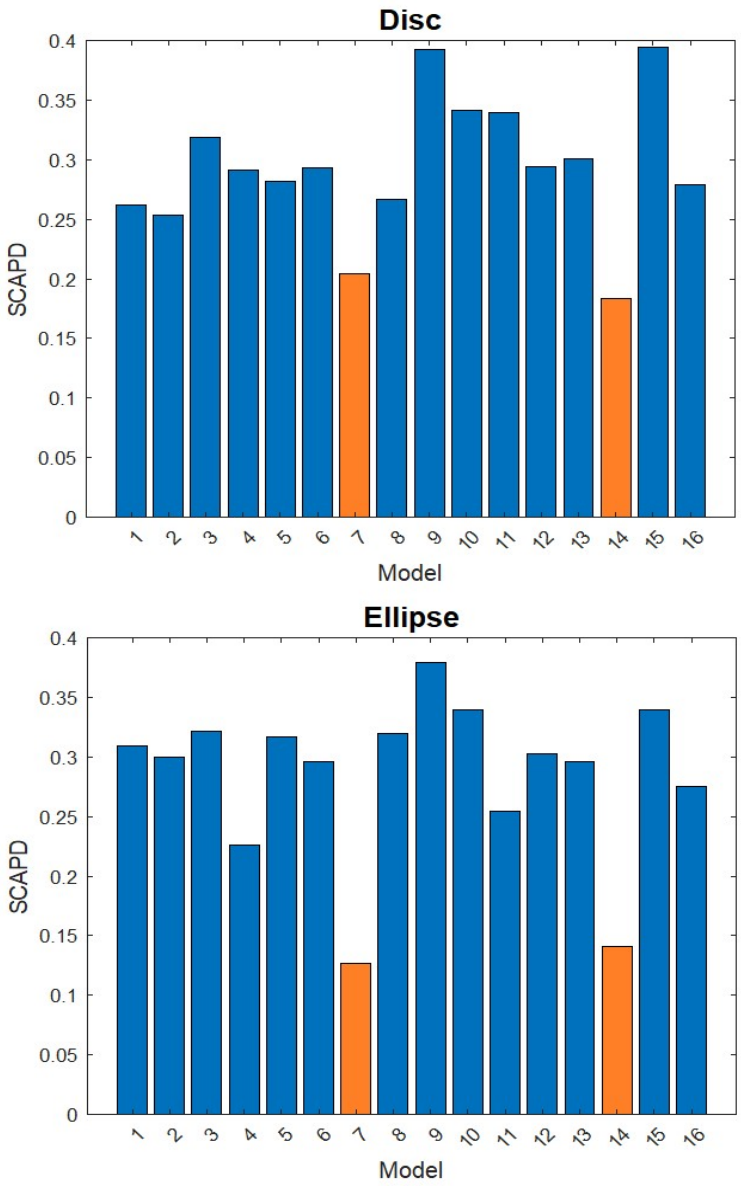


Figure 7-6: SCAPD results from 16 models for disc and ellipse micropatterns.

We tested the sensitivity of model output with respect to changes in the parameter of angle change in Model 14. As the results are shown in Table 7-2, the model output, based on the results of SCAPD, is not sensitive to the small degree of angle change. In addition, the velocity of T+ and T- cells were measured experimentally before, as described in Section 5.3. Hence, we do not include angle change value and cell velocity as one of the parameters to apply the grid search for parameter optimisation.

Table 7-2: SCAPD results of Model 14 for disc and ellipse micropatterns with different angle change values, demonstrating the model is very robust to the choice of this parameter.

Angle change value (°)	Disc	Ellipse
10	0.3591	0.6593
20	0.3992	0.6593
30	0.3546	0.6593
40	0.3640	0.6593

7.3 Models with parameter optimisation

Based on the original Model 7 and Model 14, we improved the rules by getting the ratio of velocity according to Equation (5.9). Afterwards, we tested a set of parameters combination of sensing radius (R) and standard deviation (σ) (the possible values we tested are listed in Table 5-4) and got the SCAPD results as shown in Figure 7-7. For disc models, Model 14 is slightly better than Model 7, while for ellipse models, Model 7 is slightly better than Model 14. However, the difference is extremely small (0.002213 for disc models and 0.0006 for ellipse models). The results of parameter optimisation from Model 7 and Model 14 are quite consistent from both disc and ellipse micropatterns (with slightly different optimised value for σ on ellipse micropatterns). The optimised values of R and σ show that cells might have quite a wide range of sensing

neighbouring cells to decide their directions and the randomness of their movements is quite high.

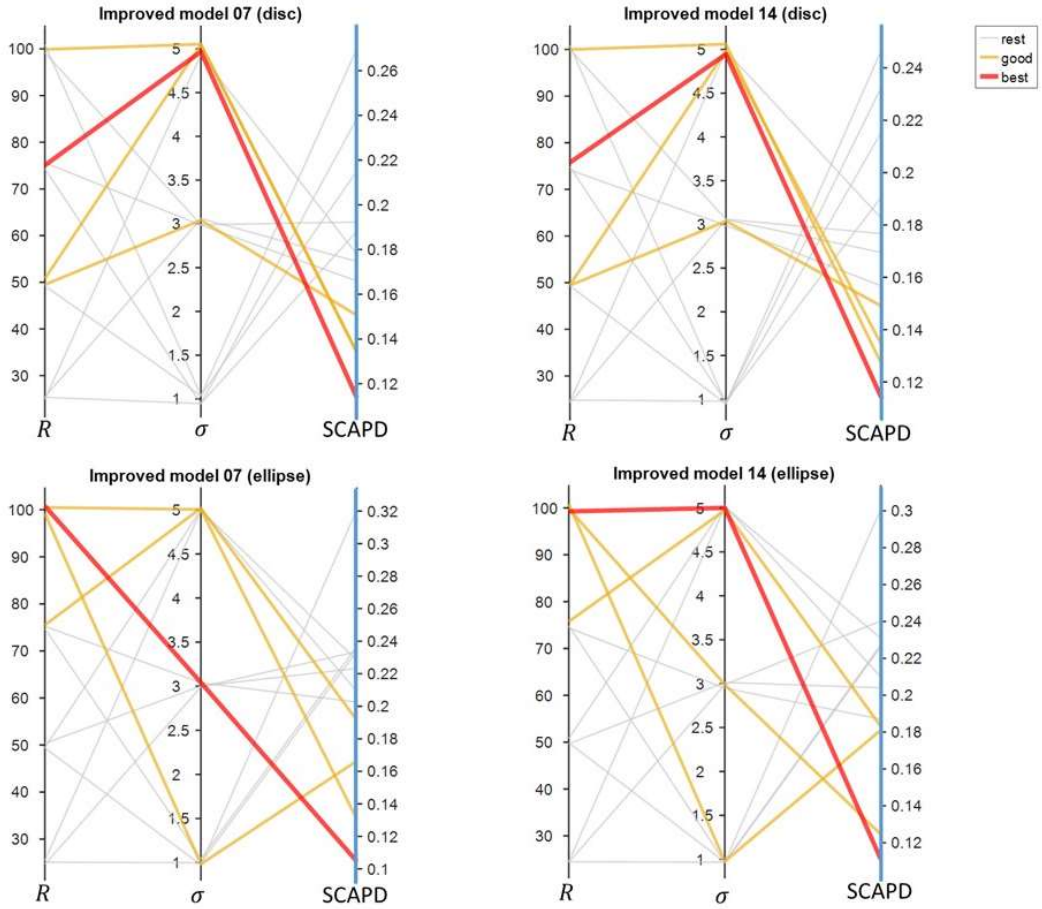


Figure 7-7: SCAPD results from models with grid search for parameter optimization. Black vertical lines stand for parameters we tested, sensing radius (R) and standard deviation (σ). The last blue vertical lines show SCAPD results. Each line crossing three vertical lines stand for each mode with specific parameters setting and quantified SCAPD. The red line is the best performing model in this group; orange lines are the next three best-performing models; grey lines stand for the remaining models.

Overall, by optimising the ratio of velocity and the optimization of parameters R and σ through grid search, we reduced SCAPD by about 38% compared to

the original best model for disc experiments, and by about 27% compared to the original best model for ellipse experiments.

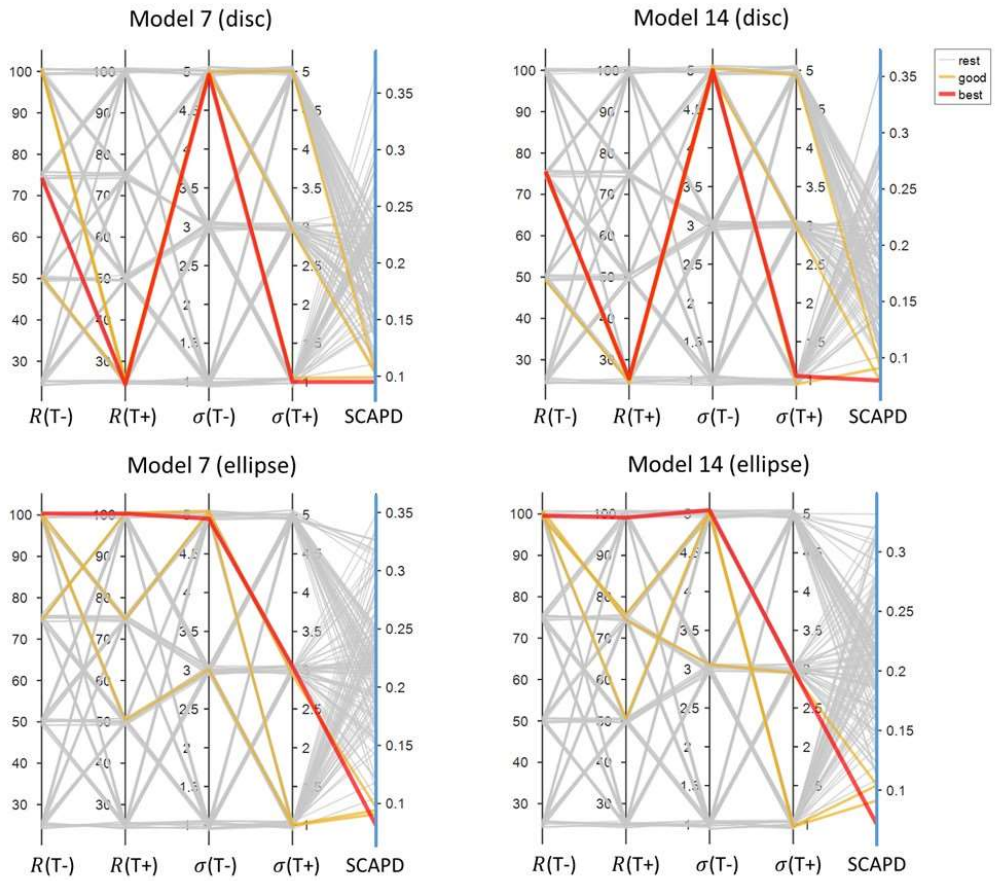


Figure 7-8: SCAPD results from models with different parameter settings for T+ and T- cells. The structure of the plots is similar to Figure 7-7. The first four vertical lines stand for four parameters (different sensing radius (R) and different standard deviation (σ) for T+ and T- cells).

7.4 Best performance models

Subsequently, we tested different sensing radius (R) and different standard deviation (σ) for T+ and T- cells separately to bring the models closer to the empirical data. Figure 7-8 shows the results of all combinations of the parameters. As we can see in Figure 7-8, sensing radius (R) play an important

role to models results, while models are not so sensitive to the standard deviation (σ). Again, we see consistent optimisation results from Model 7 and Model 14 on both disc and ellipse micropatterns. However, even though the sensing radius should be an intrinsic cell property, we have different optimised values for disc and ellipse micropatterns. This difference might cause by (1) the natural properties of the different shapes of micropatterns since part of the circle that we considered as the neighbourhood of cells would be empty due to the part of the circle would be outside of the micropatterns; (2) different crowdedness between different shaped micropatterns. Hence, it is reasonable that disc and ellipse micropatterns have different values for these parameters.

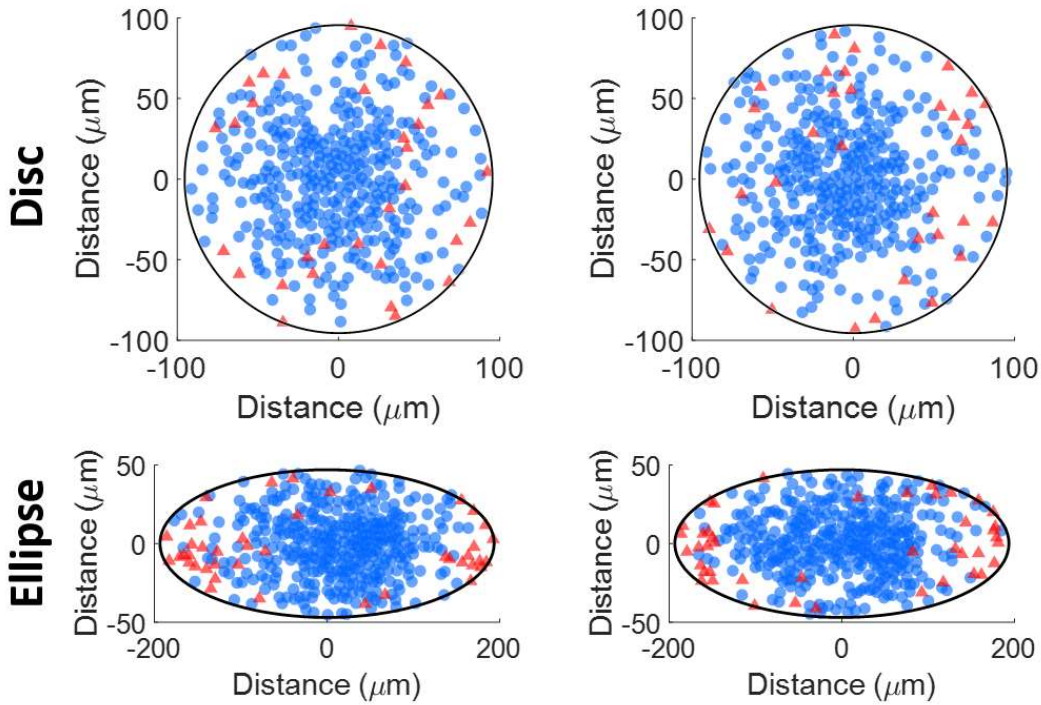


Figure 7-9: Examples of model outputs from Model 7 for disc and ellipse experiments.

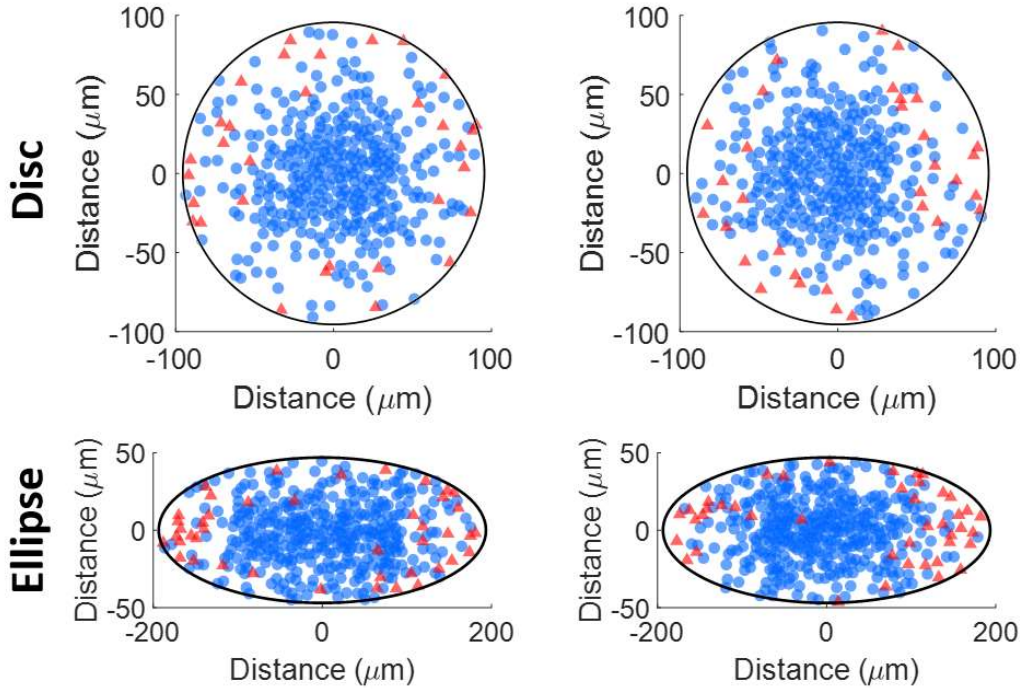


Figure 7-10: Examples of model outputs from Model 14 for disc and ellipse experiments.

Figure 7-9 and Figure 7-10 shows examples from Model 7 and Model 14 for disc and ellipse experiments with optimised parameters. Table 7-3 summarizes the final SCAPD results from the best performance models (Model 14 with rules of different velocity, directional movements and border effects) compared to the initial random models (Model 1 without any specific motility rules in the original 16 models). Since the mathematical properties of disc and ellipse, we took the results from the random model as the baseline and checked the improvements of our models by comparing them against the initial random models. The improvements in our models reach up to 70% and 77% performance improvements in terms of SCAPD for disc and ellipse micropatterns respectively.

Table 7-3: SCAPD results from random models and best performance models.

	Disc	Ellipse
Random model (Model 1)	0.26	0.31
Best performance model (Model 14)	0.08	0.07
Improvement	70%	77%

7.5 Results of time consumption measurements

In addition, we assessed the computational time required to test Model 7 and Model 14 on both disc and ellipse micropatterns. We tested out the models on a high performance computing cluster with requesting 64 GB memory. We repeated the timing computations 100 times for each model. The total time consumption for Model 7 and Model 14 is listed in Table 7-4.

Table 7-4: Time consuming of running Model 7 and Model 14 for disc and ellipse micropatterns.

Models	Total running times (22500 runs)	Average running time per run
Model 7 (disc)	800 minutes	0.0356 minutes
Model 7 (ellipse)	1970 minutes	0.0876 minutes
Model 14 (disc)	910 minutes	0.0404 minutes
Model 14 (ellipse)	23225 minutes	0.1033 minutes

Chapter 8

8 Discussion

In Chapter 8, we summarise our key findings referring back to our previous chapters to provide better content on how everything is coming together in this work. We discuss the limitations of our study as well as outlined further work that could be undertaken in this research area.

8.1 Key findings

This study investigated the potential of applying agent-based modelling in ESCs pattern formation to acquire the knowledge of cell behaviours. Based on our experimental data, we obtained a specific pattern formation in ESCs on the aggregated level (as described in Section 6.1). Towards understanding the driving power in cell behaviours that lead to observed pattern formation, four biologically plausible rules of cell behaviours (focusing on cell motility) were proposed (as described in Section 5.3). Subsequently, 16 agent-based models were constructed to test these four potential rules along with all their combinations (Section 5.2 and Section 5.4). To evaluate our model outputs, we introduced a new metric, SCAPD (Section 4.4.3), to quantify the differences between the models derived pattern formation and the experimental ESCs pattern formation. After parameter optimisation, the best of models improves fitness by 70% and 77% over the random models for a discoidal or an ellipsoidal geometrical confinement respectively.

Previous studies on cell sorting and tissue morphogenesis have described motility driven pattern formation and provided some insights at the molecular level (Halbleib and Nelson, 2006). A review of previous biological studies

working on understanding cell behaviours and pattern formation from the molecular level was provided in Section 2.1. Different from previous studies, in this study we focused on investigating the impact of cell motility on pattern formation on a cell level. We hypothesized that high-level (cell level) models of cell social interaction and decisions can give sufficient predictive power. A series of minimal probabilistic models were constructed based on our hypothesis. Furthermore, we obtained meaningful outputs from our models. A more detailed interpretation of our results will be provided later in this section.

With regard to the modelling approaches, multiple different modelling methods have been applied to study stem cells at a population level to reproduce the pattern dynamics by generating minimal models (Pir and Novère, 2015). More detailed descriptions of former work were provided in Section 2.3. For example, Libby and colleagues have applied cellular Pott models to human pluripotent stem cells, enabling a machine learning optimisation approach to predict experimental conditions that yield targeted multicellular patterns (Libby *et al.*, 2019). Multiple mathematical models were applied to increase the precision of modelling stem cell proliferation (Tabatabai *et al.*, 2011) and investigating stem cell self-renewal (Stiehl and Marciniak-czochra, 2017). Compared to cellular Pott models, a modelling approach that is targeted to achieve the pattern emergence with the possible simplest computation, agent-based modelling provides higher freedom, since cells are free to move and interact with other cells and their environment. Compared to equational models, the behavioural rules for agent-based models, with appropriate designing, can be explained to people straightforwardly since the rules are naturally understandable. The rules describing interactions and behaviours of cells are easier to interpret into biological terms. Hence, agent-based modelling is a suitable modelling method for investigating cell motility. In addition, agent-based modelling have been widely applied in cell biology. For example, Briers and colleagues applied agent-based modelling to study specific pattern formation in ESCs differentiation (Briers *et al.*, 2016). In this study, we applied agent-based modelling to investigate specific observed pattern formation in ESCs with focusing cell motility, and tested a wide range of motility rules to obtain the

minimal rules that can reproduce the pattern formation. Agent-based modelling allows us to model cell motility intuitively and gives insight of cell motility.

As one of the challenges of applying agent-based modelling in morphogenesis is the evaluation of patterning (Glen, Kemp and Voit, 2019), in this study, we applied multiple existing metrics for evaluation to assess model performance, including KL divergence, EMD, and CRPS. The results from these metrics, as described in Section 5.5.1, show the shortcoming of these existing approaches in our case due to the high randomness in the experimental data. The underlying problem was explained in Section 4.4.2 by taking artificial data as an example. Hence, we need a new metric, which can capture the key features of the desired patterns, to assess model performance in this specific application. We introduced a new metric, SCAPD, quantifying differences between model-derived pattern formation and the experimental ESCs pattern formation. SCAPD evaluates the models' results by calculating the distance between the density plot of models' results and empirical data. SCAPD captures the key features of the desired patterns by generating the borders of HDA and evaluates the probabilistic modelling by quantifying the aggregated model outputs according to the features from the desired patterns. Based on SCAPD, we quantified our models' results with different rules instead of visually estimating the results. The high variability in empirical data could be the reason that existing approaches for evaluation resulting in outcomes that do not follow the visual impression we had from the resulting patterns.

SCAPD is a novel metric to calculate the distance between probabilistic ground truth and probabilistic models. Beyond cell biology, SCAPD might be generalizable and applicable to other domains to assess probabilistic model performance based on the specific pattern formation observed in the probabilistic ground truth.

We applied the grid search for parameter optimisation to bring our model outputs closer to the empirical data. We tested different values of sensing radius and standard deviation. We also tested different values for angle change in rule iv and showed that the model output is not sensitive to this

parameter. In consequence, we observed that the SCAPD of Model 7 and Model 14 after parameter optimisation is considerably lower compared to the original outputs. Specifically, we computed SCAPD as 0.08 and 0.07 for disc and ellipse micropatterns respectively after revising rules and parameters optimisation. Compared to the initial random models, the models' performance was improved by 70% and 77% for disc and ellipse micropatterns, respectively. The represented models can probabilistically produce broadly realistic pattern formation (when compared to the empirical data).

In summary, this study has touched upon many topics in both biological and mathematical fields. Even though the rules in our minimal models might differ from the mechanisms in the real world, our models provide valuable output in proposing new testable rules, providing predictive tools, and holding the potential to reduce the cost required in biological labs. Some key contributions from this study are highlighted below:

- We generated probabilistic models to reproduce the observed pattern formation in ESCs. We demonstrated that we could replicate the pattern formation with a quantified level of uncertainty.

- Since the evaluation results based on existing metrics do not follow the visual impression we got from the outputs, we developed a new metric SCAPD to evaluate the models' results at a quantified level.

- Based on the results of SCAPD from 16 models, we found that Model 7 and Model 14 have the best performance, which consistent for both disc and ellipse micropatterns. Model 7 is built using different velocity for T+/T- cells and directional movements based on their cell type and their neighbouring cells, while Model 14 has an additional rule compared to Model 7 including the border effect.

- We proposed new testable rules for understanding the mechanisms of pattern formation based on the presented work. The new hypotheses

are that the pattern formation can be achieved by engineering cell speed and the level of adherence of cells by using synthetic biology.

8.2 Limitations

We want to acknowledge some limitations of our exploratory work. Firstly, some methodological limitations are directly due to the limitations imposed by the experimental data available to this study. Due to the sample size of the experimental data made available to this study (186 images for disc micropatterns and 152 images for ellipse micropatterns), we were not able to apply the approaches that require a big sample dataset to investigate the relationship between cell behaviours and pattern formation (e.g. advanced machine learning tools). However, with agent-based modelling, we provide more translatable rules to biologists for further experiments. Secondly, our experimental data does not include images for initial cell seeding, hence, we do not have corresponded data for the relationship between cell seeding and final cell patterning. In our models, cells were randomly seeded within the disc or ellipse micropatterns, which is the same as in reality. However, since we do not have corresponding images of the initial and final states of cell growth, it is not possible to verify the effects of cell seeding. Moreover, our experimental data was only collected as images 48 hours after seeding, which means we do not have data on the process of cell growing and patterning. Hence, we were not able to investigate the state transition and comparing model outputs and experimental data throughout the time. It is possible that our models only work for a specific timestamp (48 hours of cell growth according to our experimental data). Even though we cannot prove the consistency between our model and cell growth in real life throughout the 48 hours of cell growth, our models provide a predictive tool for patterning in a period with specific settings.

In addition, our models were based on a series of assumptions aiming at investigating the impacts of cell motility rules. For example, our models are 2D

and we did not take into account cell division, differentiation and apoptosis. Similarly, cell shape, the chemistry of cell signalling, and the environment have not been explored in the simulations in our models. Hence, our models do not suggest that the cell behaviours in reality are as same as the rules we demonstrated in our models. However, our models provide valued contributions to understanding ESCs pattern formation, engineering ESCs and further studies on understanding the mechanisms of ESCs (more explanations in Section 8.1 and Section 8.4).

Basic to our approach is the hypothesis that behavioural rules which do not deeply reference chemistry, environment or cellular structure can account substantially for the patterning observe in benchmark data sets. Our modelling aims to find a minimal set of rules that are as simple as we can make them while also having broad plausibility in terms of the underlying physics and biology of the systems concerned (even though we do not directly model the physics or biology). Even our models represent the idealisations of the physical systems which are not unique for the complex systems in reality, the model outputs are valuable because they can be used as tools for predicting pattern formation in related contexts and, in more distantly related contexts, they can be re-parameterised with less effort than would be required with more complex models.

In this study, we demonstrated that we could replicate the pattern formation with a quantified level of uncertainty. We proposed new testable rules for understanding the mechanisms of pattern formation based on the presented work.

8.3 Future work

More experiments and tests will allow realising a deeper analysis of cell behaviours and obtaining a better understanding of the driving power in pattern formation. However, due to the limitation of time, many different adaptations,

tests, and experiments have been left for future work. Some biological experiments will need to be supported by biologists. Here we list some ideas for future research (for both biological and computational modelling studies):

- **Test our models on data from newly shaped micropatterns:** Our model outputs are consistent with disc and ellipse-shaped micropatterns. It would be interesting to test the universality of our rules by using other different shaped micropatterns (e.g. an ellipse with a hole in the middle). Additional experiments will be required to collect data from newly shaped micropatterns. Ideally, these extra experiments and data collection will be carried out with the same conditions and processes as on disc and ellipse micropatterns. Our models can be adjusted to micropatterns with new shapes easily by changing the setting for environment agents. The same approach of verification can be applied to new data to compare the results from experiments and simulations. If the pattern produced from our model is close enough to the new experimental data, then we can further verify the universality of our rules.

- **Test our new testable hypotheses of cell behaviours:** Based on our model outputs, we proposed the possible combination of motility rules that lead to the observed pattern formation. These rules are testable but still waiting for validation in the wet lab. As demonstrated by Model 7 in our study, differential speed and directed movements based on cells neighbours can lead to the pattern formation we observed. Hence, the new hypotheses are that the pattern formation can be achieved by engineering cell speed and the level of adherence of cells by using synthetic biology (Cachat *et al.*, 2014; Davies, 2017). Further experiments will be required to verify these hypotheses.

- **Explore the threshold chosen in calculating SCAPD for model evaluation:** High-density areas were defined to calculate SCAPD for quantifying the distance between experimental data and model outputs.

Currently, the threshold of defining the border of the high-density areas are the mean of the max and min value of density in all grid spaces. The evaluation results might be affected by the threshold value. Further analysis is required to explore the effect from the threshold chosen and the possibility of improving model evaluation results by choosing different threshold.

- **Explore the aggregation process for disc micropatterns:** Different patterns are observed in disc experimental data. T+ cells stay on the whole border of the disc in some colonies, while in some colonies they only stay in one direction of the border instead of the whole ring shape. Some asymmetric distribution may be hidden after the current aggregation process. It is worth exploring different processes for aggregation and have a further investigation of the asymmetricity of the patterns on disc micropatterns.

- **Inverting models to analyse the corresponding relationship between the initial state and final state:** With inverting models, we can propose the potential initial states for desired patterns with specific cell behaviours. The corresponding relationships between the initial state (how do we seed the cells) and desired final state (the formation of the desired pattern) can be verified by more experiments. These researches can further support the future study on engineering cell motility by using synthetic biology to obtain desired patterns.

- **Investigate the state transitions in our models and dynamic cell behaviours in real data throughout the time:** If some videos of patterning can be collected from wet labs, it would allow us to investigate cell behaviours throughout the time. We can generate time series plots of multiple different metrics from both experimental data and our simulations. The distance between experimental data results and simulation outputs can give us a hint of the similarity of our proposed rules and the cell behaviours in reality.

Through these experiments and tests, we will provide more contributions to regenerative medicine by increasing the robustness of achieving desired pattern formation and having a better understanding of the driving power of pattern formation from the population level.

8.4 Conclusion

Contemporary work has demonstrated the importance of studying ESCs since they hold great potential for potentially developing novel treatments for a large number of diseases. Even though there are many studies carried out to understand cell behaviours from the molecular level, the mechanisms of pattern formation during embryonic development remain poorly understood due to the complexity of cell dynamic behaviours (as described in Section 2.1). In this study, we fill in the gap of studying ESCs behaviours from the cell level by investigating their spatial pattern formation upon geometrical confinement.

A key hypothesis in our study was that there is sufficient predictive power in cell behaviours and interactions at a cellular level. The complexity of cell behaviours at a molecular level need not be a barrier if we can model the pattern formation simply by modelling their social interaction decisions at a higher level. Based on our hypothesis, we generated a minimalist, agent-based probabilistic model to reproduce the observed pattern formation in ESCs.

In this study, we analysed our experimental data collected from a wet lab after randomly seeding mESCs on different shaped micropatterns. We described the interesting patterns we observed from disc and ellipse micropatterns along with the explanations of the variability in our experimental data. Besides, we assessed the proximity of different types of mESCs by applying two different measures (minimum spanning tree and average distance for each query object to five nearest targets). The results of the proximity assessment support us to propose cell behaviour rules for our models.

The focus of our study was to build a probabilistic agent-based model to reproduce the pattern formation with a minimal set of cell behaviour rules. We constructed our models from scratch by setting up two types of agents to represent cells and the environment. We proposed four biologically plausible rules of cell motility and tested all combinations of these rules. Our model comes with a user-friendly interface that allows users to modify the parameters easily.

We evaluated our model outputs by comparing them against our experimental data. We tested multiple existing metrics and the results do not follow the visual impression we got. Hence, we proposed a novel evaluation metric SCAPD. After parameter optimisation with applying grid search, we brought our model outputs closer to our experimental data. Through revising cell motility rules and parameters optimisation, we improved the model performance by about 70% and 77% compared to the initial random model for disc and ellipse micropatterns respectively.

Towards better understanding and controlling embryonic development, our study fills in the specific gap of analysing spatial pattern formation upon geometrical confinement on a cell level. Our models, representing an idealised physical system, illustrated a minimal set of plausible behavioural rules responsible for the observed pattern formation. Even though the consistency between these plausible rules and the mechanisms in reality is waiting for further biological experiments, new testable rules were proposed based on model outputs. Our models provide opportunities for engineering the cells in reality to achieve the desired pattern formation. Our study contributes to modelling ESCs and facilitates biological studies by reducing the need for extensive and costly experiments. We see this study as a step towards extending the current understanding of ESCs pattern formation, which may facilitate the development of novel stem cell therapies.

References

Aguayo-Mazzucato, C. and Bonner-Weir, S. (2010) 'Stem cell therapy for type 1 diabetes mellitus', *Nature Reviews Endocrinology*. Nature Publishing Group, pp. 139–148. doi: 10.1038/nrendo.2009.274.

Allan, R. J. (2010) *Survey of Agent Based Modelling and Simulation Tools*.

Amack, J. D. and Manning, M. L. (2012) 'Knowing the boundaries: Extending the differential adhesion hypothesis in embryonic cell sorting', *Science*. American Association for the Advancement of Science, pp. 212–215. doi: 10.1126/science.1223953.

Aoi, T. *et al.* (2008) 'Generation of pluripotent stem cells from adult mouse liver and stomach cells', *Science*. Science, 321(5889), pp. 699–702. doi: 10.1126/science.1154884.

Avior, Y., Sagi, I. and Benvenisty, N. (2016) 'Pluripotent stem cells in disease modelling and drug discovery', *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, 17(3), pp. 170–182. doi: 10.1038/nrm.2015.27.

Beccari, L. *et al.* (2018) 'Multi-axial self-organization properties of mouse embryonic stem cells into gastruloids', *Nature*. Springer US, 562(11), pp. 272–276. doi: 10.1038/s41586-018-0578-0.

Beddington, R. S. P., Rashbass, P. and Wilson, V. (1992) 'Brachyury - a gene affecting mouse gastrulation and early organogenesis', *Development*, 116, pp. 157–165.

Ben-Haim, N. *et al.* (2006) 'The Nodal Precursor Acting via Activin Receptors Induces Mesoderm by Maintaining a Source of Its Convertases and BMP4', *Developmental Cell*. Elsevier, 11(3), pp. 313–323. doi: 10.1016/j.devcel.2006.07.005.

Bender, E. A. and Willianson, S. G. (2010) *Lists, Decisions and Graphs - With an Introduction to Probability*. University of California at San Diego. Available at: http://freecomputerbooks.com/Lists-Decisions-and-Graphs.html (Accessed: 8 May 2020).

Bianconi, E. *et al.* (2013) 'An estimation of the number of cells in the human body', *Annals of Human Biology*. Taylor & Francis, 40(6), pp. 463–471. doi: 10.3109/03014460.2013.807878.

Bilodeau, S. *et al.* (2009) 'SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state', *Genes and Development*, 23, pp. 2484–2489. doi: 10.1101/gad.1837309.Cole.

Blin, G. *et al.* (2018) ‘Geometrical confinement controls the asymmetric patterning of Brachyury in cultures of pluripotent cells’, *Development*, 145, p. dev.166025. doi: 10.1242/dev.166025.

Blin, G. *et al.* (2019) ‘Nessys: A new set of tools for the automated detection of nuclei within intact tissues and dense 3D cultures’, *PLoS Biology*. Public Library of Science, 17(8), p. e3000388. doi: 10.1371/journal.pbio.3000388.

Botev, Z. I., Grotowski, J. F. and Kroese, D. P. (2010) ‘Kernel density estimation via diffusion’, *Annals of Statistics*, 38(5), pp. 2916–2957. doi: 10.1214/10-AOS799.

Brennan, J. *et al.* (2001) ‘Nodal signalling in the epiblast patterns the early mouse embryo’, *Nature*, 411(6840), pp. 965–969. doi: 10.1038/35082103.

Briers, D. *et al.* (2016) ‘Pattern Synthesis in a 3D Agent-Based Model of Stem Cell Differentiation’, in *2016 IEEE 55th Conference on Decision and Control (CDC)*. Las Vegas: IEEE, pp. 4202–4207. doi: 10.1109/CDC.2016.7798907.

Brink, S. C. Van Den *et al.* (2014) ‘Symmetry breaking , germ layer specification and axial organisation in aggregates of mouse embryonic stem cells’, *Development*, 141, pp. 4231–4242. doi: 10.1242/dev.113001.

Britton, G. *et al.* (2019) ‘A novel self-organizing embryonic stem cell system reveals signaling logic underlying the patterning of human ectoderm’, *Development*, 146(20). doi: 10.1242/dev.179093.

Cachat, E. *et al.* (2014) ‘A library of mammalian effector modules for synthetic morphology’, *Journal of Biological Engineering*. BioMed Central Ltd., 8(1), p. 26. doi: 10.1186/1754-1611-8-26.

Chacón-Martínez, C. A., Koester, J. and Wickström, S. A. (2018) ‘Signaling in the stem cell niche: regulating cell fate, function and plasticity’, *Development*. The Company of Biologists, 145(15), p. dev165399. doi: 10.1242/dev.165399.

Chen, Y.-C. (2017) ‘A tutorial on kernel density estimation and recent advances’, *Biostatistics & Epidemiology*. Taylor & Francis, 1(1), pp. 161–187. doi: 10.1080/24709360.2017.1396742.

Cheng, D. *et al.* (2017) ‘Regulation of human and mouse telomerase genes by genomic contexts and transcription factors during embryonic stem cell differentiation’, *Scientific Reports*. Nature Publishing Group, 7(1), pp. 1–12. doi: 10.1038/s41598-017-16764-w.

Cover, T. M. and Thomas, J. A. (2006) *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. USA: Wiley-Interscience.

D’Inverno, M. and Saunders, R. (2005) ‘Agent-based modelling of stem cell self-organisation in a niche’, *Lecture Notes in Computer Science (including*

subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 3464 LNAI, pp. 52–68. doi: 10.1007/11494676_4.

Danos, V. *et al.* (2008) 'Rule-Based Modelling of Cellular Signalling', *Web*, pp. 1–25. doi: 10.1007/978-3-540-74407-8_3.

Davidson, M. D., Ware, B. R. and Khetani, S. R. (2015) 'Stem cell-derived liver cells for drug testing and disease modeling', *Discovery Medicine*. Solariz, Inc., 19(106), pp. 349–358. Available at: /pmc/articles/PMC5768200/?report=abstract (Accessed: 17 September 2020).

Davies, J. (2017) 'Using synthetic biology to explore principles of development', *Development (Cambridge)*. Company of Biologists Ltd, pp. 1146–1158. doi: 10.1242/dev.144196.

Davis, R. P. *et al.* (2011) 'Pluripotent stem cell models of cardiac disease and their implication for drug discovery and development', *Trends in Molecular Medicine*. Elsevier Ltd, 17(9), pp. 475–484. doi: 10.1016/j.molmed.2011.05.001.

Deglincerti, A. *et al.* (2016) 'Self-organization of the in vitro attached human embryo', *Nature*. Nature Publishing Group, 533(7602), pp. 251–254. doi: 10.1038/nature17948.

Discher, D. E., Mooney, D. J. and Zandstra, P. W. (2009) 'Growth factors, matrices, and forces combine and control stem cells.', *Science (New York, N.Y.)*, 324(5935), pp. 1673–7. doi: 10.1126/science.1171643.

Dobrovolska'ia-Zavadskaa, N. (1927) 'Sur la mortification spontanEe de la queue chez la souris nouveau-nEe et sur l'existence d'un caractEre heriditaire "non-viable"', *C R Soc Biol*, 97, pp. 114–116.

Evans, M. J. and Kaufman, M. H. (1981) 'Establishment in culture of pluripotential cells from mouse embryos', *Nature*, pp. 154–156. doi: 10.1038/292154a0.

Falconnet, D. *et al.* (2006) 'Surface engineering approaches to micropattern surfaces for cell-based assays', *Biomaterials*. Elsevier, pp. 3044–3063. doi: 10.1016/j.biomaterials.2005.12.024.

Fazzio, T. G., Huff, J. T. and Panning, B. (2008) 'An RNAi Screen of Chromatin Proteins Identifies Tip60-p400 as a Regulator of Embryonic Stem Cell Identity', *Cell*, 134(1), pp. 162–174. doi: 10.1016/j.cell.2008.05.031.An.

Feynman, R. (1965) *The Character of Physical Law*. Available at: https://books.google.co.uk/books?hl=en&lr=&id=SJNPDgAAQBAJ&oi=fnd&pg=PP6&dq=gravitational+laws+feynman&ots=Vb2jPNrlrq&sig=F66T9z30pH_g5xglqe9zsaOsVGc&redir_esc=y#v=onepage&q=gravitational laws feynman&f=false (Accessed: 15 November 2020).

Flaim, C. J., Chien, S. and Bhatia, S. N. (2005) 'An extracellular matrix microarray for probing cellular differentiation', *Nature Methods*, 2(2), pp. 119–125. doi: 10.1038/NMETH736.

Foty, R. A. and Steinberg, M. S. (2005) 'The differential adhesion hypothesis: A direct evaluation', *Developmental Biology*. Academic Press Inc., 278(1), pp. 255–263. doi: 10.1016/j.ydbio.2004.11.012.

Freund, C. and Mummery, C. L. (2009) 'Prospects for pluripotent stem cell-derived cardiomyocytes in cardiac cell therapy and as disease models', *Journal of Cellular Biochemistry*. John Wiley & Sons, Ltd, 107(4), pp. 592–599. doi: 10.1002/jcb.22164.

Fuchs, E., Tumbar, T. and Guasch, G. (2004) 'Socializing with the neighbors: Stem cells and their niche', *Cell*, 116(6), pp. 769–778. doi: 10.1016/S0092-8674(04)00255-7.

Gan, Q. *et al.* (2007) 'Concise Review: Epigenetic Mechanisms Contribute to Pluripotency and Cell Lineage Determination of Embryonic Stem Cells', *Stem Cells*, 25(1), pp. 2–9. doi: 10.1634/stemcells.2006-0383.

Garijo, N. *et al.* (2012) 'Stochastic cellular automata model of cell migration, proliferation and differentiation : Validation with in vitro cultures of muscle satellite cells', *Journal of Theoretical Biology*. Elsevier, 314, pp. 1–9. doi: 10.1016/j.jtbi.2012.08.004.

Gattazzo, F., Urciuolo, A. and Bonaldo, P. (2014) 'Extracellular matrix: A dynamic microenvironment for stem cell niche', *Biochimica et Biophysica Acta - General Subjects*. Elsevier B.V., 1840(8), pp. 2506–2519. doi: 10.1016/j.bbagen.2014.01.010.

Gierer, A. and Meinhardt, H. (1972) 'A theory of biological pattern formation', *Kybernetik*, 12(1), pp. 30–39. doi: 10.1007/BF00289234.

Ginis, I. *et al.* (2004) 'Differences between human and mouse embryonic stem cells', *Developmental Biology*, 269(2), pp. 360–380. doi: 10.1016/j.ydbio.2003.12.034.

Glen, C. M., Kemp, M. L. and Voit, E. O. (2019) 'Agent-based modeling of morphogenetic systems: Advantages and challenges', *PLoS Computational Biology*. Public Library of Science, 15(3). doi: 10.1371/journal.pcbi.1006577.

Gneiting, T. and Raftery, A. E. (2007) 'Strictly Proper Scoring Rules , Prediction , and Estimation', *Journal of the American Statistical Association*, 102(477), pp. 359–378. doi: 10.1198/016214506000001437.

Gong, J. *et al.* (2008) 'Effects of extracellular matrix and neighboring cells on induction of human embryonic stem cells into retinal or retinal pigment epithelial progenitors', *Experimental Eye Research*. Academic Press, 86(6), pp. 957–965. doi: 10.1016/j.exer.2008.03.014.

Gorochowski, T. E. (2016) ‘Agent-based modelling in synthetic biology’, *Essays in Biochemistry*. Portland Press Ltd, 60(4), pp. 325–336. doi: 10.1042/EBC20160037.

Guilak, F. *et al.* (2009) ‘Control of Stem Cell Fate by Physical Interactions with the Extracellular Matrix’, *Cell Stem Cell*. Elsevier Inc., 5(1), pp. 17–26. doi: 10.1016/j.stem.2009.06.016.

Guo, H. *et al.* (2013) ‘Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing’, *Genome Research*, 23, pp. 2126–2135. doi: 10.1101/gr.161679.113.

Halbleib, J. M. and Nelson, W. J. (2006) ‘Cadherins in development: Cell adhesion, sorting, and tissue morphogenesis’, *Genes and Development*, 20, pp. 3199–3214. doi: 10.1101/gad.1486806.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Hitchcock, F. L. (1941) ‘The Distribution of a Product from Several Sources to Numerous Localities’, *Journal of Mathematics and Physics*. Wiley, 20(1–4), pp. 224–230. doi: 10.1002/sapm1941201224.

Hooke, R. (1665) *Micrographia*. Available at: http://www.gutenberg.org/files/15491/15491-h/15491-h.htm (Accessed: 2 June 2020).

Horn, A. (1951) ‘On Sentences Which are True of Direct Unions of Algebras’, *J. Symbolic Logic*. Association for Symbolic Logic, 16(1), pp. 14–21. Available at: https://projecteuclid.org:443/euclid.jsl/1183731038.

Ivanova, N. *et al.* (2006) ‘Dissecting self-renewal in stem cells with RNA interference’, *Nature*, 442(3), pp. 533–538. doi: 10.1038/nature04915.

Ivey, K. N. *et al.* (2008) ‘MicroRNA Regulation of Cell Lineages in Mouse and Human Embryonic Stem Cells’, *Cell Stem Cell*, 2(3), pp. 219–229. doi: 10.1016/j.stem.2008.01.016.

Jang, J. H. and Schaffer, D. V. (2006) ‘Microarraying the cellular microenvironment’, *Molecular Systems Biology*. Nature Publishing Group, p. 39. doi: 10.1038/msb4100079.

Kailath, T. (1967) ‘The Divergence and Bhattacharyya Distance Measures in Signal Selection’, *IEEE Transactions on Communication Technology*, 15(1), pp. 52–60. doi: 10.1109/TCOM.1967.1089532.

Kamm, R. D. *et al.* (2018) ‘Perspective: The promise of multi-cellular engineered living systems’, *APL Bioengineering*, 2(4), p. 040901. doi: 10.1063/1.5038337.

Karsenti, E. (2008) ‘Self-organization in cell biology: a brief history’, *Na*, 9(March), pp. 1–8. Available at: papers2://publication/uuid/B1319C59-E350-44EB-BF57-0AEA080C93E3%5Cnfile:///data/Work/Bibliography/Papers2/Articles/Unknown/2008/2008 nrm2357.pdf.

Keener, J. and Sneyd, J. (2009a) *Mathematical Physiology I: Cellular Physiology*. Second Edi. Springer.

Keener, J. and Sneyd, J. (2009b) *Mathematical Physiology II: Systems Physiology*. Second Edi, *Book*. Second Edi. Springer. doi: 10.1007/978-0-387-79388-7.

Kruskal, J. (1956) ‘On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem’, *American Mathematical Society*, 7(1), pp. 48–50.

Kullback, S. and Leibler, R. A. (1951) ‘On Information and Sufficiency’, *Annals of Mathematical Statistics*. Institute of Mathematical Statistics, 22(1), pp. 79–86. doi: 10.1214/AOMS/1177729694.

Ladewig, J., Koch, P. and Brüstle, O. (2013) ‘Leveling Waddington: the emergence of direct programming and the loss of cell fate hierarchies.’, *Nature reviews. Molecular cell biology*. Nature Publishing Group, 14(4), pp. 1–12. doi: 10.1038/nrm3543.

Leeper, N. J., Hunter, A. L. and Cooke, J. P. (2010) ‘Stem cell therapy for vascular regeneration: Adult, embryonic, and induced pluripotent stem cells’, *Circulation*. Lippincott Williams & Wilkins, pp. 517–526. doi: 10.1161/CIRCULATIONAHA.109.881441.

Libby, A. R. G. *et al.* (2019) ‘Automated Design of Pluripotent Stem Cell Self-Organization’, *Cell Systems*. Elsevier Inc., 9, pp. 1–13. doi: 10.1016/j.cels.2019.10.008.

Van Liedekerke, P. *et al.* (2015) *Simulating tissue mechanics with agent-based models: concepts, perspectives and some novel results*, *Computational Particle Mechanics*. Springer International Publishing. doi: 10.1007/s40571-015-0082-3.

Lutolf, M. P. and Hubbell, J. A. (2005) ‘Synthetic biomaterials as instructive extracellular microenvironments for morphogenesis in tissue engineering’, *Nature Biotechnology*. Nature Publishing Group, pp. 47–55. doi: 10.1038/nbt1055.

M Batty (2000) *GeoComputation*, *books.google.com*. Available at: https://books.google.com/books?hl=en&lr=&id=zTc7Rl8F3sUC&oi=fnd&pg=PA96&dq=Geocomputation+using+cellular+automata&ots=SpYqmAmYji&sig=kHYCNFndZ-9rfCblNNja_jnX8VY (Accessed: 26 March 2020).

Madl, C. M. and Heilshorn, S. C. (2018) 'Engineering Hydrogel Microenvironments to Recapitulate the Stem Cell Niche', *Annual Review of Biomedical Engineering*. Annual Reviews, 20(1), pp. 21–47. doi: 10.1146/annurev-bioeng-062117-120954.

Maini, P. K. (2004) 'Using mathematical models to help understand biological pattern formation', *Comptes Rendus - Biologies*, 327(3), pp. 225–234. doi: 10.1016/j.crvi.2003.05.006.

Maini, P. K. *et al.* (2012) 'Turing's model for biological pattern formation and the robustness problem', *Interface Focus*, 2(4), pp. 487–496. doi: 10.1098/rsfs.2011.0113.

Malvino, A. and Bates, D. (2016) *Electronic principles*. Eighth edition. New York NY: McGraw-Hill Education.

Marée, A. F. M., Grieneisen, V. A. and Hogeweg, P. (2007) 'The Cellular Potts Model and Biophysical Properties of Cells, Tissues and Morphogenesis', in *Single-Cell-Based Models in Biology and Medicine*. Birkhäuser Basel, pp. 107–136. doi: 10.1007/978-3-7643-8123-3_5.

Marikawa, Y. *et al.* (2009) 'Aggregated P19 mouse embryonal carcinoma cells as a simple in vitro model to study the molecular regulations of mesoderm formation and axial elongation morphogenesis', *genesis*. John Wiley & Sons, Ltd, 47(2), pp. 93–106. doi: 10.1002/dvg.20473.

Martin, G. R. (1981) 'Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells', *Proceedings of the National Academy of Sciences of the United States of America*, 78(12), pp. 7634–7638. doi: 10.1073/pnas.78.12.7634.

Meier, J. J., Bhushan, A. and Butler, P. C. (2006) 'The potential for stem cell therapy in diabetes', *Pediatric Research*. Nature Publishing Group, pp. 65–73. doi: 10.1203/01.pdr.0000206857.38581.49.

Meinhardt, H. and Gierer, A. (2000) 'Pattern formation by local self-activation and lateral inhibition.', *BioEssays : news and reviews in molecular, cellular and developmental biology*, 22(8), pp. 753–60. doi: 10.1002/1521-1878(200008)22:8<753::AID-BIES9>3.0.CO;2-Z.

Miller, H. (2009) 'The SAGE handbook of spatial analysis'.

Monteagudo, Á. and Santos, J. (2015) 'Treatment analysis in a cancer stem cell context using a tumor growth model based on cellular automata', *PLoS ONE*. Public Library of Science, 10(7). doi: 10.1371/journal.pone.0132306.

Mori, H. *et al.* (2009) 'Self-organization of engineered epithelial tubules by differential cellular motility.', *Proceedings of the National Academy of Sciences*, 106(35), pp. 14890–14895. doi: 10.1073/pnas.0901269106.

Nava, M. M., Raimondi, M. T. and Pietrabissa, R. (2012) 'Controlling self-renewal and differentiation of stem cells via mechanical cues', *Journal of Biomedicine and Biotechnology*, 2012. doi: 10.1155/2012/797410.

Von Neumann, J. and Burks, A. W. (1966) *Theory of self-reproducing automata*. Urbana, University of Illinois Press. Available at: https://archive.org/details/theoryofselfrepr00vonn_0 (Accessed: 12 November 2020).

Niazi, M. A. and Hussain, A. (2017) 'Agent-based computing from multi-agent systems to agent-based Models: a visual survey', *Scientometrics*, 89(2), pp. 479–499. doi: 10.1007/s11192-011-0468-9.

Niessen, C. M. and Gumbiner, B. M. (2002) 'Cadherin-mediated cell sorting not determined by binding or adhesion specificity', *Journal of Cell Biology*. The Rockefeller University Press, 156(2), pp. 389–399. doi: 10.1083/jcb.200108040.

Nishikawa, S. I., Goldstein, R. A. and Nierras, C. R. (2008) 'The promise of human induced pluripotent stem cells for research and therapy', *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, pp. 725–729. doi: 10.1038/nrm2466.

Okano, H. *et al.* (2013) 'Steps toward safe cell therapy using induced pluripotent stem cells', *Circulation Research*. Lippincott Williams & WilkinsHagerstown, MD, pp. 523–533. doi: 10.1161/CIRCRESAHA.111.256149.

Okita, K., Ichisaka, T. and Yamanaka, S. (2007) 'Generation of germline-competent induced pluripotent stem cells', *Nature*. Nature Publishing Group, 448(7151), pp. 313–317. doi: 10.1038/nature05934.

Parzen, E. (1962) 'On Estimation of a Probability Density Function and Mode', *Annals of Mathematical Statistics*. Institute of Mathematical Statistics, 33(3), pp. 1065–1076. doi: 10.1214/AOMS/1177704472.

Pauklin, S., Pedersen, R. A. and Vallier, L. (2011) 'Mouse pluripotent stem cells at a glance', *Journal of Cell Science*, 124(22), pp. 3727–3732. doi: 10.1242/jcs.074120.

Phadnis, S. M. *et al.* (2015) 'Dynamic and social behaviors of human pluripotent stem cells.', *Scientific Reports*. Nature Publishing Group, 5(14209), pp. 1–12. doi: 10.1038/srep14209.

Pir, P. and Novère, N. Le (2015) 'Mathematical models of pluripotent stem cells: at the dawn of predictive regenerative medicine', 1386(December 2015). doi: 10.1007/978-1-4939-3283-2.

Poleszczuk, J. and Enderling, H. (2014) 'A High-Performance Cellular Automaton Model of Tumor Growth with Dynamically Growing Domains',

Applied Mathematics. Scientific Research Publishing, Inc, 05(01), pp. 144–152. doi: 10.4236/am.2014.51017.

Poleszczuk, J., Macklin, P. and Enderling, H. (2016) 'Agent-based modeling of cancer stem cell driven solid tumor growth', in *Methods in Molecular Biology*. Humana Press Inc., pp. 335–346. doi: 10.1007/7651_2016_346.

Politis, M. and Lindvall, O. (2012) 'Clinical application of stem cell therapy in Parkinson's disease', *BMC Medicine*. BioMed Central, p. 1. doi: 10.1186/1741-7015-10-1.

Prakasam, A. K., Maruthamuthu, V. and Leckband, D. E. (2006) 'Similarities between heterophilic and homophilic cadherin adhesion', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 103(42), pp. 15434–15439. doi: 10.1073/pnas.0606701103.

Prasad, A. *et al.* (2016) 'A review of induced pluripotent stem cell , direct conversion by reprogramming and oligodendrocyte differentiation', *Regenerative Medicine*, 11, pp. 181–191.

Pratt, V. (1987) 'Direct Least-Squares Fitting of Algebraic Surfaces', in *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*. New York, NY, USA: Association for Computing Machinery (SIGGRAPH '87), pp. 145–152. doi: 10.1145/37401.37420.

Prim, R. C. (1957) 'Shortest Connection Networks And Some Generalizations', *the Bell System Technical Journal*, 36(6), pp. 1389–1401. doi: 10.1002/j.1538-7305.1957.tb01515.x.

Rabajante, J. F. *et al.* (2015) 'Mathematical modeling of cell-fate specification: From simple to complex epigenetics', *Stem Cell Epigenetics*, (April), pp. 1–10. doi: 10.14800/sce.752.

Rabajante, J. F. and Babierra, A. L. (2015) 'Branching and oscillations in the epigenetic landscape of cell-fate determination', *Progress in Biophysics and Molecular Biology*, 117(2–3), pp. 240–249. doi: 10.1016/j.pbiomolbio.2015.01.006.

Ramalho-Santos, M. *et al.* (2002) '"Stemness": Transcriptional profiling of embryonic and adult stem cells', *Science*. American Association for the Advancement of Science, 298(5593), pp. 597–600. doi: 10.1126/science.1072530.

Rao, M. (2004) 'Conserved and divergent paths that regulate self-renewal in mouse and human embryonic stem cells', *Developmental Biology*. Academic Press Inc., pp. 269–286. doi: 10.1016/j.ydbio.2004.08.013.

Romito, A. and Cobellis, G. (2016) 'Pluripotent stem cells: Current understanding and future directions', *Stem Cells International*. Hindawi

Publishing Corporation, 2016. doi: 10.1155/2016/9451492.

Ronaghi, M. *et al.* (2009) 'Challenges of Stem Cell Therapy for Spinal Cord Injury: Human Embryonic Stem Cells, Endogenous Neural Stem Cells or Induced Pluripotent Stem Cells?', *Stem Cells*. John Wiley & Sons, Ltd, 28(1), p. N/A-N/A. doi: 10.1002/stem.253.

Rosental, B. *et al.* (2017) 'Coral cell separation and isolation by fluorescence-activated cell sorting (FACS)', *BMC Cell Biology*. BioMed Central Ltd., 18(1), p. 30. doi: 10.1186/s12860-017-0146-8.

Rosenthal, A., Macdonald, A. and Voldman, J. (2007) 'Cell patterning chip for controlling the stem cell microenvironment', *Biomaterials*. Elsevier, 28(21), pp. 3208–3216. doi: 10.1016/j.biomaterials.2007.03.023.

Rossant, J. and Joyner, A. L. (1989) 'Towards a molecular-genetic analysis of mammalian development', *Trends Genet*, 5(8), pp. 277–283.

Rubner, Y., Tomasi, C. and Guibas, L. J. (2000) 'The Earth Mover's Distance as a Metric for Image Retrieval', *International Journal of Computer Vision*, 40(2), pp. 99–121.

Saha, S. *et al.* (2006) 'Inhibition of human embryonic stem cell differentiation by mechanical strain', *Journal of Cellular Physiology*. John Wiley & Sons, Ltd, 206(1), pp. 126–137. doi: 10.1002/jcp.20441.

Savill, N. J. and Merks, R. M. H. (2007) 'The Cellular Potts Model in Biomedicine', in *Single-Cell-Based Models in Biology and Medicine*. Birkhäuser Basel, pp. 137–150. doi: 10.1007/978-3-7643-8123-3_6.

Schelling, T. C. (1971) 'Dynamic models of segregation', *The Journal of Mathematical Sociology*. Taylor & Francis Group , 1(2), pp. 143–186. doi: 10.1080/0022250X.1971.9989794.

Semrau, S. *et al.* (2017) 'Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells', *Nature Communications*. Springer US, 8(1096), pp. 1–16. doi: 10.1038/s41467-017-01076-4.

Setty, Y. (2012) 'Multi-scale computational modeling of developmental biology', *Bioinformatics*, 28(15), pp. 2022–2028. doi: 10.1093/bioinformatics/bts307.

Shannon, C. E. (1948) *A Mathematical Theory of Communication*, *The Bell System Technical Journal*.

Sheather, S. J. (1992) 'The performance of six popular bandwidth selection methods on some real data sets (with discussion)', *Computational Statistics*, 7, pp. 225–258, 271–281.

Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*,

Monographs on Statistics and Applied Probability, London: Chapman and Hall1986.

Simunovic, M. and Brivanlou, A. H. (2017) 'Embryoids , organoids and gastruloids : new approaches to understanding embryogenesis', *Development*, 144, pp. 976–985. doi: 10.1242/dev.143529.

Slack, J. M. W. (2007) 'Metaplasia and transdifferentiation: From pure biology to the clinic', *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, 8(5), pp. 369–378. doi: 10.1038/nrm2146.

Smith, J. (1997) 'Brachyury and the T-box genes', *Current Opinion in Genetics & Development*, 7(4), pp. 474–480.

Steinberg, M. S. (1970) 'Does differential adhesion govern self-assembly processes in histogenesis? Equilibrium configurations and the emergence of a hierarchy among populations of embryonic cells', *Journal of Experimental Zoology*. John Wiley & Sons, Ltd, 173(4), pp. 395–433. doi: 10.1002/jez.1401730406.

Stiehl, T. and Marciniak-czochra, A. (2017) 'Stem cell self-renewal in regeneration and cancer : Insights from mathematical modeling', *Current Opinion in Systems Biology*. Elsevier Ltd, 5, pp. 112–120. doi: 10.1016/j.coisb.2017.09.006.

Suzuki, A. *et al.* (2006) 'Maintenance of embryonic stem cell pluripotency by Nanog-mediated reversal of mesoderm specification', *Nature Clinical Practice Cardiovascular Medicine*, 3, pp. S114–S122. Available at: https://doi.org/10.1038/ncpcardio0442.

Tabar, V. and Studer, L. (2014) 'Pluripotent stem cells in regenerative medicine: challenges and recent progress', *Nat Rev Genet*. Nature Publishing Group, 15(2), pp. 82–92. doi: 10.1038/nrg3563.

Tabatabai, M. A. *et al.* (2011) 'Mathematical modeling of stem cell proliferation', *Medical and Biological Engineering and Computing*, 49(3), pp. 253–262. doi: 10.1007/s11517-010-0686-y.

Takahashi, K. *et al.* (2007) 'Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors', *Cell*, 131(5), pp. 861–872. doi: 10.1016/j.cell.2007.11.019.

Takahashi, K. and Yamanaka, S. (2006) 'Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors', *Cell*, 126(4), pp. 663–676. doi: 10.1016/j.cell.2006.07.024.

Tewary, M., Shakiba, N. and Zandstra, P. W. (2018) 'Stem cell bioengineering: building from stem cell biology', *Nature Reviews Genetics*. Nature Publishing Group, pp. 595–614. doi: 10.1038/s41576-018-0040-z.

Thomson, J. A. *et al.* (1995) ‘Isolation of a primate embryonic stem cell line’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 92(17), pp. 7844–7848. doi: 10.1073/pnas.92.17.7844.

Thomson, J. A. *et al.* (1998) ‘Embryonic stem cell lines derived from human blastocysts.’, *Science (New York, N.Y.)*. United States, 282(5391), pp. 1145–1147. doi: 10.1126/science.282.5391.1145.

Tong, X. *et al.* (2018) ‘Behaviour change in post-consumer recycling: Applying agent-based modelling in social experiment’, *Journal of Cleaner Production*. Elsevier Ltd, 187, pp. 1006–1013. doi: 10.1016/j.jclepro.2018.03.261.

Trudeau, R. J. (1993) *Introduction to Graph Theory*. Dover Pub. (Dover Books on Mathematics). Available at: https://books.google.co.jp/books?id=8nYH5OYEW24C.

Tsakiridis, A. *et al.* (2014) ‘Distinct Wnt-driven primitive streak-like populations reflect in vivo lineage precursors’, *Development*, 141(6), pp. 1209–1221. doi: 10.1242/dev.101014.

Turing, A. M. (1952) ‘The chemical basis of morphogenesis’, *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 237(641), pp. 37–72. doi: 10.1007/BF02459572.

Turner, D. A., Rue, P., *et al.* (2014) ‘Brachyury cooperates with Wnt/ -Catenin signalling to elicit Primitive Streak like behaviour in differentiating mouse ES cells.’, *BMC Biology*, 12(63), pp. 1–19. doi: 10.1101/003871.

Turner, D. A., Hayward, P. C., *et al.* (2014) ‘Wnt/ -catenin and FGF signalling direct the specification and maintenance of a neuromesodermal axial progenitor in ensembles of mouse embryonic stem cells’, *Development*, 141(22), pp. 4243–4253. doi: 10.1242/dev.112979.

Utomo, D. S., Onggo, B. S. and Eldridge, S. (2018) ‘Applications of agent-based modelling and simulation in the agri-food supply chains’, *European Journal of Operational Research*. Elsevier B.V., pp. 794–805. doi: 10.1016/j.ejor.2017.10.041.

Van De Vijver, G., Van Speybroeck, L. and Vandevyvere, W. (2003) ‘Reflecting on complexity of biological systems: Kant and beyond?’, *Acta Biotheoretica*, 51(2), pp. 101–140. doi: 10.1023/A:1024591510688.

Vining, K. H. and Mooney, D. J. (2017) ‘Mechanical forces direct stem cell behaviour in development and regeneration’, *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, pp. 728–742. doi: 10.1038/nrm.2017.108.

Walker, D. C. *et al.* (2004) ‘The epitheliome: Agent-based modelling of the social behaviour of cells’, in *BioSystems*. Elsevier, pp. 89–100. doi: 10.1016/j.biosystems.2004.05.025.

Wang, J. *et al.* (2006) 'A protein interaction network for pluripotency of embryonic stem cells', *Nature*, 444(7117), pp. 364–368. doi: nature05284 [pii]\r10.1038/nature05284.

Wang, M. *et al.* (2020) 'Predicting pattern formation in embryonic stem cells using a minimalist, agent-based probabilistic model', *Scientific Reports*. Nature Publishing Group, 10. doi: 10.1038/s41598-020-73228-4.

Wang, Z. *et al.* (2015) 'Simulating cancer growth with multiscale agent-based modeling', *Seminars in Cancer Biology*. Academic Press, pp. 70–78. doi: 10.1016/j.semcancer.2014.04.001.

Warmflash, A. *et al.* (2014) 'A method to recapitulate early embryonic spatial patterning in human embryonic stem cells', *Nature Methods*. Nature Publishing Group, 11(8), pp. 847–854. doi: 10.1038/nMeth.3016.

White, D. E. *et al.* (2013) 'Spatial Pattern Dynamics of 3D Stem Cell Loss of Pluripotency via Rules-Based Computational Modeling', *PLoS Computational Biology*, 9(3). doi: 10.1371/journal.pcbi.1002952.

Wilensky, U. and Rand, W. (2015) *An Introduction to Agent-Based Modeling*. Mit Press. Available at: http://www.jstor.org/stable/j.ctt17kk851.

Wilson, R. J. (1996) *Introduction to Graph Theory*, *John Wiley & Sons*.

Wisniewski, D., Lowell, S. and Blin, G. (2019) 'Mapping the Emergent Spatial Organization of Mammalian Cells using Micropatterns and Quantitative Imaging', *JoVE*. MyJoVE Corp, (146), p. e59634. doi: doi:10.3791/59634.

Wolfram, S. (1983) 'Statistical mechanics of cellular automata', *Reviews of Modern Physics*. American Physical Society, 55(3), pp. 601–644. doi: 10.1103/RevModPhys.55.601.

Wolpert, L. (1969) 'Positional information and the spatial pattern of cellular differentiation', *Journal of Theoretical Biology*. Academic Press, 25(1), pp. 1–47. doi: 10.1016/S0022-5193(69)80016-0.

Wu, S. M. and Hochedlinger, K. (2011) 'Harnessing the potential of induced pluripotent stem cells for regenerative medicine', *Nature Cell Biology*. Nature Publishing Group, 13(5). doi: 10.1038/ncb0511-497.

Young, R. A. (2011) 'Control of Embryonic Stem Cell State', *Cell*, 144(6), pp. 940–954. doi: 10.1016/j.cell.2011.01.032.Control.

Zhou, Q. *et al.* (2007) 'A gene regulatory network in mouse embryonic stem cells', *Proceedings of the National Academy of Sciences of the United States of America*, 104(42), pp. 16438–16443. doi: 10.1073/pnas.0701014104.

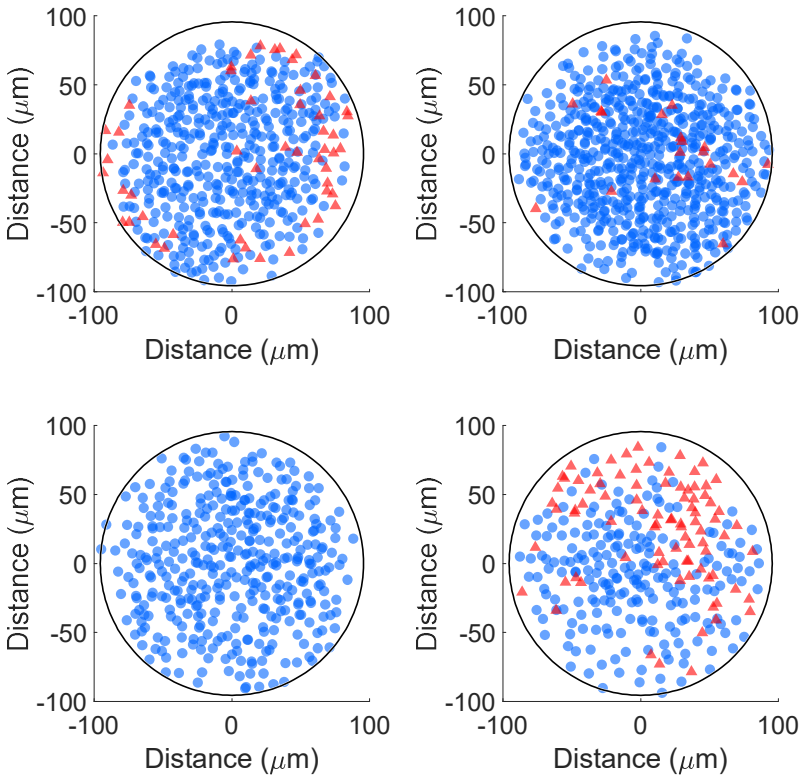
Zhou, Y. *et al.* (2016) 'Single cell studies of mouse embryonic stem cell (mESC) differentiation by electrical impedance measurements in a microfluidic device',

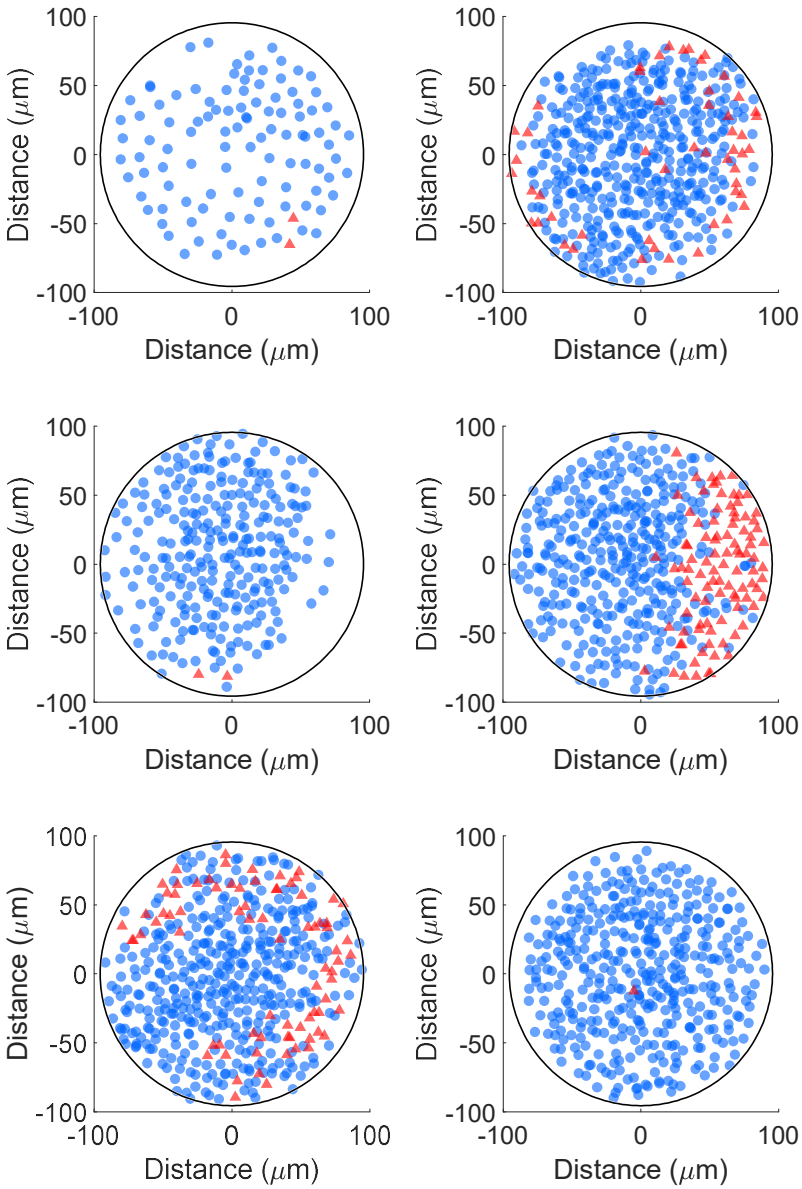
Biosensors and Bioelectronics. Elsevier, 81, pp. 249–258. doi: 10.1016/j.bios.2016.02.069.

Appendix A: Randomly selected samples from experimental data

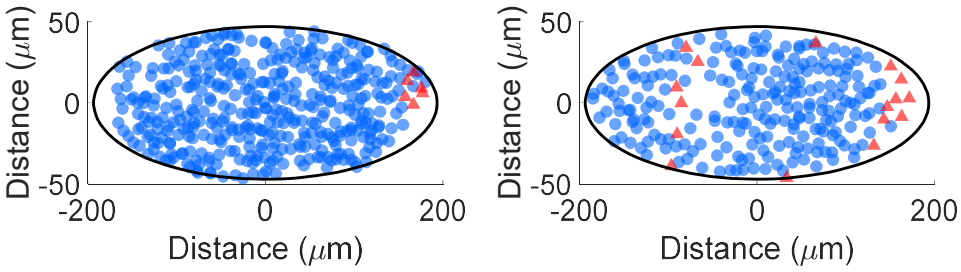
Ten indicative, randomly selected examples of cell colonies on disc and ellipse micropatterns from experimental data. Red triangle markers stand for T+ cells; blue circle markers stand for T- cells.

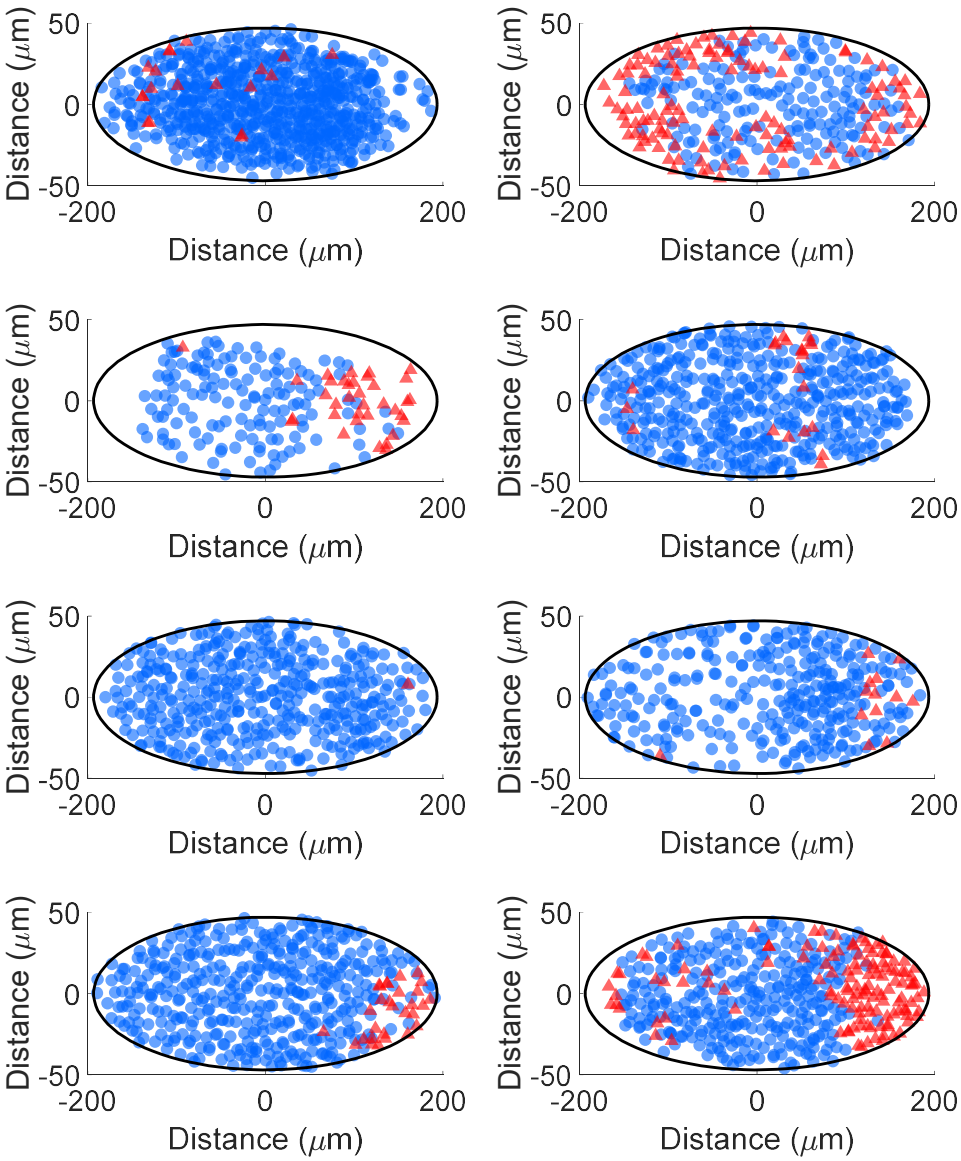
Disc experimental data





Ellipse experimental data




Distance (μm)
50
0
-50
-200
0
200
Distance (μm)

Appendix B: Pseudocode for model construction

In addition to pseudocode, the source code is available online at https://github.com/MinhongW/ESCs_models.

Algorithm: Randomly seeding cells (*initial_cells*)

```
% environment setup
Generate patches (grids) for the environment
Input the parameters for the size of the disc and ellipse
For each patch in patches
        If patch within the border of the micropattern
                Mark patch is inside the micropattern
                Mark patch as unoccupied
        Else
                Mark patch is outside the environment
End

% seed cells randomly within the micropattern
Input the number of T+ and T- cells
Input cell radius
For each T+ cell
        Set cell radius
        Set cell shape as a circle
        Set cell type as T+ cell
        Set cell speed as T+ cell max speed
        Move cell to one of the unoccupied patch
        Mark neighbouring patches within cell radius as occupied
End
For each T- cell
        Set cell radius
        Set cell shape as a circle
        Set cell type as T- cell
        Set cell speed as T- cell max speed
        Move cell to one of the unoccupied patch
        Mark neighbouring patches within cell radius as occupied
End
```

Algorithm: Transition

```
For each cell in cells
    If cell type == T- cell
        If rule 3 directional movement is selected
            Set cell heading by calling the function of direction movement for T-
        Else
            Set cell heading to a random direction
        If rule 4 border effect is selected
            Adjust cell heading by calling the function of border effect
        Get current destination patch by current heading and current speed
        If current destination is within the micropattern and
        the patches within the cell radius are not occupied
            Move cell to the destination
            Mark patches within the cell radius as occupied
        Else
            Cell stay at the original location

    If cell type == T+ cell
        If rule 3 directional movement is selected
            Set cell heading by calling the function of direction movement for T+
        Else
            Set cell heading to a random direction
        If rule 4 border effect is selected
            Adjust cell heading by calling the function of border effect
        Get current destination patch by current heading and current speed
        If current destination is within the micropattern and
        the patches within the cell radius are not occupied
            Move cell to the destination
            Mark patches within the cell radius as occupied
        Else
            Cell stay at the original location
End
```

Algorithm: The function of directional movement

```
Find all neighbouring cells within the sense distance R
If neighbouring cells exist
    Get a list of distance from the current cell to its all neighbouring cells
    Calculate the vectors of the forces based on cell locations and corresponding distance
    Sum up forces on x and y axis separately
    Generate noises of the forces sum on x and y axis separately
    Sum up the forces to get the final vector for cell heading direction
Else
    Set cell heading to a random direction
```

Algorithm: The function of border effect

If the destination patch is outside of the micropattern

    Find the closest patch to the cell along the micropattern border

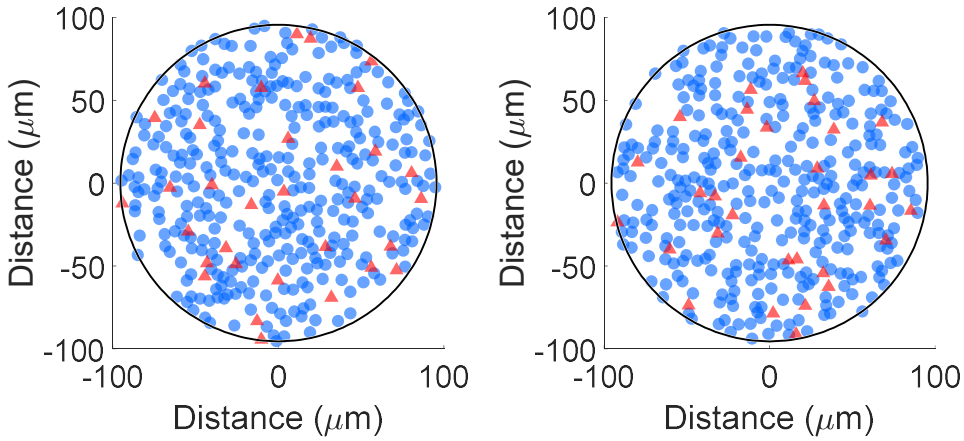
    Change the heading by comparing the current heading to the heading toward to the closest patch on the border

Appendix C: Randomly selected samples from model outputs

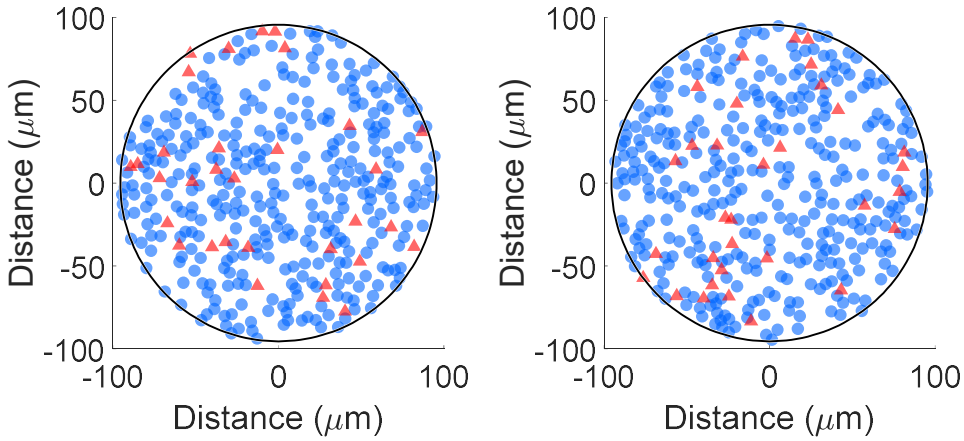
Two indicative, randomly selected model output examples for each model from both disc and ellipse experiments. Red triangle markers stand for T+ cells; blue circle markers stand for T- cells.

Disc experiments

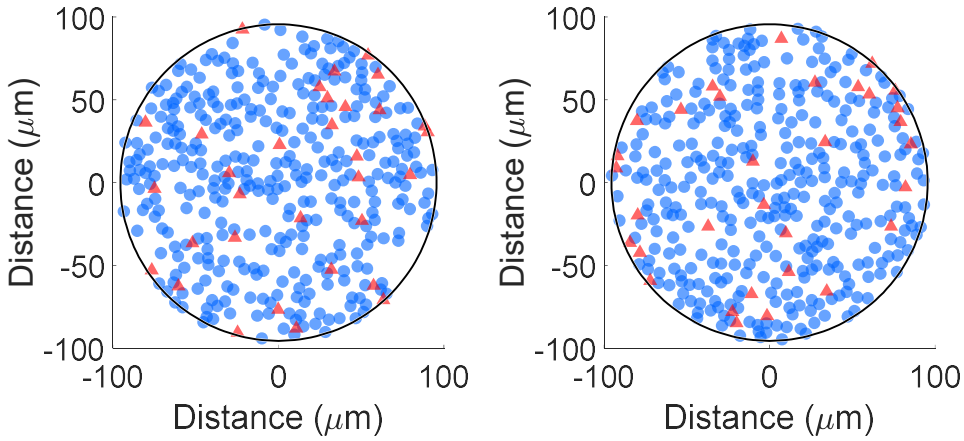
Model 1



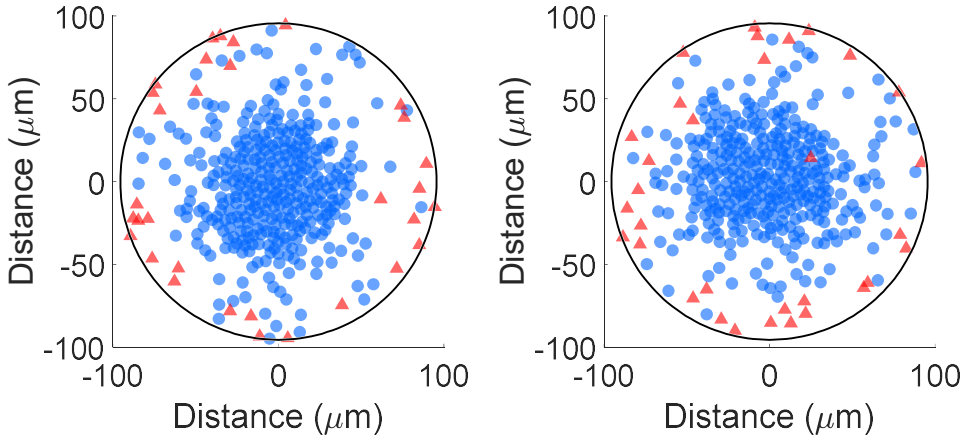
Model 2



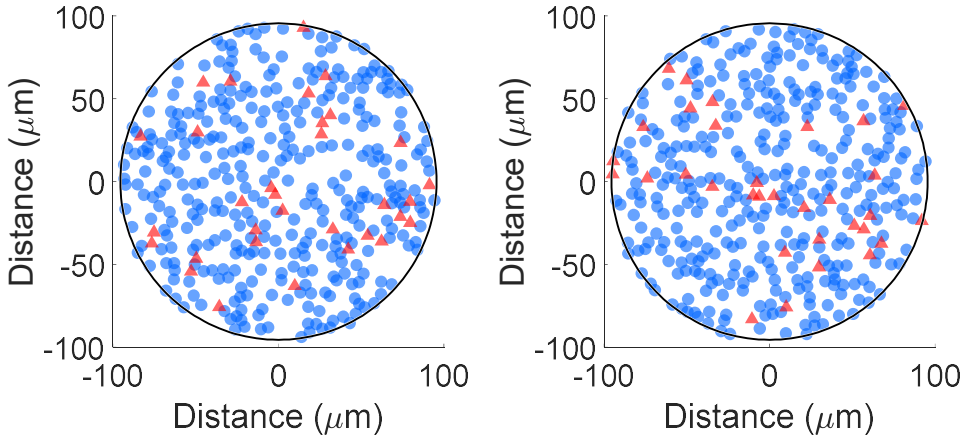
Model 3



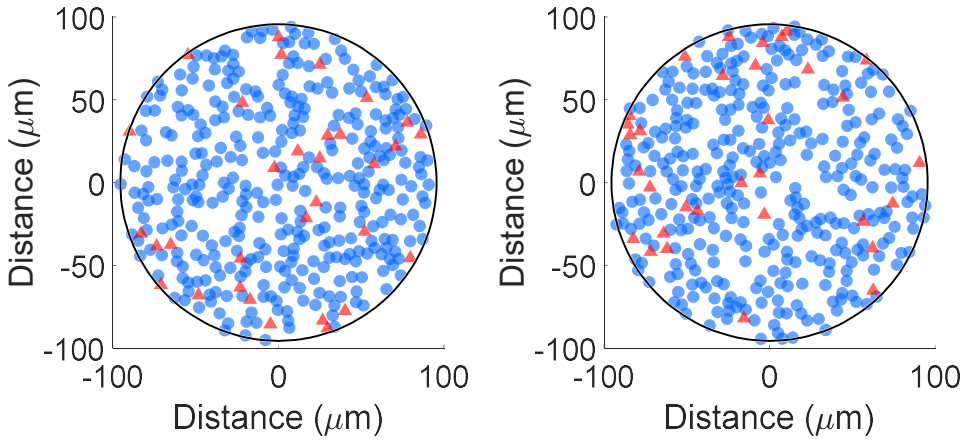
Model 4



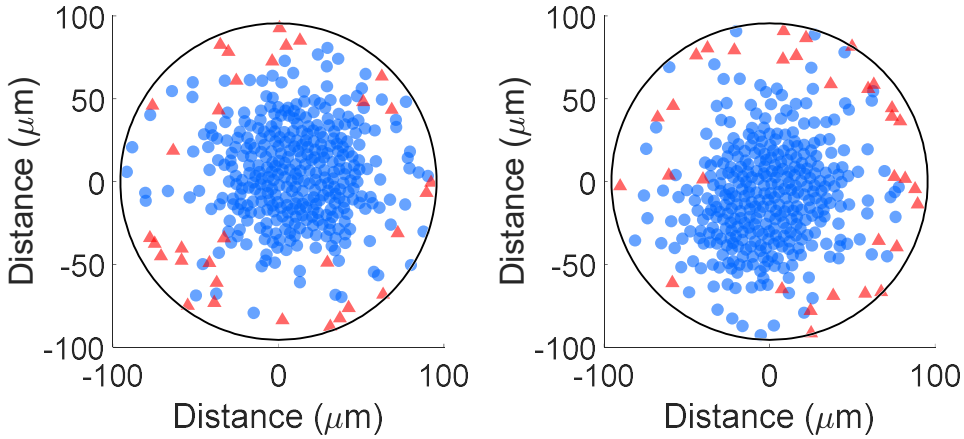
Model 5



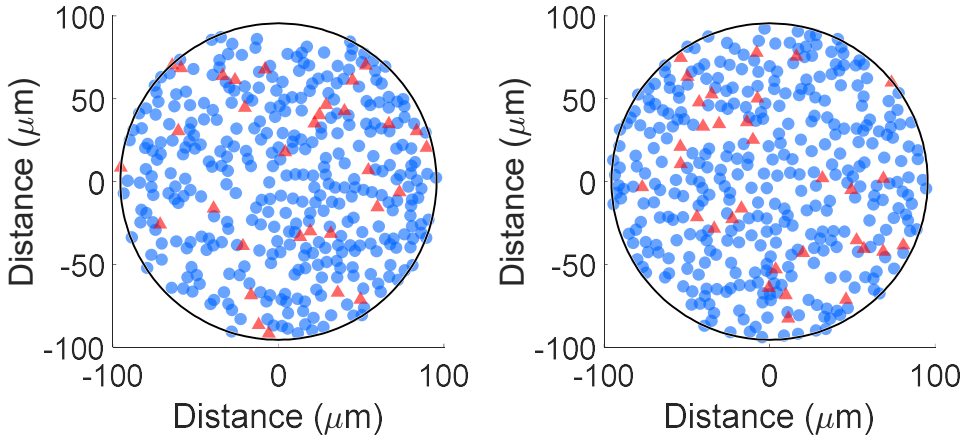
Model 6



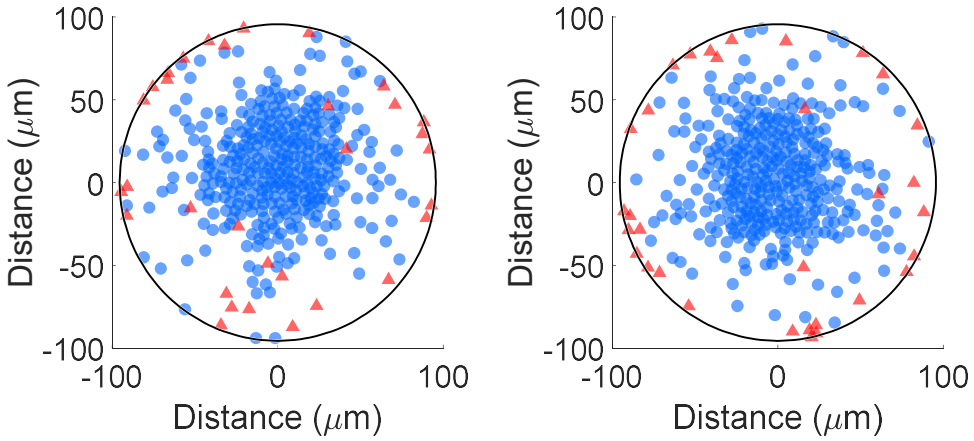
Model 7



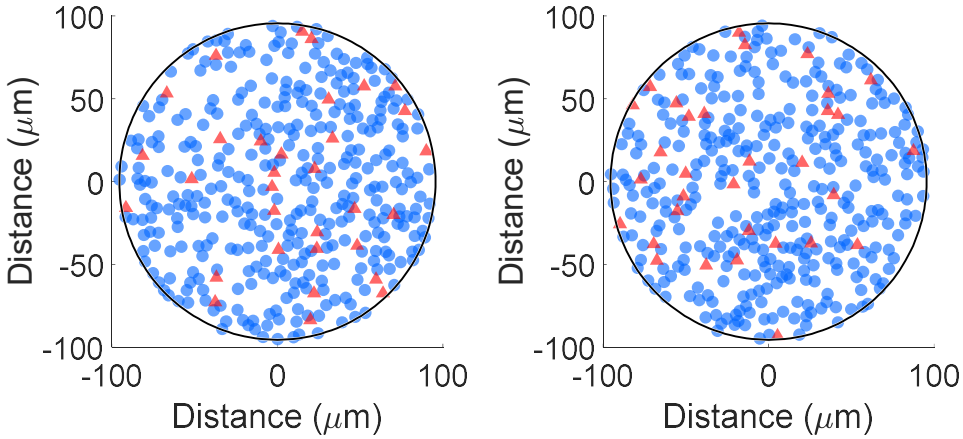
Model 8



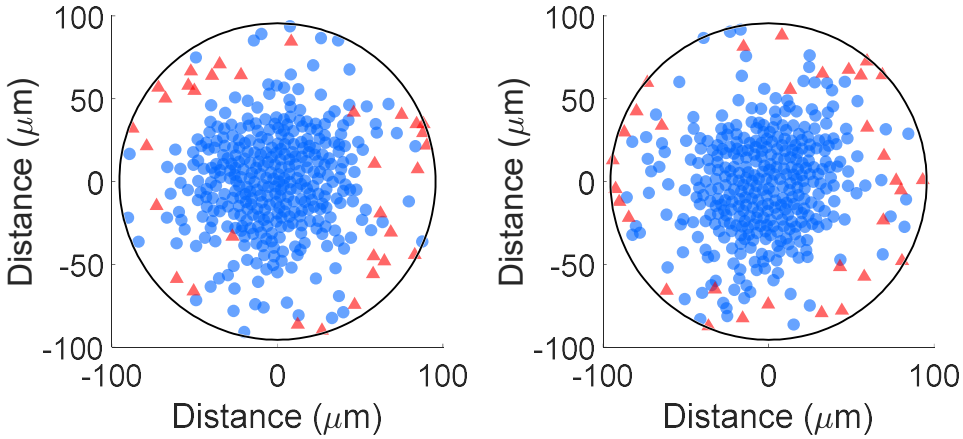
Model 9



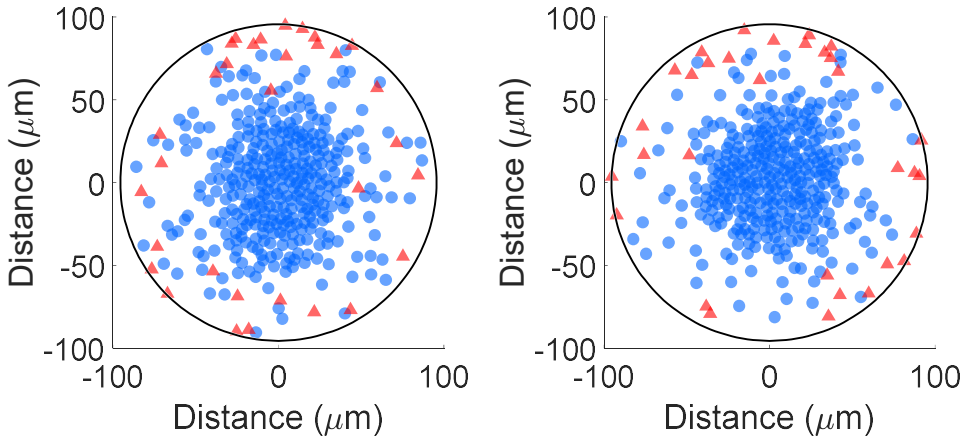
Model 10



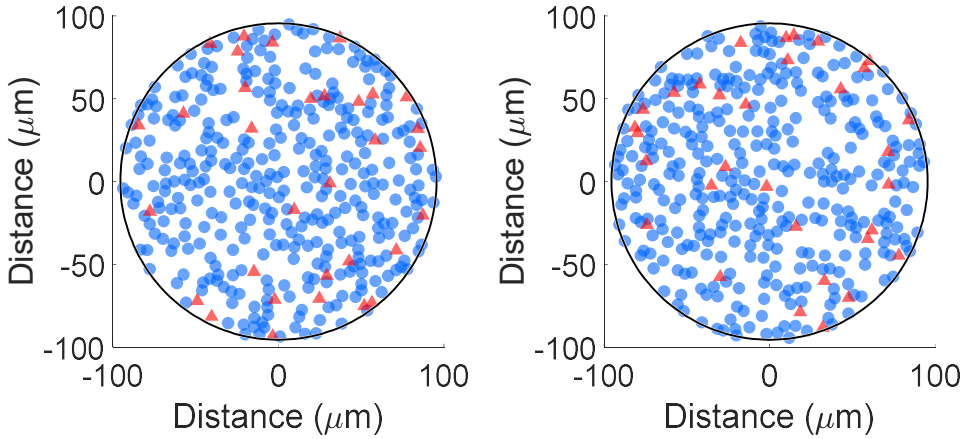
Model 11



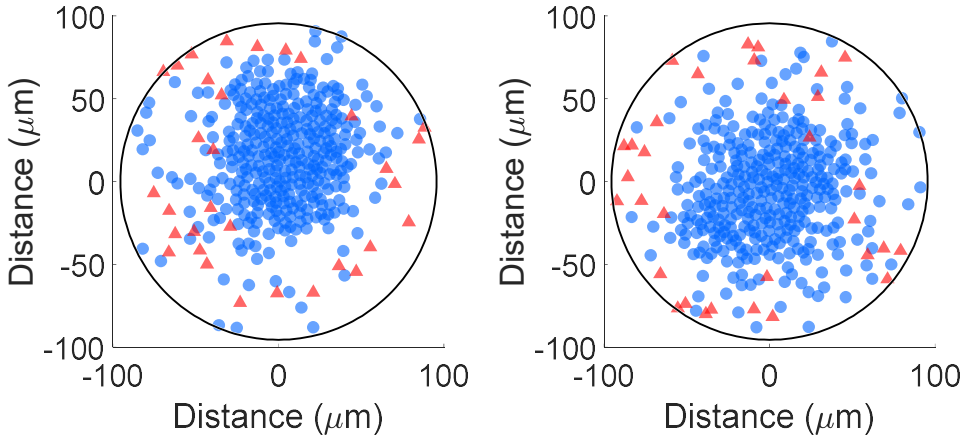
Model 12



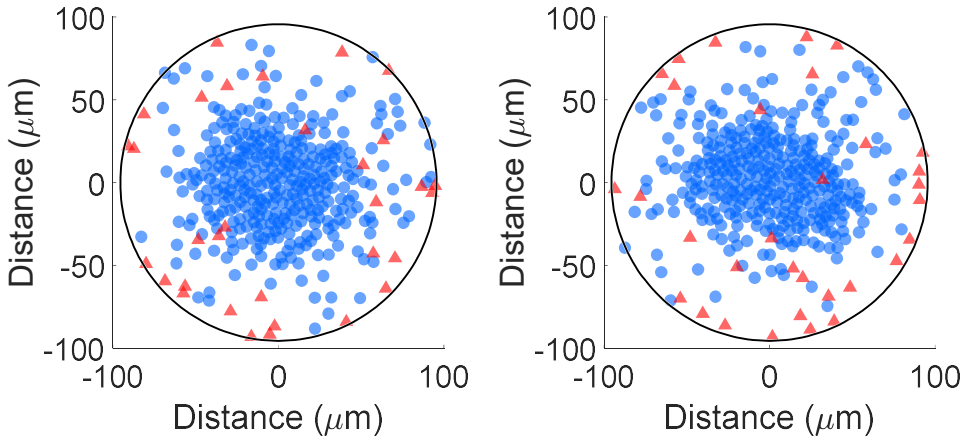
Model 13



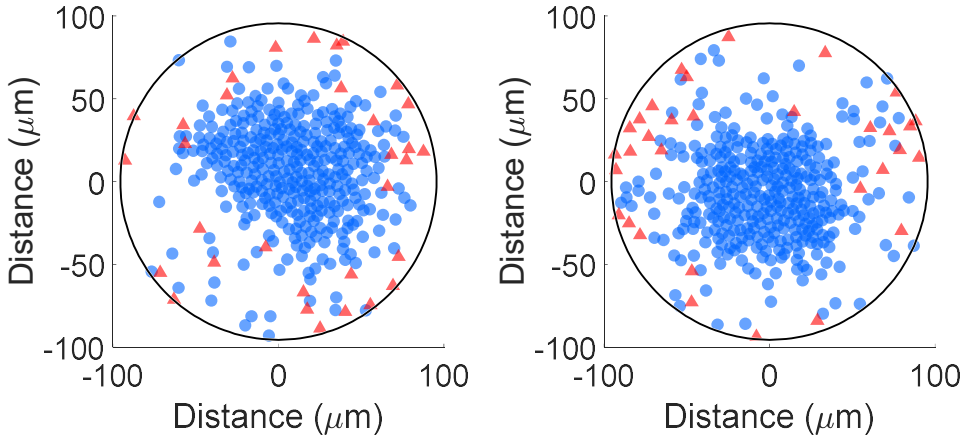
Model 14



Model 15

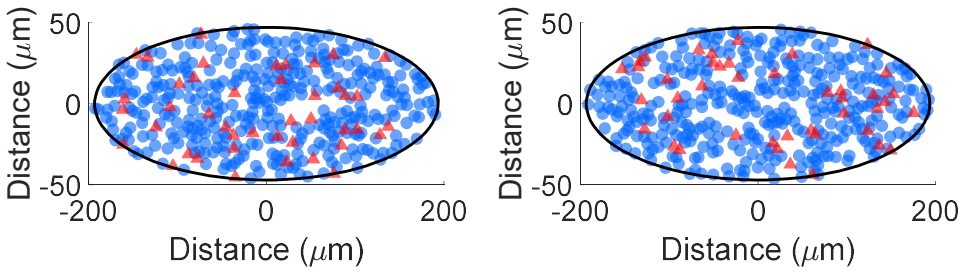


Model 16

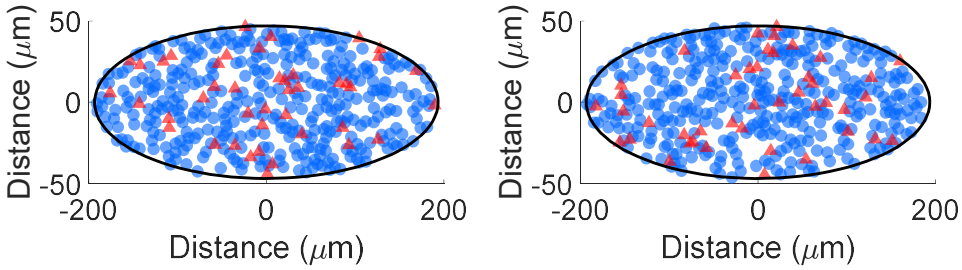


Ellipse experiments

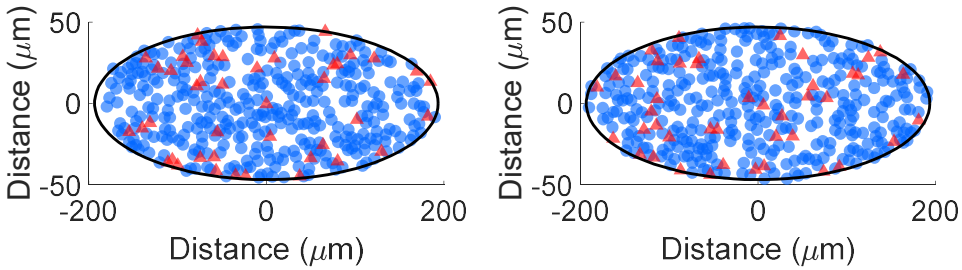
Model 1



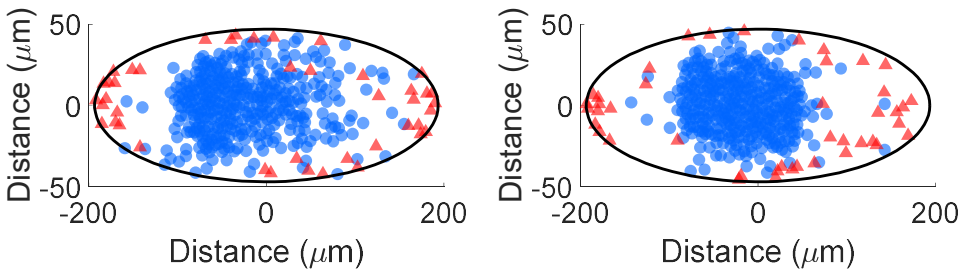
Model 2



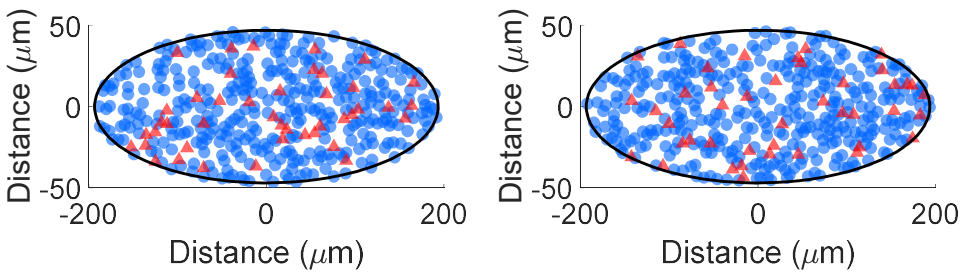
Model 3



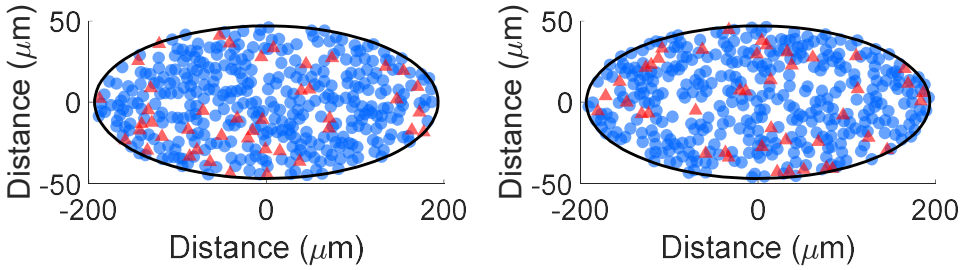
Model 4



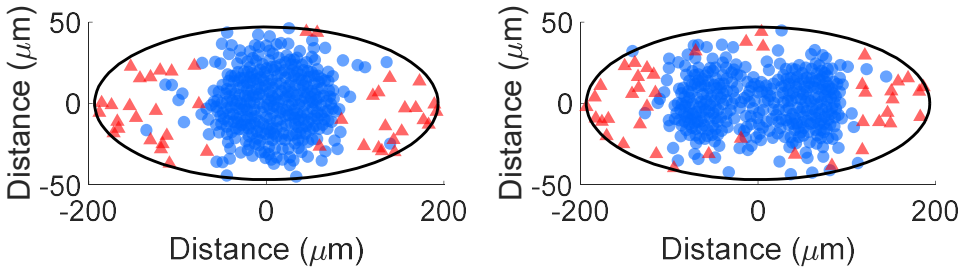
Model 5



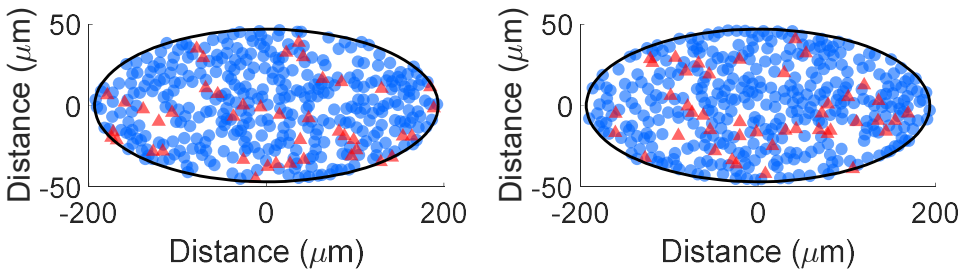
Model 6



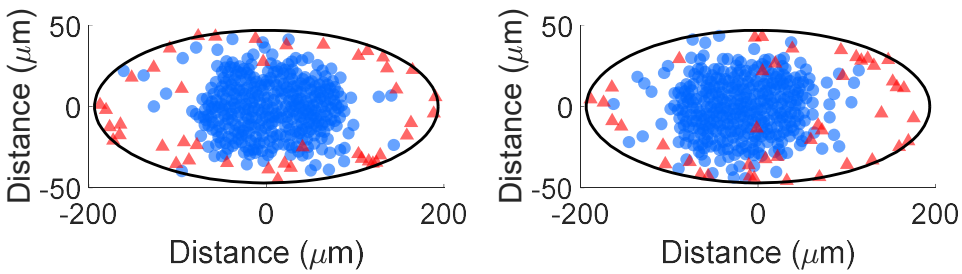
Model 7



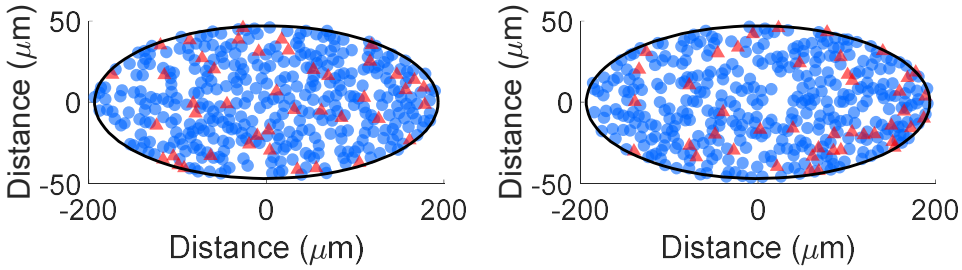
Model 8



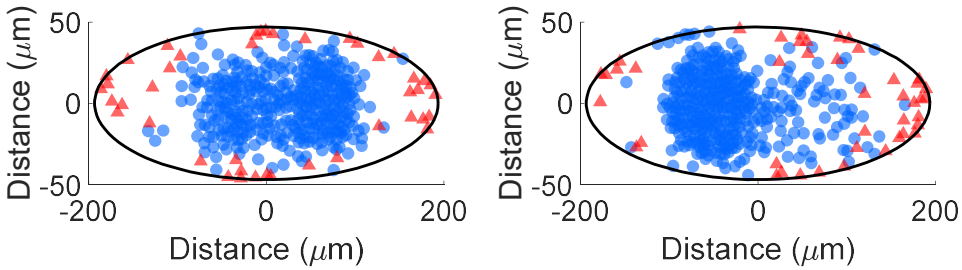
Model 9



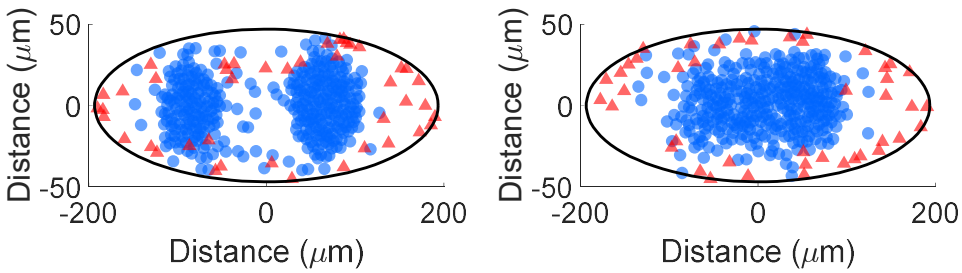
Model 10



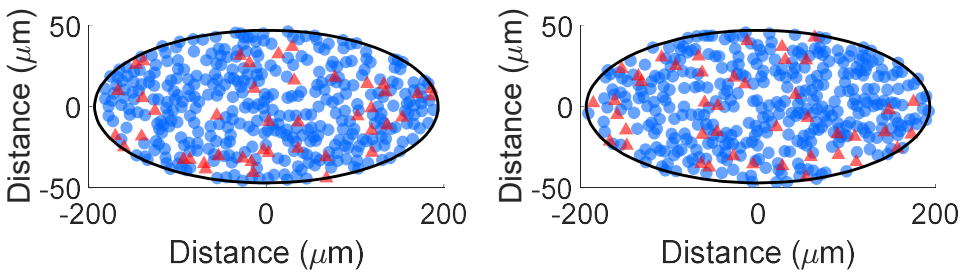
Model 11



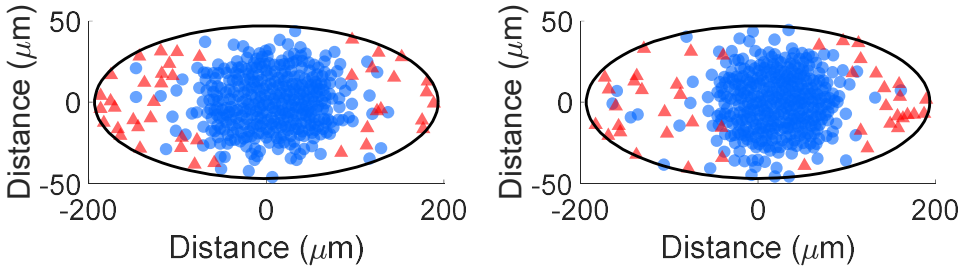
Model 12



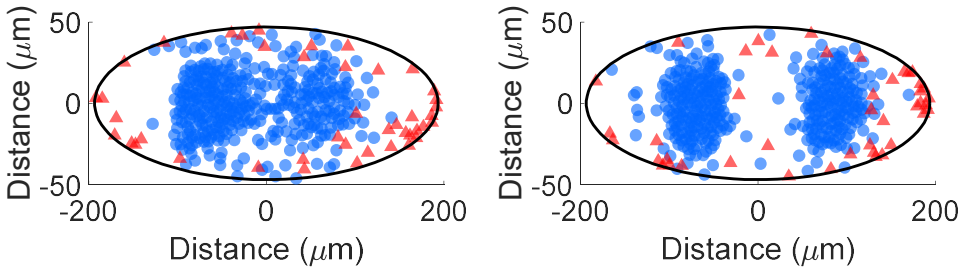
Model 13



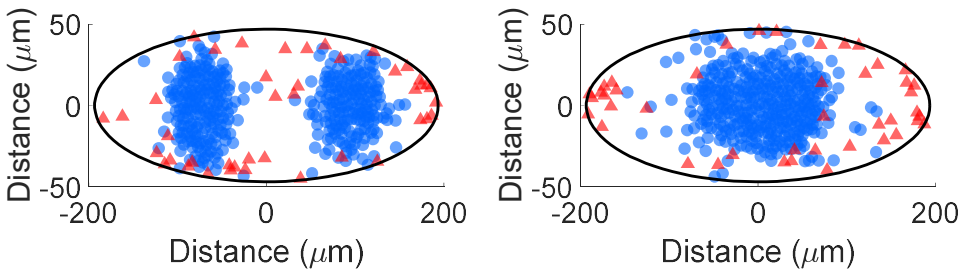
Model 14



Model 15



Model 16

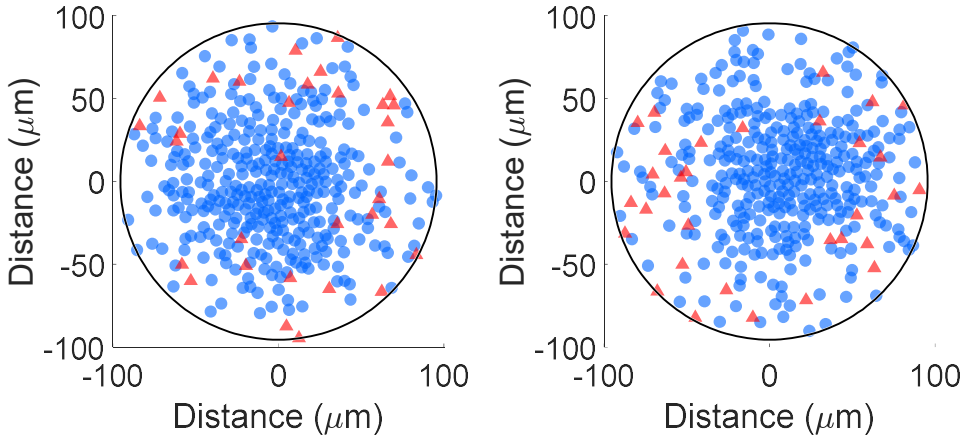


Appendix D: Randomly selected samples from best performing models

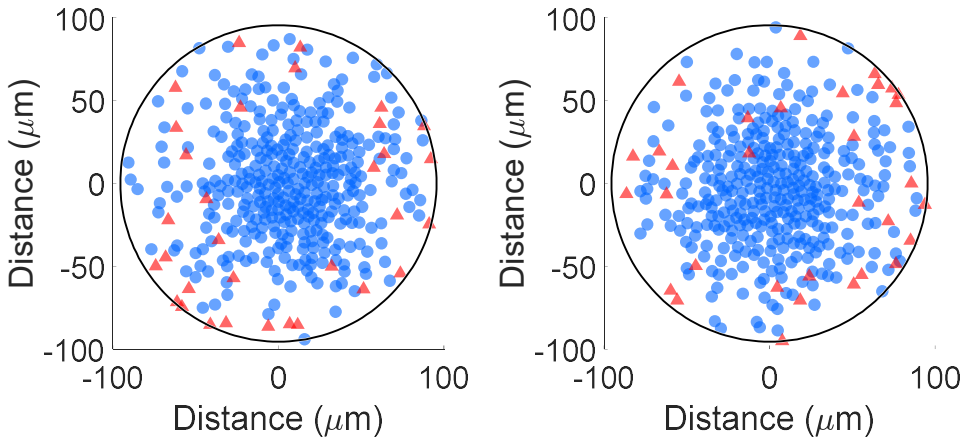
Two indicative randomly selected output examples from best performing models for both disc and ellipse experiments. Red triangle markers stand for T+ cells; blue circle markers stand for T- cells.

Disc experiments

Model 7

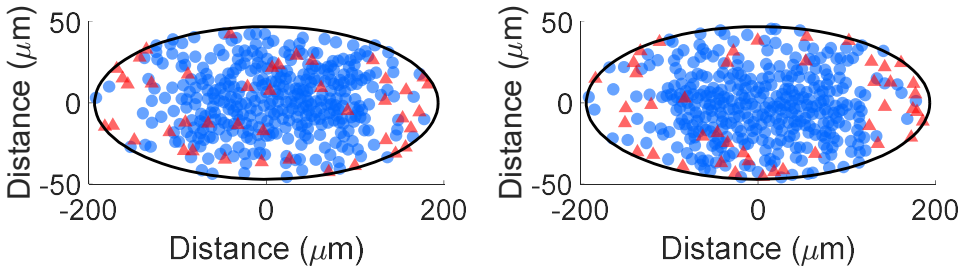


Model 14



Ellipse experiments

Model 7



Model 14