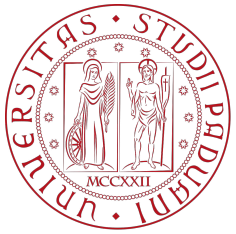
=



University of Padova

Department of Dipartimento di Fisica e Astronomia "Galileo Galilei"

Master Thesis in Physics of Data

Analysis and Modelling of Pluripotent Stem Cells Dynamics

Supervisor
Prof. Samir Suweis
University of Padova

Co-supervisor
Prof. Graziano Martello
Dr. Clelia Corridori

Master Candidate
Mariam Hergnyan

Academic Year
2023-2024

To my family.

Abstract

This thesis explores the dynamics of mouse embryonic stem cells (ESCs) during differentiation, using a combination of experimental data analysis and stochastic modeling. We begin by analyzing data from the experiment conducted by Strawbridge et al. [1], which provides key insights into cell population behavior *in vitro*. Essential parameters are inferred from these data to construct different phenomenological stochastic models of the ESC population.

The stochastic models are then implemented using the Gillespie algorithm [2] to simulate cell behavior. Initially, a simple model is constructed to validate the alignment between the simulated lifetime distribution and its analytical solution. This model also reveals that the "no change" rate, where the system remains in its previous state, has minimal impact on overall system behavior.

Building on these findings, we construct both constant rate and time-dependent rate models to describe the system's dynamics. While both models effectively replicate key metrics, such as cell count distribution and first passage times, the time-dependent model excels in capturing the non-monotonic behavior of the mean cell count over time, a characteristic observed in the experimental data.

This study provides a foundational framework for developing more sophisticated models of stem cell behavior, which could eventually guide the creation of targeted therapies for conditions such as degenerative diseases and tissue repair.

Contents

Listing of figures

Listing of tables

Listing of acronyms

PSCs . . . . . Pluripotent Stem Cells

ESCs . . . . . Embrionic Stem Cells

iPScs . . . . . Induced Pluripotent Stem Cells

GFP . . . . . Green Fluorescent Protein

CV . . . . . Coefficient of Variation

IQR . . . . . Interquartile Range

PDF . . . . . Probability Density Function

CI . . . . . Confidence Interval

FPT . . . . . First Passage Time

TS . . . . . Test Static

PBMC . . . . . Population-based Monte Carlo

SDE . . . . . Stochastic Differential Equation

1 Introduction

The field of systems biology has emerged as a powerful and interdisciplinary approach to understanding the complex interactions that govern biological systems [3, 4]. Traditional biology often focuses on studying individual components of a system in isolation, such as specific genes, proteins, or cellular processes. However, this approach can miss the broader picture of how these components interact and function together as a system. Systems biology, by contrast, seeks to integrate and analyze these interactions within biological networks, aiming to understand the emergent properties that arise from the collective behavior of their parts [5].

The interdisciplinary nature of systems biology draws heavily on principles from physics, particularly in modeling dynamics of interacting systems. Physics provides a rich set of tools for describing the complex behavior of biological systems, from nonlinear dynamics to stochastic processes [6, 7]. For example, the Hodgkin-Huxley model, which uses differential equations to describe the initiation and propagation of action potentials in neurons, has been pivotal in understanding the dynamics of biological networks [8]. Similarly, the principles of scale-free networks, first identified in physics, have been applied to biological networks, revealing their robustness and vulnerabilities [9, 10].

At the core of systems biology is the integration of experimental data, often generated from high-throughput technologies such as genomics, proteomics, and metabolomics, with mathematical models and computational simulations [11, 4]. This integration enables the description and prediction of the behavior of complex biological systems, such as gene regulatory networks, metabolic pathways, and signaling cascades. By constructing these models, systems biol-

ogy aims to identify the fundamental principles governing the dynamics of biological systems, enabling predictions of system behavior under various conditions.

Physicists have significantly advanced biological research by applying their expertise in quantitative analysis, theoretical modeling, and computational techniques. These contributions extend beyond theoretical insights to include the development of new experimental tools, such as advanced microscopy techniques, which have revolutionized our ability to observe and manipulate biological systems at the molecular level [12]. Moreover, physicists have been instrumental in the development and analysis of high-throughput technologies that enable the study of biological systems on an unprecedented scale [13].

Statistical mechanics, a branch of physics traditionally used to describe the behavior of systems with a large number of interacting particles, has become indispensable in the modeling and analysis of biological systems [14]. It provides a framework to model the stochastic nature of processes such as gene expression and protein dynamics, where random fluctuations, or "noise," play a significant role in biological regulation and function [15, 16]. These stochastic models help researchers understand how variability within a population of cells or molecules can lead to different outcomes, such as cell differentiation [17].

Moreover, statistical mechanics allows for the application of probability distributions to describe the variability within cell populations or molecular systems. In gene regulatory networks, for instance, these distributions can reveal steady states, bistability, and other emergent phenomena [18]. Statistical mechanics has also been crucial in understanding phase transitions in biological systems, analogous to transitions between different states of matter, such as the differentiation of stem cells in response to molecular signals [19, 20].

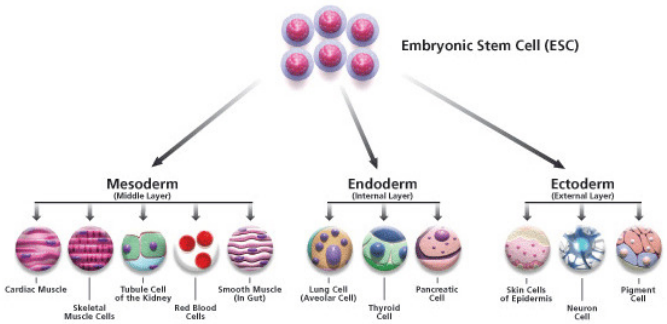


Figure 1.1: Illustration of stem cell differentiation into the ectoderm, mesoderm, and endoderm germ layers during embryonic development, showing the structures of the organism they form.

In this context, the study of pluripotent stem cells (PSCs) presents intriguing challenges and opportunities. PSCs, which include embryonic stem cells (ESCs) and induced pluripotent

stem cells (iPSCs), are unique in their capacity for self-renewal and their ability to differentiate into various specialized cell types [21]. This dual capability makes them central to growth, development, and repair mechanisms in organisms. Upon differentiation, PSCs give rise to cells from the three primary germ layers: the ectoderm, mesoderm, and endoderm(Figure 1.1). The ectoderm forms neural tissue and skin cells, the mesoderm gives rise to muscle, blood, and skeletal tissues, and the endoderm develops into the lining of the digestive and respiratory tracts, as well as organs like the liver and pancreas [21].

The differentiation process of ESCs involves transitions through various states of potential and commitment. Initially, ESCs exist in a *naïve state*, characterized by a high degree of pluripotency [22]. As they begin to differentiate, they enter an *intermediate reversible state*, where cells start to commit to specific lineages, but can still revert to the naïve state under certain conditions [23]. A lineage is defined as the sequential history of cell divisions originating from a single progenitor cell, tracing its descendants through multiple generations. Eventually, ESCs transition into a *differentiated state*, where they commit to a specific cell fate and acquire distinct functional properties [24] (see Figure 1.2).

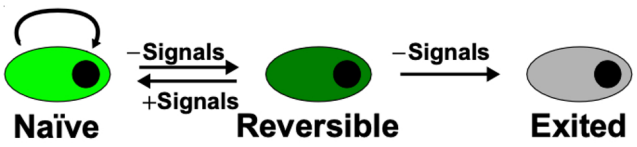


Figure 1.2: The transition of PSC from a naïve state into a differentiated state with an intermediate reversible phase.

Understanding these transitions is crucial for both developmental biology and regenerative medicine. From a computational and physical sciences perspective, studying PSCs offers valuable insights into the complex network of transcription factors that govern pluripotency and differentiation [25, 24]. These networks can be modeled as computational systems, processing biochemical signals and making biological decisions, such as self-renewal or differentiation [26].

In this work, we investigate the dynamics of mouse embryonic stem cells (ESCs) during differentiation using experimental data as our basis. Our goal is to analyze these data to study the behavior of ESCs and develop a stochastic model that simulates the underlying dynamics of the cell populations. By inferring key parameters from the data, our aim is to construct

a comprehensive yet minimal model that captures the variability and stability of ESC behavior, providing insights that complement those obtained through *in vitro* experiments alone [27, 28, 29].

The thesis is structured as follows: In Chapter 2, we analyze the data obtained from the experiments and infer key parameters essential for constructing the physical model. Chapter 3 describes the development of a stochastic model using the Gillespie algorithm, starting with a simple model to validate the alignment between the simulated and analytical lifetime distributions. Chapter 4 presents the construction of constant rate and time-dependent rate models, built to represent the actual data from the experiments, and compares their effectiveness in reproducing key quantities such as cell count distribution and first passage times, with particular emphasis on the time-dependent model's ability to capture the non-monotonic behavior observed in the experimental data. Finally, Chapter 5 discusses the implications of our findings, highlighting the potential for further refinement of these models to advance our understanding of stem cell behavior and its applications in regenerative medicine.

2 Methods

2.1 Experiments

To investigate the dynamics of mouse ESCs during differentiation, an experiment was conducted by Strawbridge et al. [1] with the objective of observing and quantifying key aspects of this process under controlled conditions. The primary focus of this experiment was to analyze the transition of ESCs from a naïve pluripotent state to a formative pluripotent state, focusing on the temporal dynamics and physical parameters associated with this transition.

This was achieved using live-cell imaging combined with single-cell tracking to monitor the behavior of individual ESCs in real-time. The differentiation process of ESCs was monitored using a Rex1-GFPd2 transgenic cell line, which produces green fluorescent protein (GFP) in pluripotent cells. This allowed researchers to track the maintenance or loss of pluripotency. A decrease in GFP fluorescence intensity below a specific threshold indicated that the cells were differentiating and losing their pluripotent characteristics, while sustained high fluorescence indicated that the cells remained in a pluripotent state [30] (Fig. 2.1, 2.2). The experiment also aimed to observe and measure several critical parameters, including cell division patterns, timing of differentiation, cell size, and motility.

Initially, all stem cells were cultured in a 2i medium environment, known to maintain the cells in a homogeneous and stable naïve pluripotent state by inhibiting key signaling pathways that would otherwise prompt differentiation [31]. The cells were seeded on laminin-coated

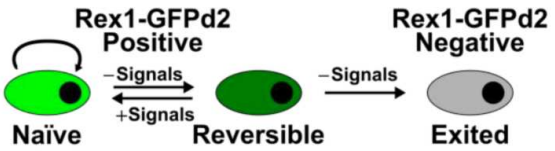


Figure 2.1: The transition of PSC from a naïve state into a differentiated state with an intermediate reversible phase characterized by the loss of the GFP fluorescence intensity.

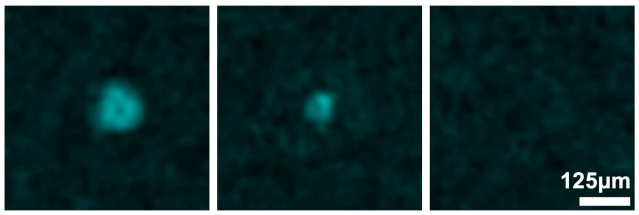


Figure 2.2: Reduction in cell visibility due to differentiation and the subsequent decline in GFP fluorescence.

surfaces to provide a supportive extracellular matrix conducive to cell adhesion and growth [32].

After an initial 12-hour period in the 2i medium, the cells were divided into two groups. One group was transferred to an N2B27 medium, a chemically defined environment designed to promote differentiation by removing the inhibitory signals that maintain naïve pluripotency [33]. The other group remained in the 2i medium as a control to ensure a direct comparison between cells induced to differentiate and those maintained in the naïve state.

Within the control group, further division was made into a **Positive Control** group and a **Negative Control** group. The **Positive Control** group remained in the 2i medium and included the Rex1-GFPd2 transgenic cell line, ensuring that these cells maintained high fluorescence intensity throughout the experiment. This consistent high fluorescence is indicative of the cells retaining their naïve pluripotent state. In contrast, the **Negative Control** group, also kept in the 2i medium, did not contain the Rex1-GFPd2 transgenic cell line. As a result, these cells exhibited low to no fluorescence intensity, serving as a baseline comparison to confirm the specific fluorescence associated with pluripotency in the **Positive Control** group.

To ensure the reliability and reproducibility of the results, the experiment was performed twice, creating two biological replicates. Each replicate followed the same protocol.

During the 72-hour differentiation period, individual ESCs were continuously tracked using

high-resolution live-cell imaging every 30 minutes. The imaging system captured confocal z-stacks at regular intervals, which were later processed into single images for each time point. This process resulted in a dataset, comprising more than 36,000 data points from more than 2,500 individual cells.

2.1.1 Data

The lineage tracking process was performed semiautomatically using a combination of imaging software and manual segmentation techniques. The output from manual segmentation served as input for the oTracks analysis software bundle, which performed further analysis of cell tracks, shapes, fluorescence intensity, and genealogical relationships.

In our work, we were provided with a dataset in the form of `.txt` files, which were generated from the oTracks output.

Each `.txt` file follows the same consistent structure, with each row representing a single cell at a specific time frame. The columns in each file include the following information:

- **Cell ID**: A unique identifier for each cell, which also contains information about the cell's lineage.
- **Cell Type**: Indicates the status of the cell, which can be one of the following:
  - `"divided"`: The cell is a product of division.
  - `"died"`: The cell has died.
  - `"unknown"`: The cell went out of frame and later returned, but tracking was not possible while it was out of frame.
  - `"lost"`: The fate of the cell is unknown.
  - `"bkgnd"`: This entry models background noise as a cell for simplicity.
- **Time Frame**: The specific time point of the observation, with each frame corresponding to 30 minutes.
- **Time (mins)**: The corresponding time in minutes for the given time frame.
- **Fluorescence Intensity**: The measured abundance of the fluorescence signal, indicating the level of GFP in the cell.
- **X Coordinate**: The X position of the cell in the tracking frame.

- **Y Coordinate**: The Y position of the cell in the tracking frame.
- **Area**: The area of the cell in μm^2.

2.1.2 Data Preprocessing

We aimed to evaluate the characteristics of the data obtained from both experiments. A notable difference was observed between the two experiments: in the first experiment, 646 cells were labeled as `"lost"`, while in the second experiment, there were none, indicating an improvement in the technique. Each frame included a background (`"bkgnd"`) modeled as a cell. Additionally, there were cells marked as `"unknown"`, whose fates could not be determined. As part of the dataset preprocessing, all cells labeled `"lost"`, `"unknown"`, and `"bkgnd"` were removed to ensure a clean and accurate analysis (Fig. 2.3).

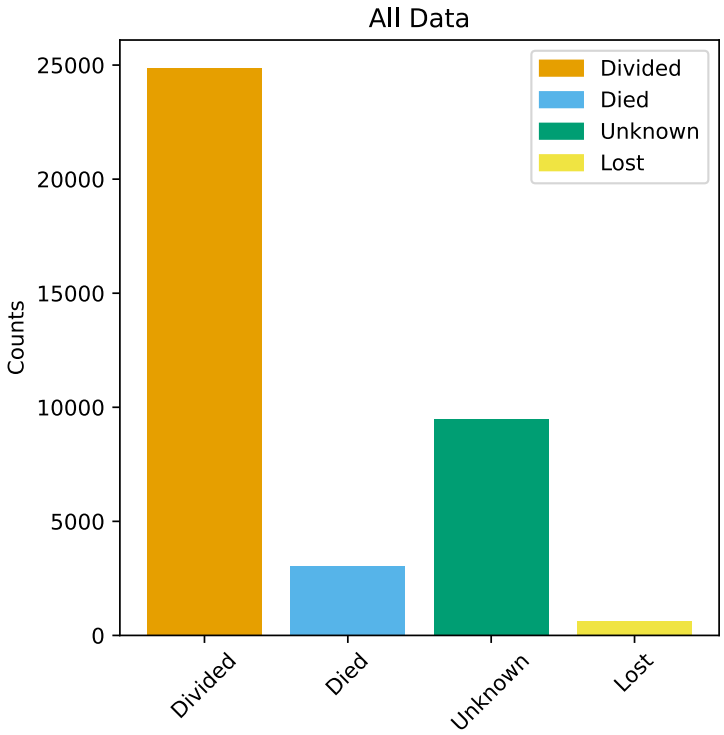
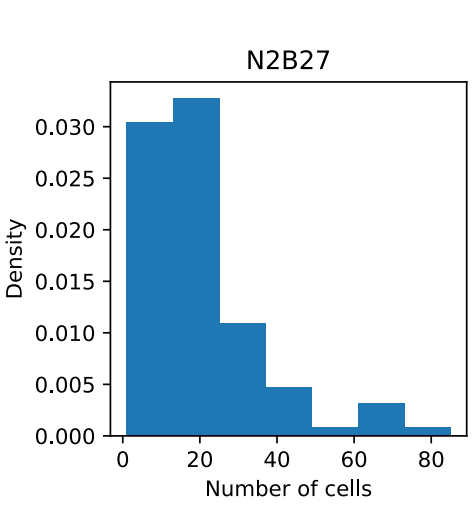
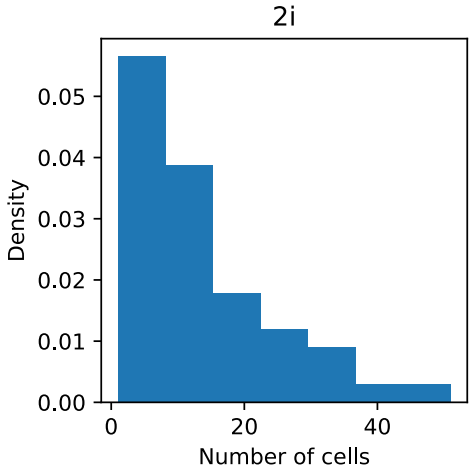


Figure 2.3: Distribution of types of cells in the whole dataset.

During the analysis, any quantities related to time were carefully pre-processed. Specifically, cells present in the first and last frames of the tracking time were disregarded, as we lack complete information about their behaviors before the initiation or after the interruption of the experiment. This step was crucial to ensure the accuracy of the analysis.



(a) Number of Cells in N2B27

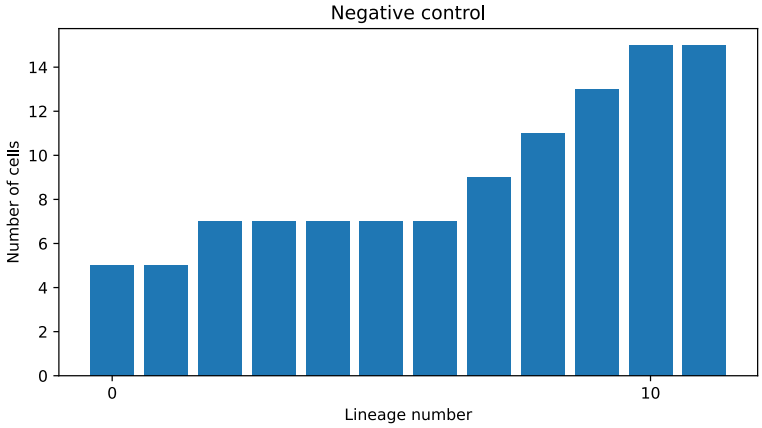


(b) Number of Cells in 2i (**Positive Control** and **Negative Control**)

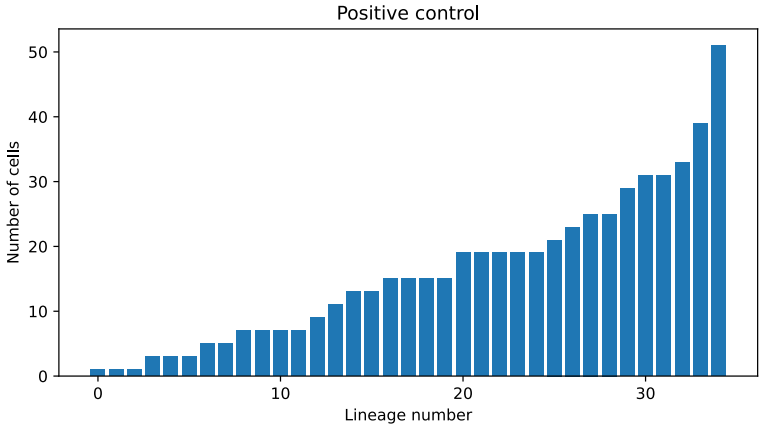
Figure 2.4: PDF of number of cells in each environment.

It is important to note that there was considerably less data available in the 2i environment compared to the N2B27 (Fig. 2.4). Specifically, the **Negative Control** in the 2i environment was utilized to analyze the data obtained from the experiments, but due to the limited sample size, the reliability of the resulting analysis may be compromised. Bootstrapping was attempted to address this issue; however, the results were unsatisfactory, and we ultimately decided not to proceed with this approach. Consequently, while the **Negative Control** is included in the analysis of the data for cells in 2i for completeness, it will be disregarded when building the physical model.

In the N2B27 condition, we initially identified **107** lineages, **33** in the **Positive Control**, and **12** in the **Negative Control**. However, following preprocessing, some lineages were eliminated, leaving **99** lineages for further analysis. This reduction was primarily due to the removal of very short lineages, resulting in empty lineages after preprocessing. It is also worth noting that the number of cells within each lineage after preprocessing can vary considerably (Figures 2.5 and 2.6).

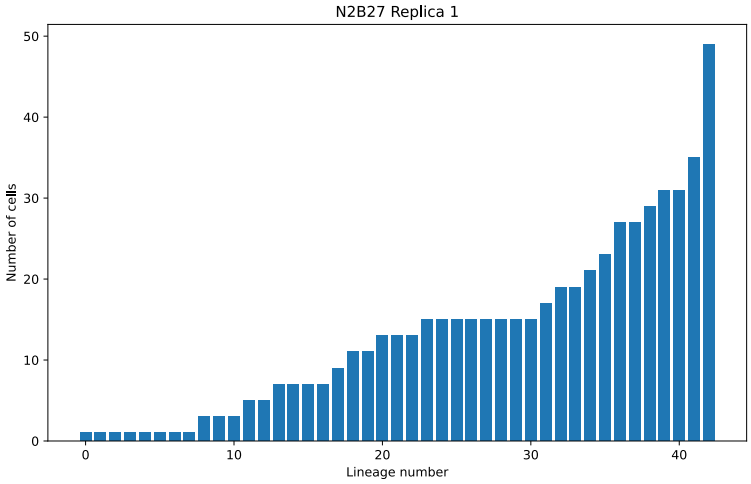


(a) Amount of cells in each lineage for Negative Control

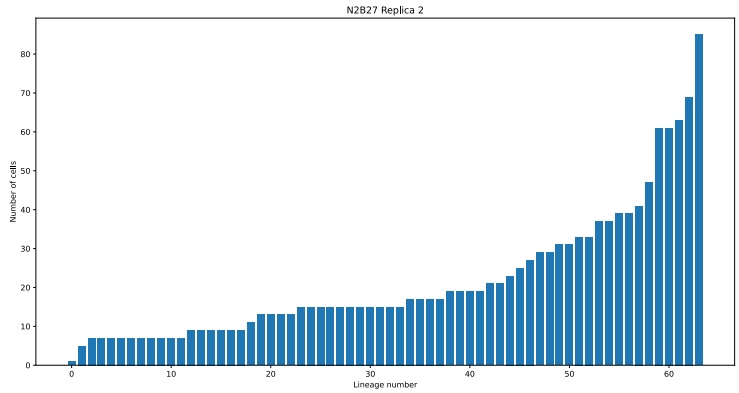


(b) Amount of cells in each lineage for Positive Control

Figure 2.5: Amount of cells in each lineage for Negative and Positive Control environments.



(a) Amount of cells in each lineage for cells N2B27 of the first experiment

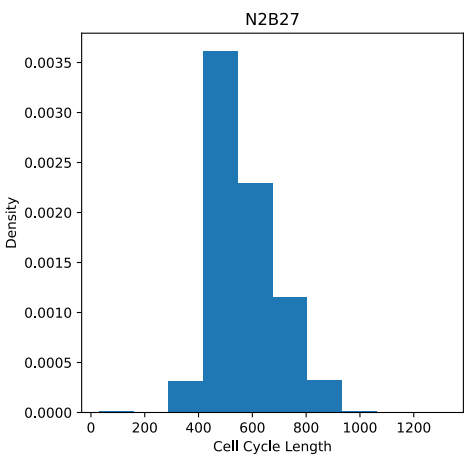


(b) Amount of cells in each lineage for cells N2B27 of the second experiment

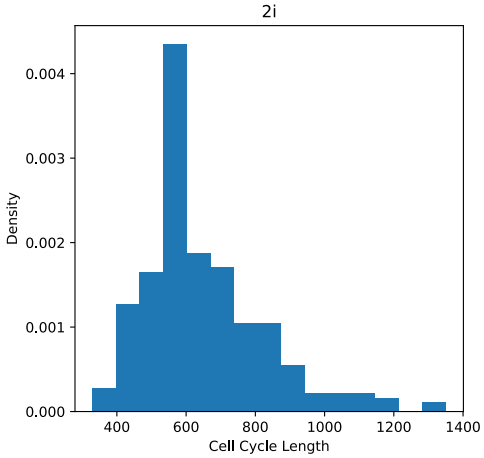
Figure 2.6: Amount of cells in each lineage for N2B27 environments in the 2 experiments.

2.2 Analysis of the Cell Cycle Time

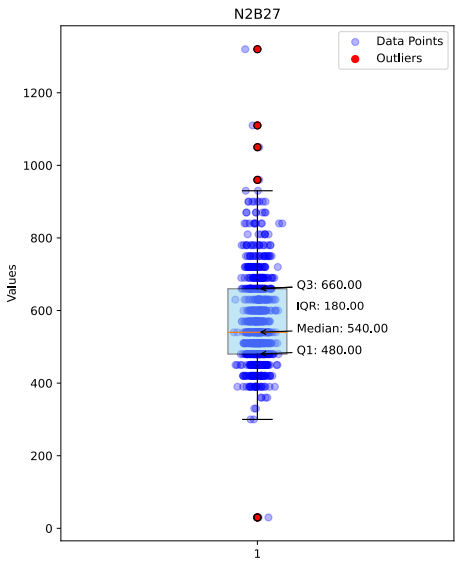
A cell cycle, also known as the cell division cycle or cell duplication time, comprises the series of events that a cell undergoes as it grows and divides into two new cells. The cell undergoing division is referred to as the **mother cell**, while the resulting two new cells are termed **daughter cells** with respect to the mother, and **sister cells** with respect to each other. In our study, we aimed to examine the variation in cell cycle lengths in both 2i and the N2B27 environmental settings and investigate whether these lengths change during the differentiation process.



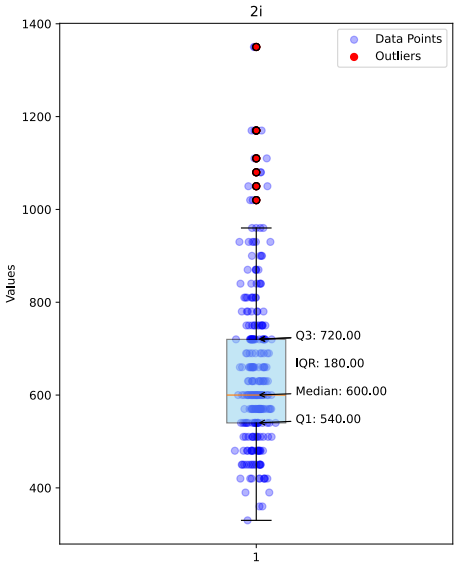
(a) Cell Cycles N2B27



(b) Cell Cycles 2i



(c) Cell Cycles N2B27



(d) Cell Cycles 2i

Figure 2.7: PDFs (2.7a, 2.7b) and boxplots (2.7c, 2.7d) of cell cycles in N2B27 and 2i environments.

The analysis revealed distributions of cell cycle durations in the N2B27 and 2i environments. In both environments, cell cycle durations were centered around 540-600 minutes (see Figure 2.7a, 2.7b). Specifically, the median cell cycle duration was 540 minutes in the N2B27 environment and 600 minutes in the 2i environment (Figure 2.7c, 2.7d). The mean cell cycle time

for cells in the N2B27 environment was 589 minutes, which will later be used as a critical parameter for constructing the physical model of cell cycle dynamics, while it was 652 minutes for cells in 2i.

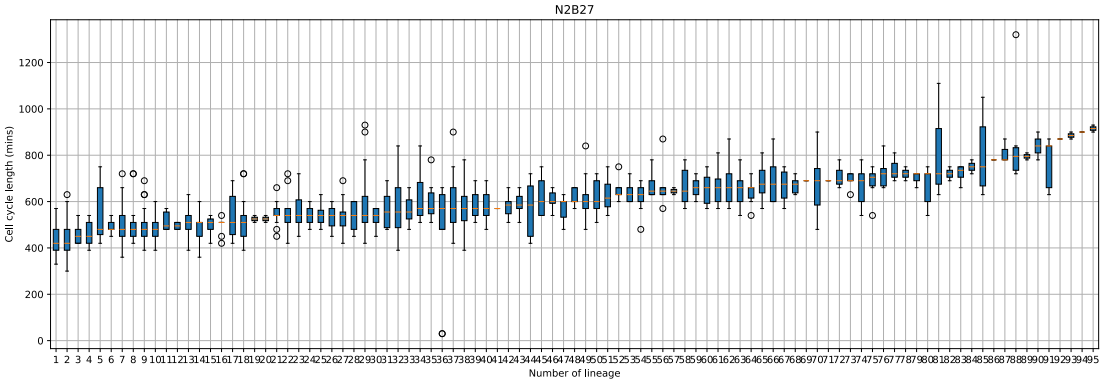
The examination of the 2i environment data indicated a wider spread in the distribution of cell cycle times compared to the N2B27 environment with the interquartile range(IQR) of [530, 650] in 2i and of [500, 580] in N2B27. The standard deviation of cell cycle durations in the 2i environment was 173 minutes, compared to 125 minutes in the N2B27 environment, highlighting greater variability in the 2i environment (Table 2.1). Additionally, a higher relative variability was observed for cells in 2i, as indicated by a significantly higher coefficient of variation(CV) of 0.265 compared to 0.106 in the N2B27 environment.

Metric	N2B27	2i	Positive Control	Negative Control
Median (minutes)	540.0	600.0	630.0	540.0
Mean (minutes)	589.0	652.0	671.4	707.6
Standard Deviation (minutes)	124.78	173.08	140.79	302.18
Coefficient of Variation (CV)	0.106	0.265	0.210	0.427
Interquartile Range (IQR) (25%, 75%)	[500, 580]	[530, 650]	[580, 680]	[440, 640]

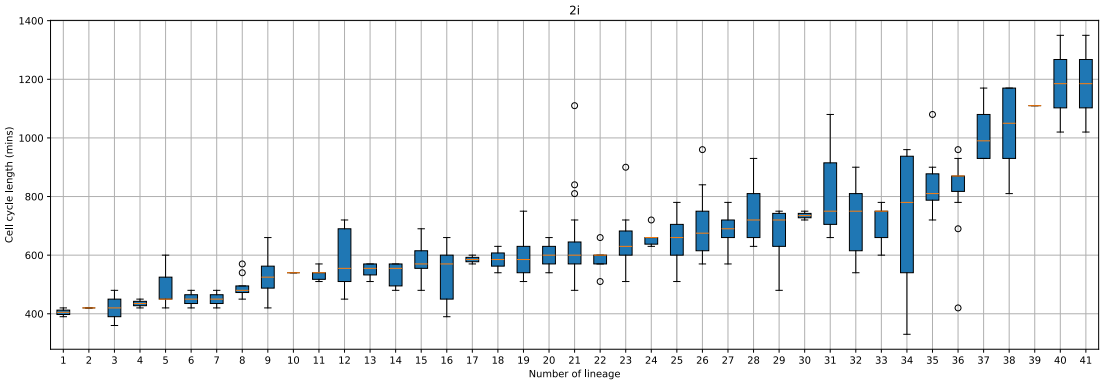
Table 2.1: Summary of key metrics for cell cycle durations in different environments.

Further analysis of the **Positive Control** and **Negative Control** subgroups within the 2i dataset revealed that the increased spread in cell cycle times was largely attributed to the **Negative Control** group, which exhibited a significantly higher standard deviation of 302 minutes compared to 141 minutes in the **Positive Control**. This high standard deviation in the **Negative Control** group is further reflected in its CV of 0.427, indicating that the substantial variability in cell cycle times is both significant in absolute terms and relative to the overall average cycle length within this group. In contrast, the **Positive Control** group had a lower CV of 0.210, suggesting that while there is variability in cell cycle times, it is more consistent and less pronounced relative to the average cycle length in this subgroup compared to the **Negative Control**.

An examination of cell cycle durations across individual lineages (Figure 2.8) revealed distinct patterns in variability and consistency between the N2B27 and 2i environments. The N2B27 environment showed more consistent cell cycle durations, with a standard deviation of 64.77 minutes and a CV of 0.218, indicating less variability within individual lineages(Table 2.2). The IQR values further support this, suggesting a more concentrated distribution of cell cycle lengths within this environment.



(a) Cell Cycle Durations in Each Lineage in N2B27



(b) Cell Cycle Durations in Each Lineage in 2i

Figure 2.8: Boxplots of Cell Cycle Durations in Each Lineage Sorted by the Ascending Order of the Median.

In contrast, the 2i environment having a higher CV and standard deviation, reflected a wider spread in cell cycle durations across lineages. However, the difference in the IQR between the two environments is comparable, with the N2B27 environment having an IQR of [500, 680] minutes and the 2i environment having an IQR of [530, 710] minutes.

Further analysis of the **Positive Control** and **Negative Control** subgroups within the 2i dataset revealed that the increased variability in the 2i environment is particularly pronounced in the **Negative Control** subgroup. The **Negative Control** subgroup showed a higher CV of 0.438 and an IQR of [440, 920] minutes, indicating a substantial range in cell cycle lengths. The **Positive Control** subgroup, although more variable than N2B27 with a CV of 0.218, showed less variability than the **Negative Control**, with a standard deviation of 76.51 minutes and an IQR of [580, 755] minutes.

Metric	N2B27	2i	Positive Control	Negative Control
Standard Deviation (minutes)	64.77	74.34	76.51	69.09
Coefficient of Variation (CV)	0.218	0.265	0.218	0.438
Interquartile Range (IQR) (25%, 75%)	[500, 680]	[530, 710]	[580, 755]	[440, 920]

Table 2.2: Summary of key metrics for cell cycle durations across lineages in different environments.

These findings suggest that while the N2B27 environment supports more uniform and stable cell cycle dynamics within lineages, the 2i environment—especially under **Negative Control** conditions—introduces significant variability, potentially reflecting a more diverse range of cellular responses to the environment.

In conclusion, the N2B27 environment fosters more stable and homogeneous cell cycle durations across lineages, whereas the 2i environment, particularly under **Negative Control** conditions, results in higher variability and a broader range of cellular responses.

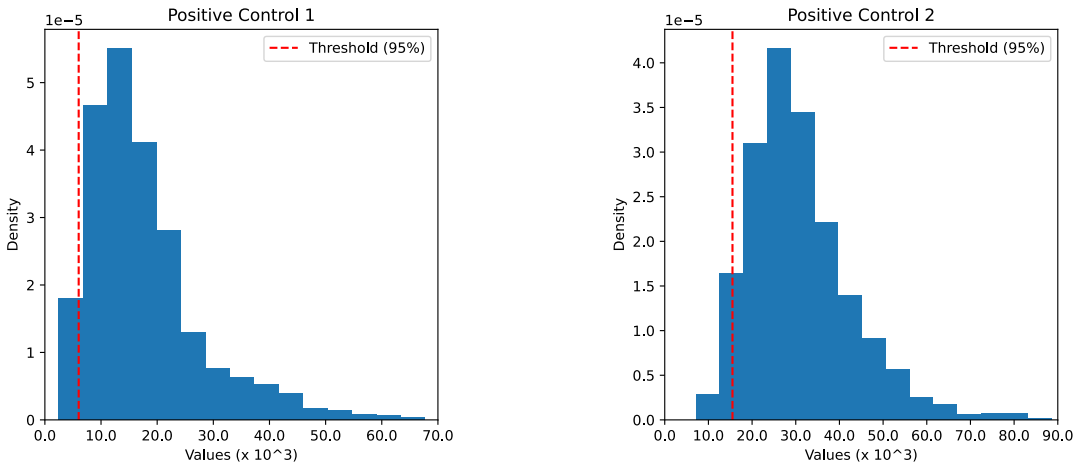
2.3 Analysis of the Differentiation Process

Our primary focus was to analyze the differentiation process of cells. For this, we have developed an experimental setup capable of identifying the naïve state of the cells. Since all the cells in N2B27 have the Rex1-GFPd2 transgenic reporter, we can monitor the differentiation process by measuring the fluorescence value of GFP. As mentioned before, cells in the naïve pluripotent state typically exhibit high GFP fluorescence levels, which gradually decrease as they start to differentiate. To identify when a cell transitions beyond the naïve state, we need to establish a threshold fluorescence value, beyond which a cell is considered to have exited the naïve state and begun differentiation.

To accurately determine the threshold, we used the data from the **Positive Control** group, since it was maintained in a 2i environment, ensuring that these cells remain pluripotent throughout the experiment. To establish the threshold, we calculated the 95% lower confidence interval (CI) of the total fluorescence intensity for the **Positive Control** datasets. This calculation was performed separately for each experiment to account for any variations between the experimental runs (Fig.2.9). In the first experiment, the lower 95% CI for fluorescence was calculated to be 6021.28 and in the second experiment it was determined to be 15511.94 (Table 2.3).

These thresholds, specific to each experiment, were then applied to the N2B27 datasets. Any cell with a fluorescence intensity below the respective threshold was considered differentiated.

Upon examination of the fluorescence profiles of the cells, we observed fluctuations in the



(a) Positive Control 1

(b) Positive Control 2

Figure 2.9: PDF of GFP fluorescence intensity values for Positive Control groups in Experiments 1 and 2. The plots show the distribution of fluorescence intensity values in the Positive Control group of Experiment 1 (a) and Experiment 2 (b). The red dashed line represents the threshold, defined as the lower bound of the 95% CI

Data	Maximum Fluorescence Intensity	Lower 95% CI
Positive Control 1	67714.81	6021.28
Positive Control 2	88690.92	15511.94

Table 2.3: Maximum fluorescence intensity and 95% lower CI for Positive Control 1 and Positive Control 2.

fluorescence levels (an example can be seen in Figure 2.10). In particular, for cells undergoing differentiation, these fluctuations can pose challenges in accurately determining when the fluorescence falls below the threshold. We noticed that selecting the first point (first negative) where the fluorescence dips below the threshold could lead to false identification of a cell as differentiated, as subsequent fluctuations may bring the fluorescence level back up. To minimize the number of false positive differentiated cells, we selected the last positive point, which is the point below the threshold after which there are no subsequent time points where the fluorescence exceeds the threshold value (Figure 2.10).

Taking into account this observation, we performed an analysis to identify the stage of the cell cycle at which differentiation happens. The results, illustrated in Figure 2.11, indicate that differentiation predominantly occurs in close proximity to the cell birth phase considering that the mean cell cycle time is 589 minutes while the average differentiation time 272 minutes.

Furthermore, we calculated the time at which differentiated cells first appear within each

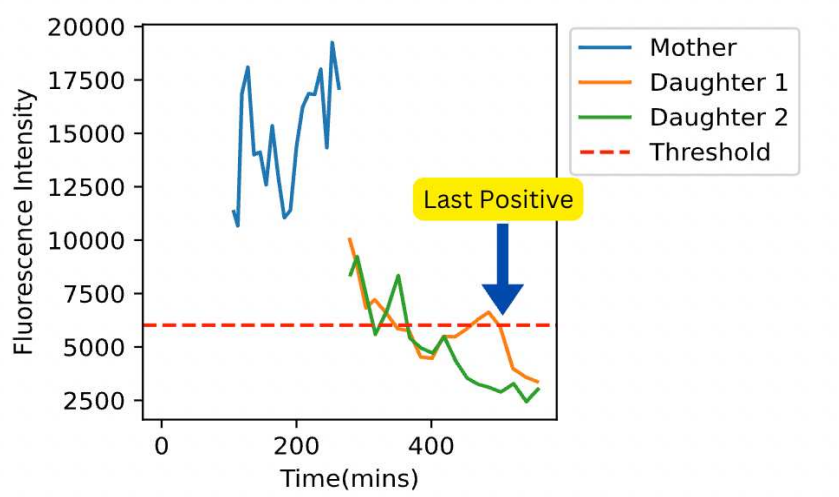


Figure 2.10: The example of fluorescence intensity profile of a mother cell undergoing division into two daughter cells which are differentiatin with an illustration of the last positive point of the Daughter 1.

lineage. The analysis revealed that most lineages begin to exhibit differentiation around the second or third cell cycle time, i.e. the second or third generation of cells (Figure 2.12).

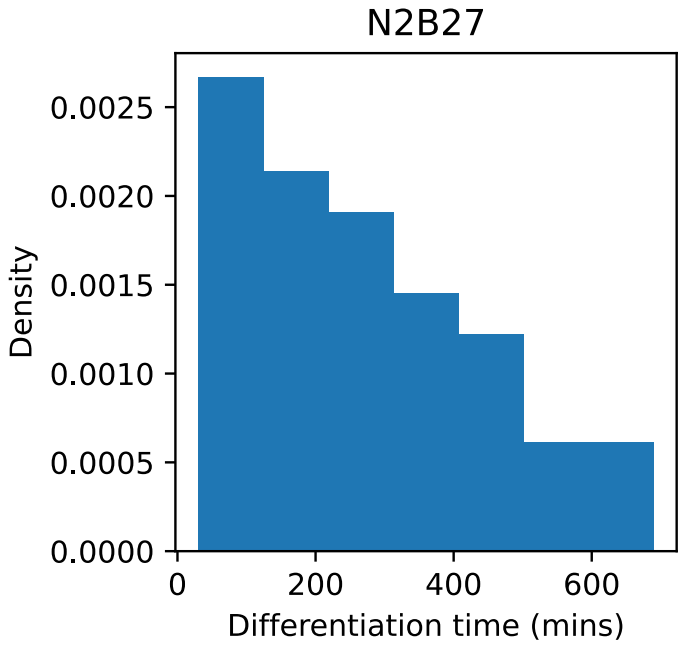


Figure 2.11: The PDF of the time that the cells take to differentiate from the moment of being born.

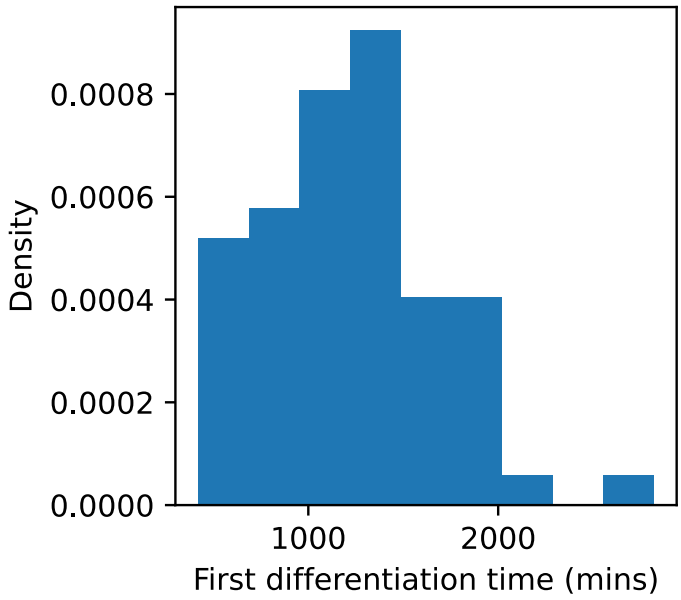


Figure 2.12: PDF of the time at which differentiated cells first appear within lineage.

2.3.1 Symmetry of Divisions

We identified six possible types of cell division, as illustrated in Figure 2.13. These division types are:

- **Symmetric division**: An undifferentiated mother cell divides into two undifferentiated daughter cells.
- **Symmetric differentiation**: An undifferentiated mother cell divides into two differentiated daughter cells.
- **Asymmetric division**: An undifferentiated mother cell divides into one differentiated and one undifferentiated daughter cell.
- **Division with one daughter cell death**: An undifferentiated mother cell divides, resulting in only one undifferentiated daughter cell, with the other cell dying.
- **Division with two daughters cell deaths**: An undifferentiated mother cell divides, resulting in no surviving daughter cells.
- **Differentiated division**: A differentiated mother cell divides into two differentiated daughter cells.

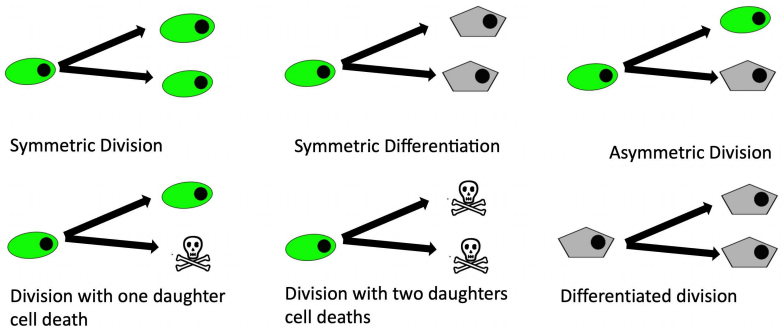


Figure 2.13: Different types of divisions observed in the dataset in the N2B27 environment.

The goal of this part was to quantify the different types of cell divisions observed and to determine which types are the most common. This analysis is crucial for identifying patterns and probabilities of division outcomes, which are essential for understanding cellular behavior and later constructing an accurate physical model. It is generally expected that symmetric divisions and symmetric differentiations are the most common types of divisions in pluripotent stem cells, as these processes are fundamental for maintaining the balance between self-renewal and differentiation [34].

We observed 314 division processes in the data, however, the differentiated division, as well as cases in which one or both daughter cells died shortly after division, were excluded from this part of the analysis due to their predictability and limited scientific significance, leaving us with 229 division processes in which we are interested in. The excluded cases, nevertheless, will later be included in the physical model as they are essential for inferring the model parameters. As shown in Figure 2.14, as expected, our analysis revealed that the majority of divisions or differentiations to be symmetric, specifically 51. 5% were **symmetric divisions** (118), 40.6% were **symmetric differentiations** (93), and 7.8% were **asymmetric divisions** (18).

After quantifying the frequency of each type of division, our objective was to validate our classification by analyzing the correlation of fluorescence intensity between sister cells in three key division types: **symmetric divisions**, **symmetric differentiations**, and **asymmetric divisions**. Initially, our classification was based on whether the fluorescence intensity values of the cells were above or below the specific threshold. Figure 2.15 provides examples of the expected fluorescence behaviors for these type of division. Although this method was effective,

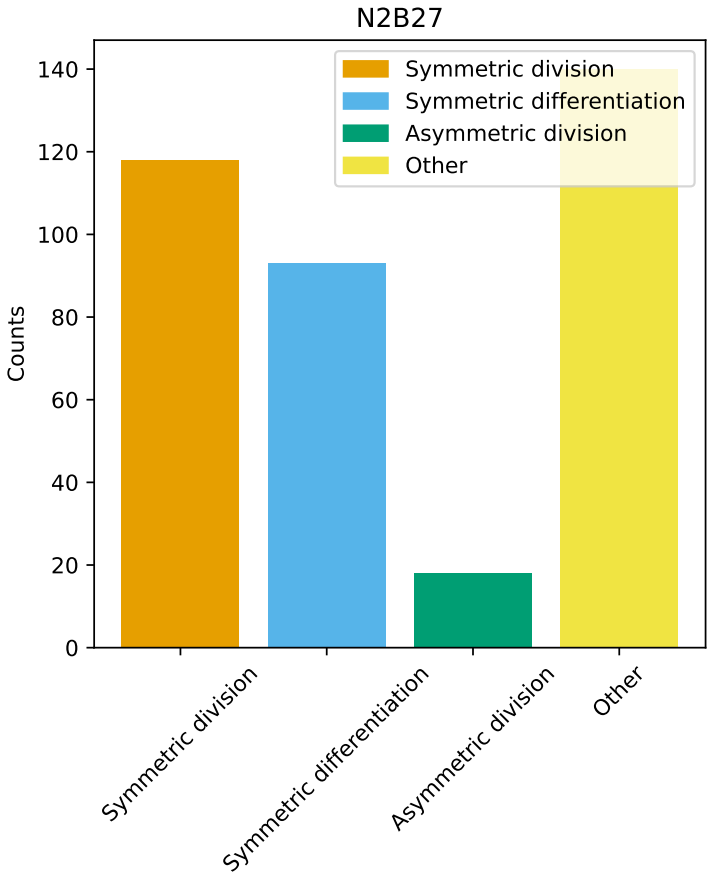
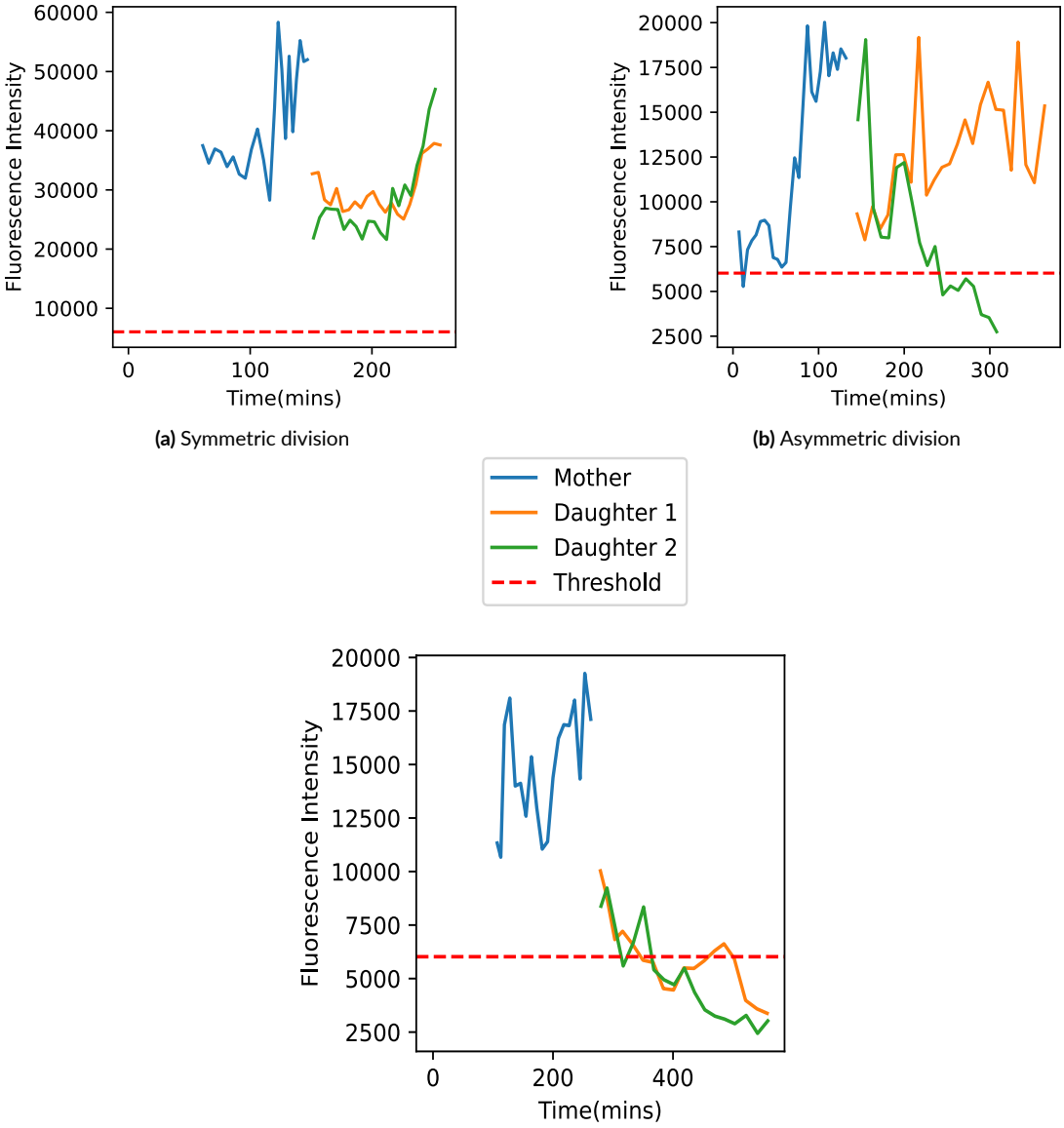


Figure 2.14: Histogram of the different types of divisions in the N2B27 environment.

we sought to ensure that the fluorescence behavior was aligned with the expected outcomes for each division type. Specifically, we expected symmetric divisions and differentiations to exhibit a high correlation in fluorescence intensity between sister cells. In contrast, asymmetric divisions should show divergence in fluorescence intensity, reflecting the distinct fates of the sister cells(Figure 2.16).

To evaluate the strength and significance of these correlations, we employed the Pearson correlation coefficient[35] and the corresponding Pearson p-value tests[36] (see Table. 2.4). The Pearson correlation coefficient is particularly well-suited for measuring the linear relationship between two continuous variables, making it ideal for assessing the relationship between fluorescence intensities in sister cells. A higher Pearson correlation coefficient indicates a strong positive correlation, which was expected for symmetric divisions and differentiations, where sis-



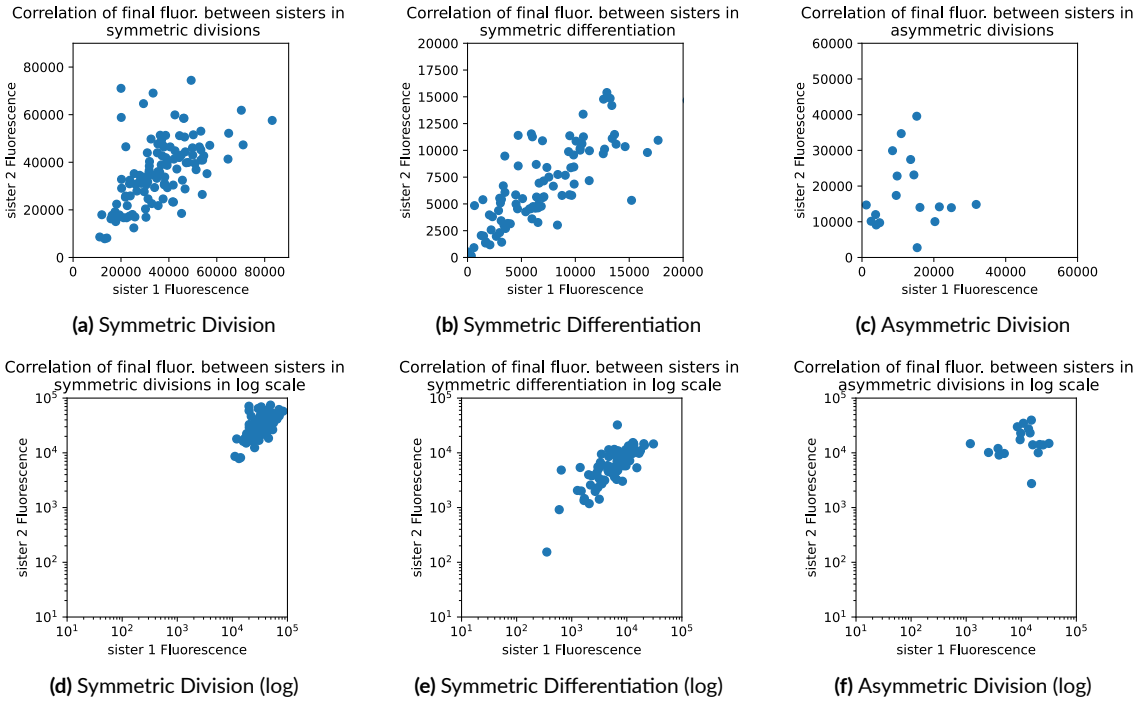
(a) Symmetric division

(b) Asymmetric division

(c) Symmetric differentiation

Figure 2.15: Examples of fluorescence profiles of a mother cell and its daughter cells undergoing a) symmetric division, b) asymmetric division, and c) symmetric differentiation.

ter cells exhibit similar fluorescence intensities. Conversely, a lower correlation was anticipated for asymmetric divisions, where the sister cells' fluorescence intensities diverge. A statistically significant correlation is confirmed by a lower Pearson p-value, suggesting that the observed



(a) Symmetric Division **(b)** Symmetric Differentiation **(c)** Asymmetric Division

(d) Symmetric Division (log) **(e)** Symmetric Differentiation (log) **(f)** Asymmetric Division (log)

Figure 2.16: Correlation of the fluorescence intensity values of the sister cells for symmetric division, symmetric differentiation, and asymmetric division in linear(a, b, c) and logarithmic(d, e, f) scales.

relationship is unlikely to be due to random chance.

In addition, we conducted Ranksum tests[37], also known as the Wilcoxon rank-sum tests, to determine whether there were significant differences in the distribution of fluorescence intensity values between sister cells across the different types of divisions. The Ranksum test, a non-parametric method, compares the distributions of fluorescence intensities between two independent groups. A higher Ranksum test statistic (TS) indicates a greater difference between the distributions, and a lower Ranksum p-value confirms that this difference is statistically significant, i.e. not random.

The results of these analysis are visualized in Figures 2.16a–2.16f and summarized in Table 2.4. The Pearson correlation tests revealed a positive correlation between the fluorescence intensity values of the sister cells in cases of **symmetric divisions** ($\rho = 0.58$) and **symmetric differentiations** ($\rho = 0.63$), with both correlations being highly significant ($P < 0.001$). In contrast, as expected, no significant correlation was observed for **asymmetric divisions** ($\rho = 0.025$, $P = 0.92$), indicating that the fluorescence intensities diverged as anticipated.

The Ranksum test results further confirmed these findings. The Ranksum coefficients and p-values indicated no significant differences in the distribution of fluorescence intensity values

between sister cells for any of the cases, suggesting that within each division type, the fluorescence intensities were consistently distributed.

These findings are consistent with our expectations, confirming that the fluorescence intensity behavior matches the predicted outcomes for symmetric and asymmetric divisions.

Division Type	ρ_{Pearson}	P-value$_{\text{Pearson}}$	Ranksum TS	Ranksum P-value
Symmetric Division	0.58	9.18×10^{-12}	-9.535×10^{-3}	0.99
Symmetric Differentiation	0.63	1.76×10^{-11}	0.11	0.91
Asymmetric Division	0.025	0.92	-1.45	0.15

Table 2.4: The Results of statistical tests carried out to analyze the correlation of the GFP fluorescence intensity of sister cells undergoing different types of divisios.

The lower than expected Ranksum TS values for asymmetric divisions, may be attributed to several factors. One significant factor could be the combination of data from two biological replicates with different fluorescence intensity levels. When these data sets are combined, the fluorescence intensity distributions of the sister cells may overlap, potentially reducing the apparent differences in fluorescence intensity between the undifferentiated and differentiated cells, which leads to lower Ranksum TS values. Although analyzing the replicates separately could help clarify these differences, the low number of asymmetric divisions overall makes such an analysis statistically unreliable. Additionally, other factors such as sample size, threshold definition, and biological noise may also contribute to these findings. Together, these factors suggest that while some difference in fluorescence intensities is expected, it may not be as pronounced when considering the combined data from both replicates.

3

Theoretical Framework

In this chapter, we explore the theoretical foundations necessary for understanding the dynamics of stem cell populations. Biological systems, particularly at the cellular level, are inherently complex and influenced by a multitude of factors. To gain insights into these systems, modeling becomes an indispensable tool. Models allow us to simulate and predict behaviors that are difficult, if not impossible, to observe directly in experiments. However, the choice of model is critical, as it dictates the accuracy and relevance of the predictions made.

Statistical physics provides a powerful framework for studying systems composed of a large number of interacting components, such as cells in a biological environment. It offers tools to describe the macroscopic behavior of a system based on the microscopic rules governing individual components. In biological contexts, where cellular processes are subject to intrinsic randomness, stochastic processes become particularly valuable. Unlike deterministic models, which predict a single outcome from a given initial condition, stochastic models account for randomness in biological systems, providing a distribution of possible outcomes[38].

In the context of stem cell dynamics, stochastic modeling is especially relevant. Stem cells are influenced by a variety of random factors, including fluctuations in gene expression and microenvironmental signals. These factors can lead to significant variability in the behavior of the cell population, which deterministic models may not capture. To accurately simulate the dynamics of a large population of undifferentiated stem cells, we require a model that incorporates this inherent stochasticity.

To address these challenges, we employ the Gillespie algorithm, a well-established method for

simulating the time evolution of systems where randomness plays a crucial role. The Gillespie algorithm[2], also known as the Stochastic Simulation Algorithm (SSA), allows the exact numerical simulation of systems by modeling the sequence and timing of individual events. This approach is particularly suited for our study, where the timing and sequence of cellular events, such as cell division and differentiation, are critical to understanding the overall dynamics of the system.

3.1 Stochastic Modelling of Cell Dynamics

To accurately describe the dynamics of undifferentiated stem cells, we rely on a stochastic approach that captures the inherent randomness of biological processes. Central to this approach is the **Master Equation**, which governs the time evolution of the probability distribution of the system's state. This equation provides a comprehensive framework for understanding how the probabilities of different system states change over time, influenced by the birth and death processes within the cell population.

To model the processes occurring in the system based on experimental data, we start by examining a simplified model that includes two fundamental processes: birth and death. These processes can be described by their transition rates, which define the probability of the system transitioning from one state to another:

- **Birth (Cell Division)**: An undifferentiated stem cell divides, increasing the cell population by one.
- **Death (Cell Death or Differentiation)**: An undifferentiated stem cell either undergoes cell death or differentiates into a specialized cell type, reducing the undifferentiated cell population by one.

The transition rates for these processes are given by the following:

$$T(n \to n+1) = b \cdot n \tag{3.1}$$

$$T(n \to n-1) = d \cdot n \tag{3.2}$$

Here, b represents the birth rate, d represents the death rate, and n denotes the number of undifferentiated cells at a given time.

Incorporating these transition rates into the Master Equation, we obtain:

$$\frac{dP(n,t)}{dt} = b\cdot(n-1)\cdot P(n-1,t) + d\cdot(n+1)\cdot P(n+1,t) - (b\cdot n + d\cdot n)\cdot P(n,t) \quad (3.3)$$

This equation characterizes the time-dependent probability distribution of the number of undifferentiated cells, capturing how the probability $P(n,t)$ of having n undifferentiated cells at time t changes over time due to birth and death processes. The first two terms on the right-hand side represent the probability flow into the state with n cells, while the last term represents the probability flow out of this state.

3.1.1 The Simulation Algorithm: Gillespie

To overcome the challenges of solving the Master Equation analytically, we employ the **Gillespie algorithm**. This algorithm provides a powerful method for generating exact numerical solutions by simulating individual reaction events in a step-by-step manner.

The Gillespie algorithm proceeds through the following steps[39]:

1. **Initialization**: Begin with an initial set of conditions, including the number of particles and the rates at which different reactions or events can occur.

2. **Calculate Propensities**: For each possible event or reaction, calculate its propensity function $a_j(x)$. The propensity is a product of the reaction's rate constant and the number of possible reactive combinations in the system. It represents the likelihood of that event occurring in a small time interval dt.

3. **Determine the Time to Next Reaction**: The time τ until the next reaction occurs is drawn from an exponential distribution, where the mean of the distribution is the inverse of the sum of all propensities $\sum_j a_j(x)$. Mathematically, τ is given by:

$$\tau = \frac{1}{\sum_j a_j(x)} \ln\left(\frac{1}{r_1}\right)$$

   where r_1 is a uniformly distributed random number between 0 and 1.

4. **Select the Reaction to Occur**: The specific reaction that occurs next is determined by comparing another random number r_2 to the cumulative sum of the normalized propensities. The reaction is chosen so that:

$$\sum_{i=1}^{j-1} a_i < r_2 \cdot \sum_j a_j \leq \sum_{i=1}^{j} a_i$$

5. **Update the System**: Once the reaction is selected, the state of the system is updated according to the stoichiometry of the reaction, and the simulation time is advanced by τ.

6. **Repeat**: Steps 2-5 are repeated until a predefined stopping criterion is met, such as reaching a certain time or number of reactions.

This algorithm is particularly useful in simulating the stochastic behavior of undifferentiated stem cells as they undergo birth and death processes. By capturing the sequence and timing of individual events, the Gillespie algorithm allows us to obtain a numerical approximation of the system's state probability distribution over time, providing valuable insights into the dynamics of stem cell populations.

3.2 Analytical Results

Since our focus is on analyzing the differentiation process of stem cells, a crucial concept to consider is the time it takes for a cell to differentiate from its birth. A key quantity in this context is the **first passage time** (FPT), which refers to the time taken for a state variable to reach a specific value[40]. In our work, we define the **First Passage Time (FPT)**, referred to as the **lifetime duration** of a lineage, as the time it takes for the number of undifferentiated cells in a lineage to reach zero—meaning all cells have either differentiated or died (Figure 3.1).

To validate our model, we aim to analytically calculate the distribution of the lifetime duration and compare it with the results from our simulations. To achieve this, we begin by considering the stochastic dynamics of undifferentiated stem cell populations through the Fokker-Planck equation.

The Fokker-Planck equation describes the time evolution of the probability distribution of a continuous stochastic process. It captures how the probability density of a system's state changes over time, accounting for both deterministic forces (drift) and random fluctuations (diffusion). The equation is given by:

$$\frac{\partial P(x,t)}{\partial t} = -\frac{\partial}{\partial x}\left[A(x)P(x,t)\right] + \frac{\partial^2}{\partial x^2}\left[B(x)P(x,t)\right], \tag{3.4}$$

where $P(x,t)$ is the probability density function of finding the system in state x at time t, $A(x)$ is the drift coefficient representing deterministic forces, and $B(x)$ is the diffusion coefficient representing stochastic fluctuations.

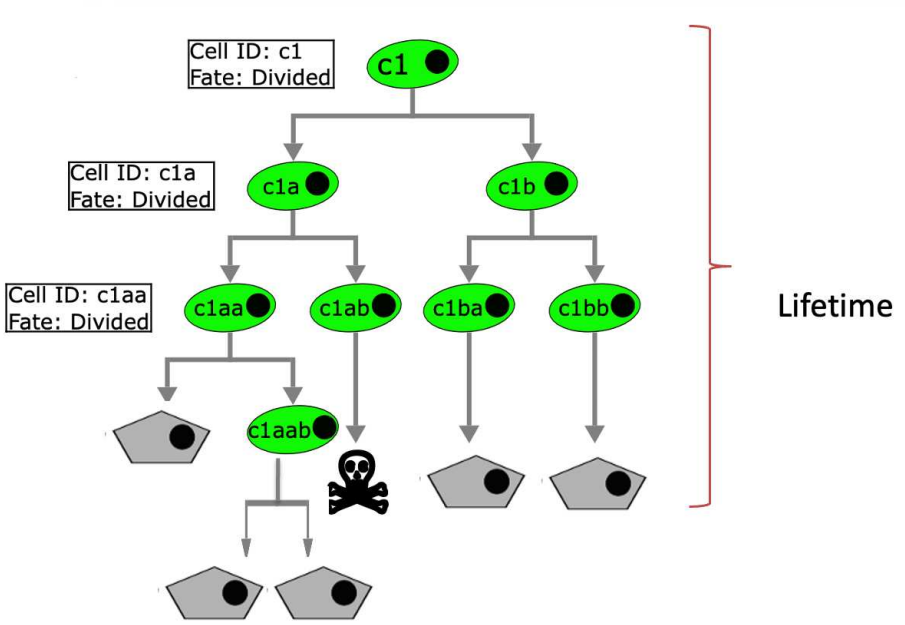


Figure 3.1: An illustration of the definition of the lifetime in a lineage.

However, directly solving the Fokker-Planck equation for our system is challenging due to the complexity introduced by absorbing boundary, when the number of undifferentiated cells reaches zero.

To address this, we use the probability distribution $p_{\text{abs}}(x, t \mid x_0, 0)$, derived by Azaele et al. [41], which describes the likelihood of having x undifferentiated cells at time t, given an initial condition of x_0 cells at $t = 0$. This is the time-dependent absorbing solution of the probability distribution function of finding x individuals at time t in the lineage and is given by:

$$p_{\text{abs}}(x, t \mid x_0, 0) = \frac{1}{D\tau} \cdot \frac{1}{1 - e^{-t/\tau}} \exp\left[-\frac{1}{D\tau} \cdot \frac{x + x_0 e^{-t/\tau}}{1 - e^{-t/\tau}}\right] \cdot \left(\frac{x}{x_0} e^{t/\tau}\right)^{-1/2} I_{1-\frac{b}{D}}\left[\frac{2}{D\tau} \cdot \frac{\sqrt{x_0 x e^{t/\tau}}}{e^{t/\tau} - 1}\right] \tag{3.5}$$

Where:

- $p_{\text{abs}}(x, t \mid x_0, 0)$ represents the probability density of observing x cells at time t, given an initial count x_0.
- $D = \frac{\text{birth rate} + \text{death rate}}{2}$ accounts for demographic stochasticity.

- $\tau = \frac{1}{\text{death rate} - \text{birth rate}}$ denotes the characteristic timescale of the system, capturing how quickly the population dynamics respond to stochastic fluctuations.
- b is the migration rate, which is set to zero in our model since our biological system does not involve migration.
- $I_\sigma(z)$ is the modified Bessel function of the first kind, which arises naturally in the solution of stochastic differential equations.

It is important to note that even though we have an absorbing condition, the probability does not vanish at $x = 0$.

By integrating Equation 3.5 with respect to x from 0 to infinity, and then taking the derivative with respect to time t, applying a negative sign to the result, we obtain the formula that describes the lifetime distribution of the system:

$$P_{\text{lifetime}}(t) = -\frac{2e^{\frac{t}{\tau}+\frac{x_0}{D\tau - De^{\frac{t}{\tau}}\tau}}\sqrt{e^{\frac{t}{\tau}}}x_0}{D\left(-1+e^{\frac{t}{\tau}}\right)^3\tau^2\sqrt{2\left(\frac{t}{2\tau}\right)}} \tag{3.6}$$

3.3 Numerical Simulations

Except from obtaining the lifetime distribution by using the analytical solution, we can compute it through simulation as well. We start by simulating the system by incorporating the transition rates into the Gillespie algorithm, calculating propensities based on the rates provided in equations (3.1) and (3.2). We start by modeling a simple birth and death process, implementing the algorithm with a birth rate coefficient $b = 3 \times 10^{-2}$ min^{-1} and a death rate coefficient $d = 5 \times 10^{-2}$ min^{-1}. To analyze system behavior, we simulate the number of undifferentiated cells in each lineage over time, starting with an initial cell population of $x_0 = 100$ at $t = 0$, and running the simulation for $t_{\text{max}} = 10^4$ minutes for 10^4 lineages. Although this duration and number of lineages exceed those observed in the experiment, they were chosen to ensure statistically robust results and to capture stochastic fluctuations effectively.

Since we simulate the number of undifferentiated cells in each lineage at all time points, we calculate the mean cell count over the entire duration of the simulation, as shown in Fig. 3.2. As expected, the mean decreases over time due to the higher death rate compared to the birth rate. The initial part of the simulation, up to around 250 minutes, represents a rapid drop in the mean cell count and a higher standard deviation, reflecting significant stochastic fluctuations. As the simulation progresses beyond this period, the mean cell count stabilizes.

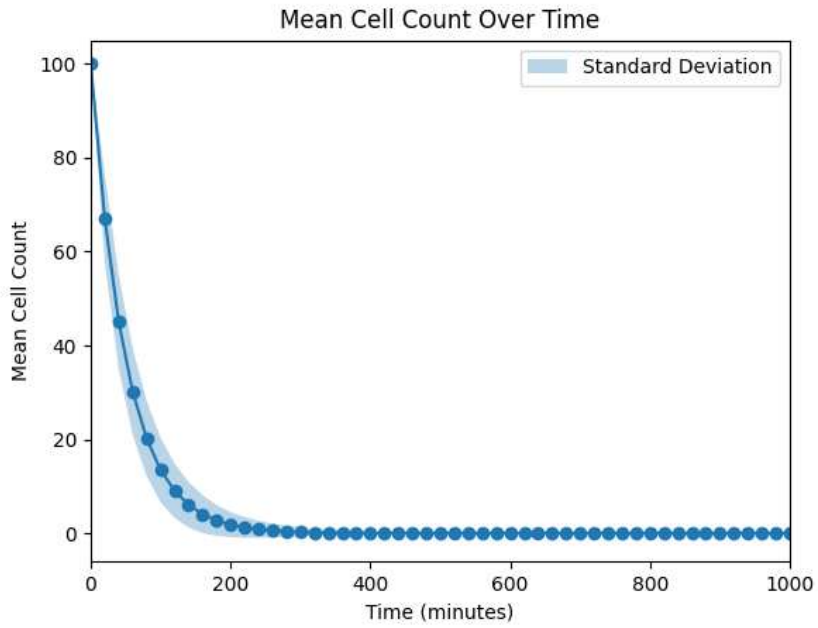


Figure 3.2: Mean cell count for 10^4 lineages.

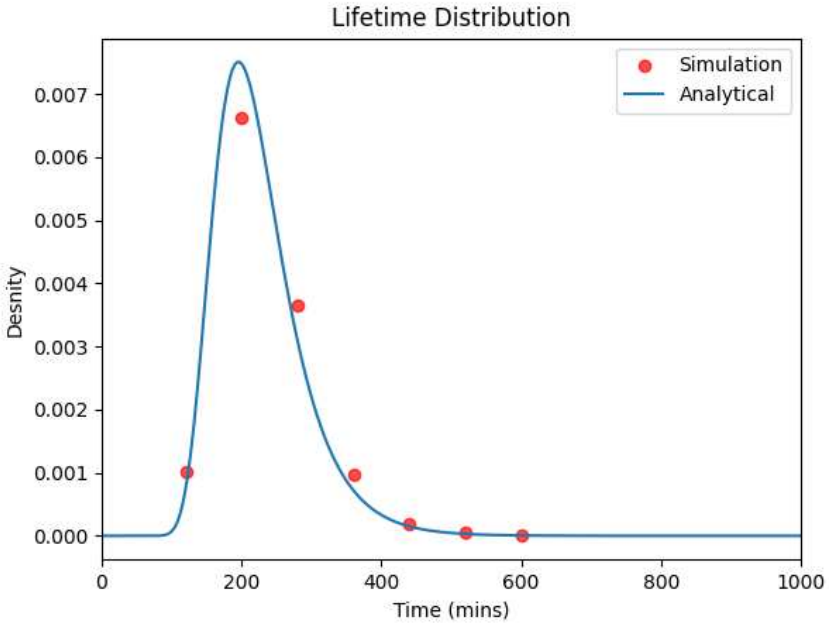


Figure 3.3: The lifetime distributions of the simple birth-death model computed for 10^4 lineages.

To validate our simulation results, we compare them with the analytical solution for the lifetime duration of the lineage. Using the analytical formula for lifetime distribution given by Equation 3.6, we calculate the lifetime distribution for our model. The values of the parameters D and τ are calculated based on the birth and death rates as follows: $D = 0.04\,\text{min}^{-1}$ and $\tau = 50$ minutes. By applying these parameters, we can generate the lifetime distribution for a range of x values and validate it against the simulation results to assess the model's accuracy. As shown in Figure 3.3, there is a general agreement between the simulation and the analytical solution, though some points deviate slightly from the curve. These minor discrepancies can be attributed to the binning of the points. Additionally, the deviations might be caused by the finite number of lineages that we are able to simulate.

Most cells exhibit a short lifetime, with a peak observed around 150-250 minutes. This is consistent with the model's design, which incorporates a higher death rate than birth rate, and aligns with the rapid decline in the mean cell count shown in Figure 3.2.

To further investigate the system's behavior, we modify the model by introducing a new reaction called "no change" that is present in the real biological system. When this process occurs, the system remains in its current state, with the number of cells n staying constant. The transition rate of "no change" can be described as follows:

$$T(n \to n) = q \cdot n \tag{3.7}$$

where q is the "no change" rate. With the introduction of the "no change" process, the Master Equation takes the following form:

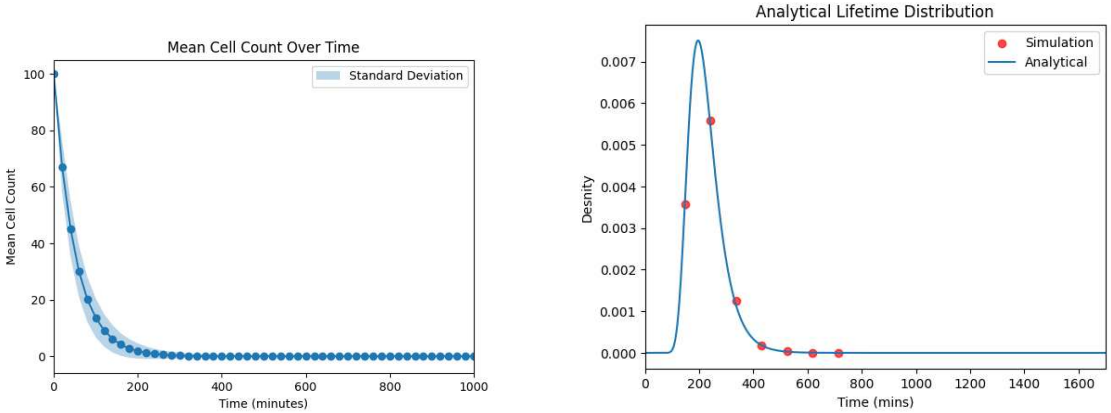
$$\begin{aligned}\frac{\partial P_n(t)}{\partial t} &= T(n-1 \to n)P_{n-1}(t) + T(n+1 \to n)P_{n+1}(t) \\ &\quad - \left[T(n \to n+1) + T(n \to n-1)\right] P_n(t) + P_n(t) \cdot T(n \to n)\end{aligned} \tag{3.8}$$

With the "no change" process now incorporated, we simulate the system using the Gillespie algorithm to assess its impact on the overall outcome. We set $q = 0.01$ and compare the results with the analytical distribution of the initial model that only includes birth and death processes. The plot of the mean cell count (Fig.3.4a) exhibits a similar pattern to that of the birth-death model (Fig. 3.2): a rapid decrease due to the higher death rate, followed by rapid drop of the mean cell count in period lasting until around 250 minutes, and eventually reaching a steady state with a near-zero mean cell count. Our observations indicate that the introduction of the

"no change" rate has a minimal effect on the overall outcome. Additionally, when comparing the lifetime duration to the analytical formula that does not account for the "no change" process, we find that the influence of the "no change" rate is negligible as the fit between the simulation results and the analytical predictions remains adequately good (Fig. 3.4b). These findings suggest that the "no change" process does not significantly alter the dynamics of the system.

To validate confirm this observation regarding the impact of the "no change" rate on the system's behavior, we experimented with different ratios of the birth and death rates to the "no change" rate. Keeping the birth and death rates constant, we ran simulations with $q = 0.05$, $q = 0.005$, and $q = 0.001$, in addition to the previously used $q = 0.01$. The results consistently showed that the introduction of the "no change" process has minimal effect on the system. In all cases, the mean cell count exhibited the same behavior: a rapid initial drop of the mean number of cells until around 250 minutes, and a subsequent steady state. The lifetime distribution in each scenario showed a good fit with the analytical solution (Figures 3.5a - 3.5f). Although there were variations in the placement of the peak and the spread of the distribution across different "no change" rates, no clear trend was observed, indicating that these shifts do not significantly impact the overall system dynamics.

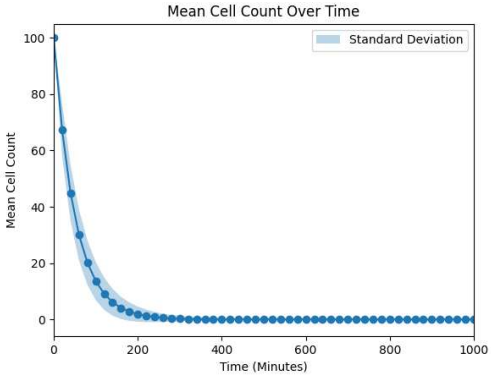
The model's outcomes are in agreement with the theoretical predictions, confirming its validity. Furthermore, for the birth and death process, the inclusion of a "no change" process, which maintains the system in its current state, was found to have a minimal effect on the overall behavior of the system. The lifetime distributions across different "no change" rates consistently align with the analytical solution, showing that varying the "no change" rate primarily influences the peak and spread of the lifetime distribution, but does not significantly alter the overall dynamics. This suggests that while the "no change" process introduces some variability in the timing of events, it does not fundamentally impact the system's long-term behavior.



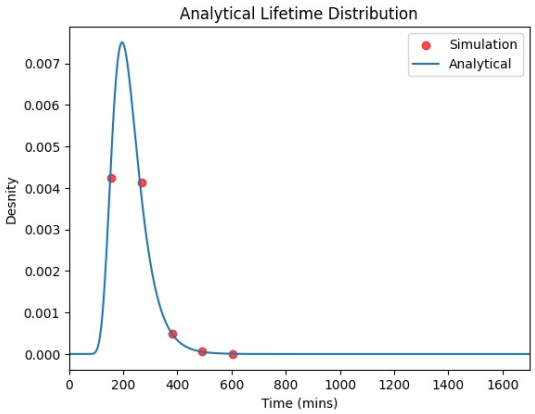
(a) Mean cell count of the model with $q = 0.01$

(b) Lifetime distribution of the model with $q = 0.01$

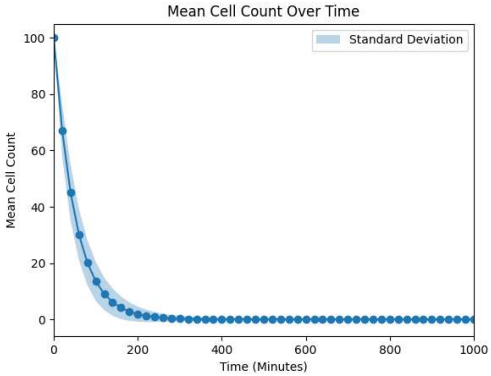
Figure 3.4: Comparison of mean cell counts and lifetime distribution for $q = 0.01$.



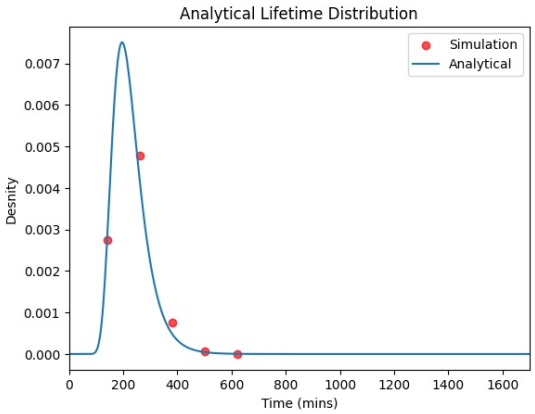
(a) Mean cell count of the model with $q = 0.05$



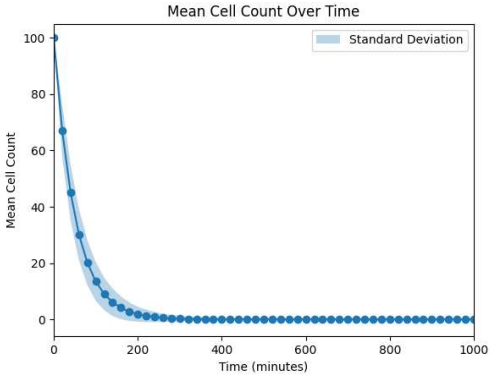
(b) Lifetime distribution of the model with $q = 0.05$



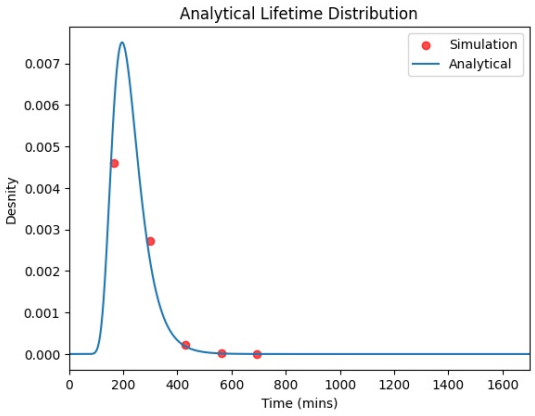
(c) Mean cell count of the model with $q = 0.005$



(d) Lifetime distribution of the model with $q = 0.005$



(e) Mean cell count of the model with $q = 0.001$



(f) Lifetime distribution of the model $q = 0.001$

Figure 3.5: Comparison of mean cell counts over time and lifetime distributions of a birth, death, ”no change” model across different values of the ”no change” rate q.

4 Results

4.1 Inferring Model Parameters from the Data

After examining a simple model, we can now implement the Gillespie algorithm for simulation to construct a model that accurately represents the behavior of the biological system observed in the experiment. The first step is to identify the types of processes that occur within the system during the experiment. As discussed in Chapter 2, cell divisions in the data result in six different outcomes. While divisions leading to the death of one or both daughter cells may not hold direct scientific interest for the analysis from a biological point of view, they are crucial for constructing the physical model, as they impact the population of undifferentiated stem cells.

All the division types observed in the data, excluding those where the cells are already differentiated, can be categorized into three processes similar to those used in the simple model with birth, death and "no change"(Figure 4.1):

- **Birth**: Symmetric division.
- **Death**: Symmetric differentiation and division where both daughter cells die.
- **No change**: Division with one daughter cell death and asymmetric division.

In Chapter 3, we discussed the transition rates and the Master Equation that describe these processes. These processes are defined by different types of divisions, therefor we can rewrite the transition rates to reflect the data more accurately. Since the transition rates represent velocities,

which define the probability of the system transitioning from one state to another, the birth, death, and ”no change” rates can be calculated by multiplying the probability of each type of division by the inverse of the mean cell cycle time, which we define as the division rate.

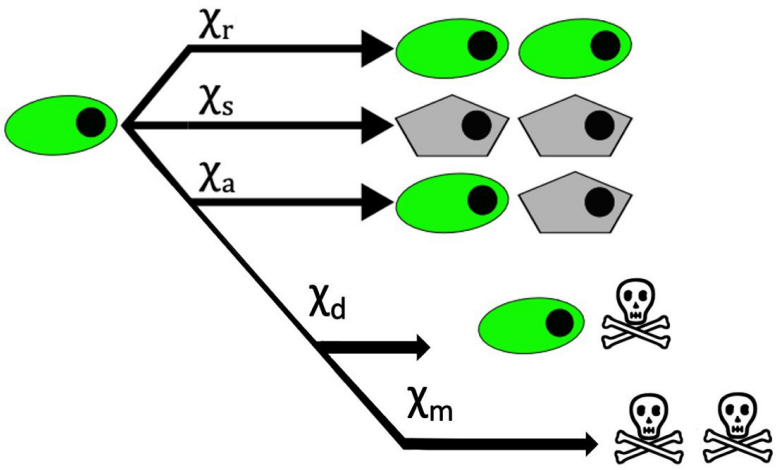


Figure 4.1: A scheme of the types of divisions and the probabilities that define them used to construct the model.

The probabilities of each division type, denoted by the χ terms, can be determined by calculating the number of each type of division observed in the data and dividing it by the total number of divisions. These χ values, along with the division rate, allow us to express the transition rates as follows:

$$T(n \to n+1) = \lambda \chi_r n \tag{4.1}$$

$$T(n \to n-1) = \lambda(\chi_s + \chi_m) n \tag{4.2}$$

$$T(n \to n) = \lambda(\chi_a + \chi_d) n \tag{4.3}$$

where:

- λ - Division rate
- n - Number of undifferentiated cells
- χ_r - Probability of symmetric division

- χ_s - Probability of symmetric differentiation
- χ_a - Probability of asymmetric division
- χ_d - Probability of division with one dead daughter
- χ_m - Probability of division with two dead daughters

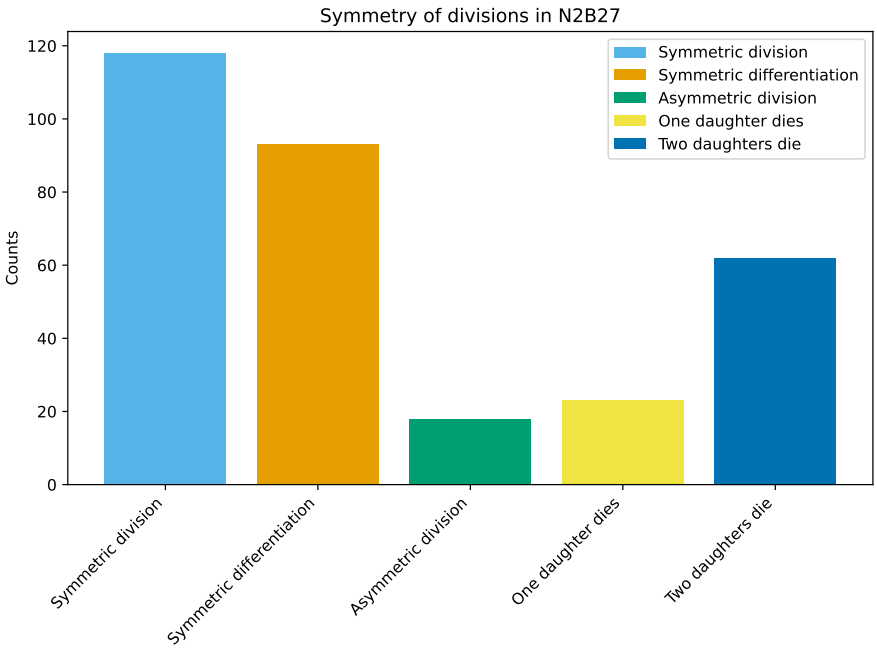


Figure 4.2: Histogram of different types of divisions used for constructing the physical model.

Considering that the χ terms represent probabilities, their sum equals 1:

$$\chi_r + \chi_s + \chi_a + \chi_d + \chi_m = 1 \tag{4.4}$$

We begin by constructing a model with constant transition rates. From our analysis, we infer that during the experiment, a total of 314 divisions were observed. Of these, 37.6% (118) were symmetric divisions, 29.6% (93) were symmetric differentiations, 5.7% (18) were asymmetric divisions, 7.3% (23) were divisions where one daughter cell died, and 19.7% (62) were divisions where both daughter cells died(Figure 4.2). The values of the χ terms are calculated by dividing the frequency of each type of division by the total number of divisions(Table 4.1).

From Chapter 2, we know that the mean cell cycle time for cells in the **N2B27** environment is 589 minutes. Using this information, we calculate the division rate as $\lambda = 1.698 \times 10^{-3}$

Type of Division	Symbol	Probability (χ)
Symmetric Division	χ_r	$\frac{118}{314}$
Symmetric Differentiation	χ_s	$\frac{93}{314}$
Asymmetric Division	χ_a	$\frac{18}{314}$
Division (One Daughter Cell Died)	χ_d	$\frac{23}{314}$
Division (Both Daughter Cells Died)	χ_m	$\frac{62}{314}$

Table 4.1: Types of cell divisions, their symbols, and corresponding probabilities.

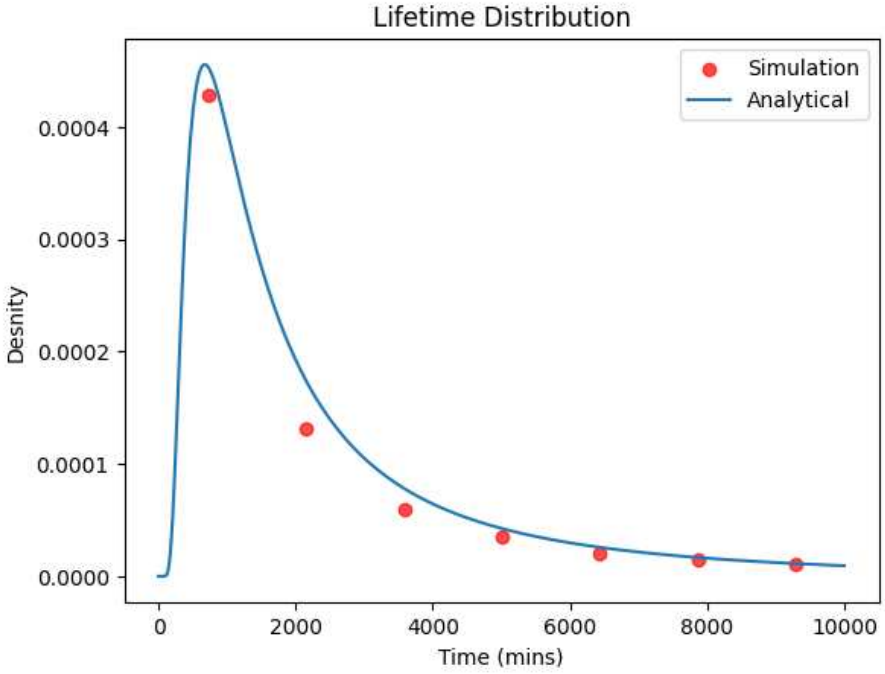


Figure 4.3: The pdf of simulated and analytical lifetime distributions of the constant rate model.

min^{-1}, birth rate as 6.38×10^{-4} min^{-1}, the death rate as 8.38×10^{-4} min^{-1}, and the "no change" rate as 2.22×10^{-4} min^{-1}. The simulation was run for the same duration as the models in Chapter 3 $t_{\text{max}} = 10^4$ minutes across 10^4 lineages and initiated with once cell, reflecting the initial conditions in the experimental data.

We use the formula 3.6 to compare the analytical and simulated lifetime distributions for the constant rate model. However, the fit between the two is not perfect (Figure 4.3). The deviations between the simulation and analytical results, particularly at shorter times, suggest that the constant rate model may not fully capture the system's dynamics, especially in the early stages. These discrepancies indicate that the model could be oversimplifying the underlying

biological processes, possibly missing key factors that influence stem cell differentiation.

The early peak in the simulation data, compared to the smoother analytical curve, may reflect stochastic fluctuations that the constant rate model does not adequately account for. This could indicate that the real system exhibits time-dependent behavior or other complex dynamics not captured by the constant rate model.

Based on these findings, we constructed a model with time-dependent rates to more accurately reflect the observed dynamics. To calculate the time-dependent division rate, we first determined the average value of the cell cycles for every 600-minute interval from the data. The interval of 600 minutes was chosen as a characteristic time since the mean cell cycle time for cells in **N2B27** is 589 minutes, which was rounded to 600 minutes for simplicity. Similarly, to calculate the time-dependent probabilities for each type of division, we counted the number of each type of division within each interval and divided it by the total number of divisions during that interval. By multiplying the division rate by the corresponding probability in each interval, we derived the time-dependent birth, death, and "no change" rates.

For consistency, the simulation for the time-dependent model was also run across 10^4 lineages for $t_{\max} = 10^4$ minutes. The time-dependent rates of the processes were then mapped onto the time of the experiment to ensure comparability between the simulated and observed data (Figures 4.4 and 4.5). In both constant and time-dependent cases, all lineages were initiated with one cell, reflecting the initial conditions in the experimental data.

While the analytical formula 3.6 cannot be applied to describe the lifetime distribution of the model with time-dependent rates, as it is based on the assumption of constant rates, the time-dependent model allows us to account for the dynamic nature of the cell cycle and division processes observed in the experiment.

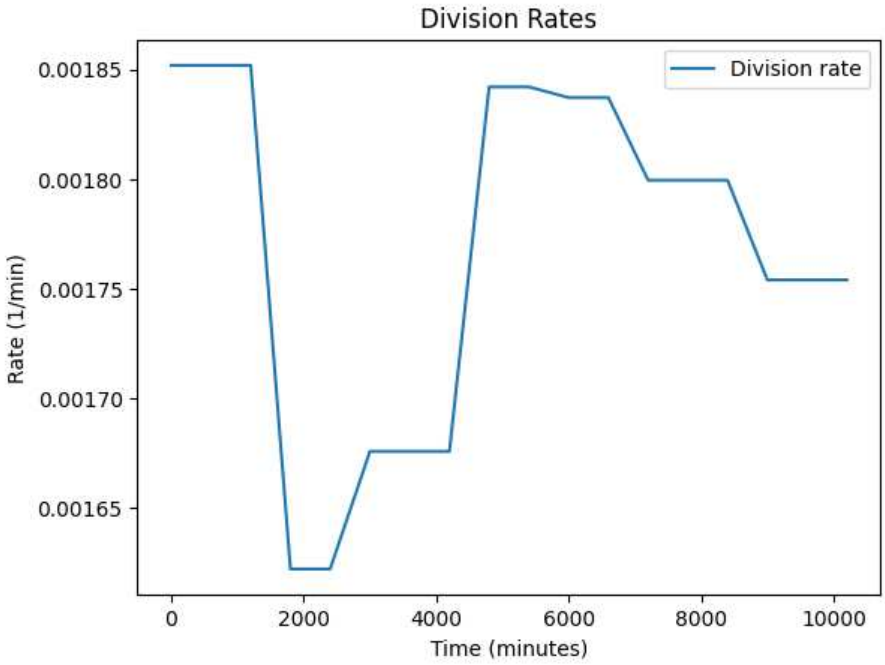


Figure 4.4: The division rates over time mapped onto the time of the simulation.

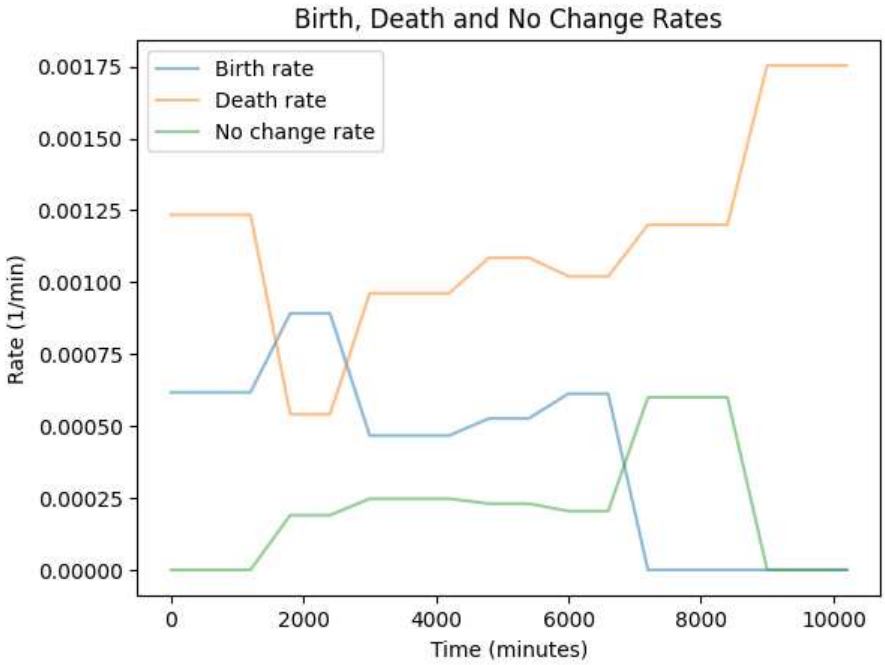


Figure 4.5: The birth, death and ”no change” rates over time mapped onto the time of the simulation.

4.2 Comparison of the Models and the Data

As mentioned before, our objective for constructing both the constant rate and time-dependent rate models is to identify the model that best represents the experimental data, allowing us to study the cell system *in silico*. In order to determine the most suitable model, we need to compare some key quantities that describe the behavior of the cells. However, considering the different timeframes of the experiment and the simulations, it would be meaningless to directly compare the mean cell count over time.

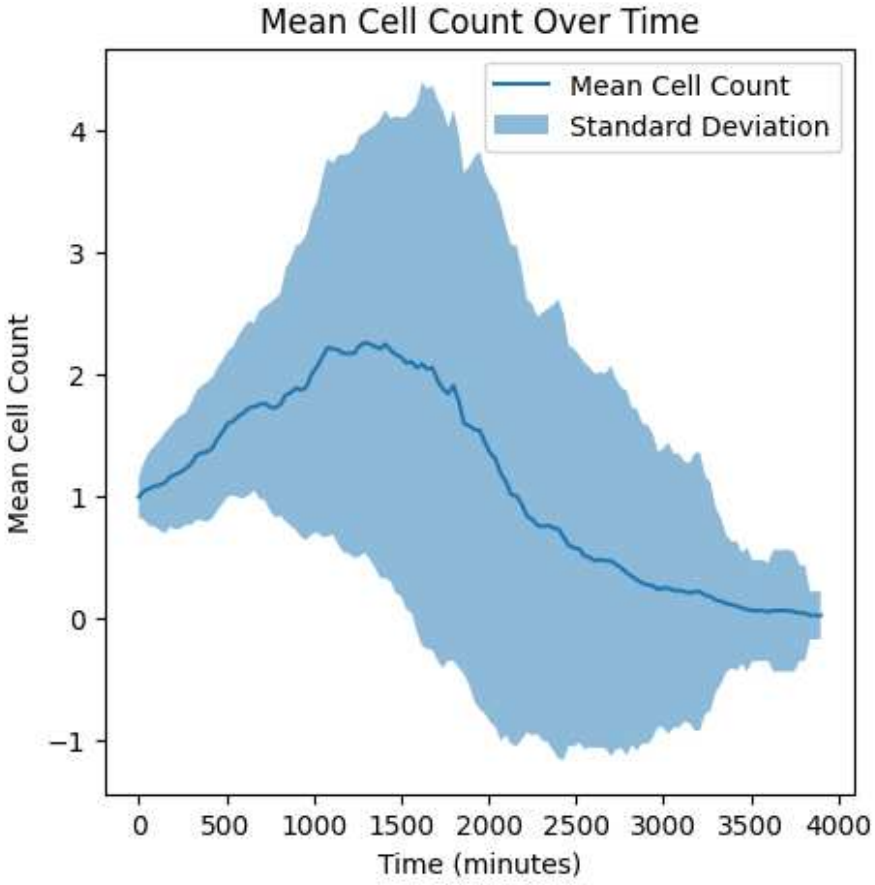


Figure 4.6: The mean cell count over time of the experimental data.

This is because the simulations are run for a much longer duration ($t_{\max} = 10^4$ minutes) than the actual experiment (72 hours), leading to differences in the observed dynamics. For instance, the mean cell count in the simulation may exhibit behavior that stabilizes or evolves over time in ways that are not observable within the shorter experimental timeframe. Addition-

ally, the accumulation of stochastic fluctuations in the simulation over the longer period can further skew direct comparisons. Therefore, while we cannot directly compare the absolute values of the mean cell count over time, we can still analyze and compare the overall trends and behavior of the cell populations between the models and the experimental data.

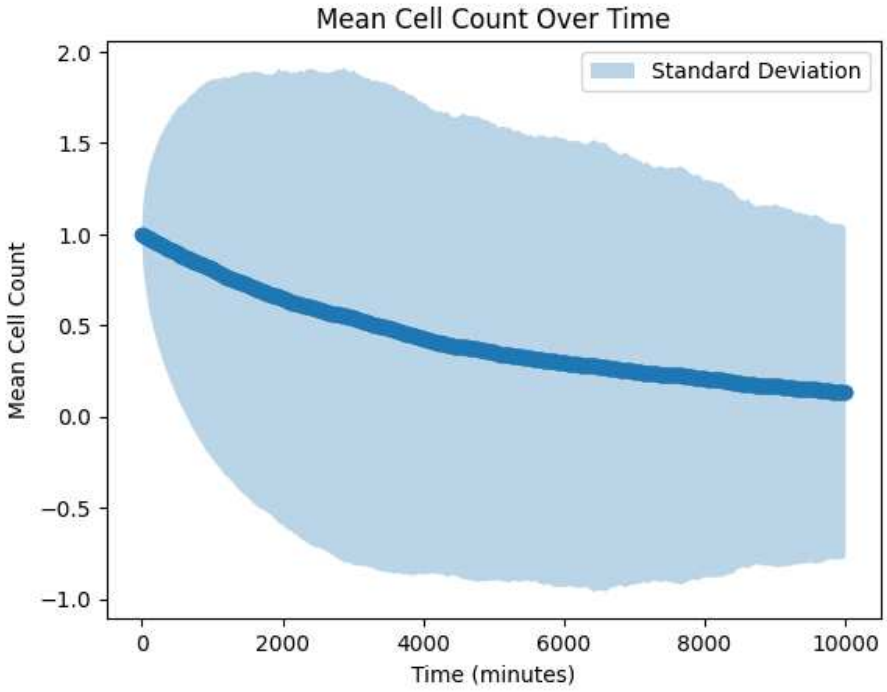


Figure 4.7: The mean cell count over time of the constant rate model.

The experimental data reveal a non-monotonic trend in the mean cell count over time (Figure 4.6). In contrast, the model with constant rates exhibits a monotonic decline, which is expected given the higher death rate compared to the birth and ”no change” rates (Figure 4.7). However, the time-dependent rate model produces a non-monotonic pattern in mean cell count, aligning more closely with the observed data, although the range of cell counts differs (Figure 4.8). Furthermore, the time-dependent model captures the fluctuations and variability seen in the experimental data, suggesting that the underlying biological processes influencing cell cycle duration and division rates may also be time dependent.

In addition to comparing the mean cell count over time, we can also compare quantities that remain consistent even when the duration of the simulation exceeds that of the experiment since they show a distribution. One such quantity is the distribution of the cell population, which shows the frequency of occurrence of a given number of cells in the lineages.

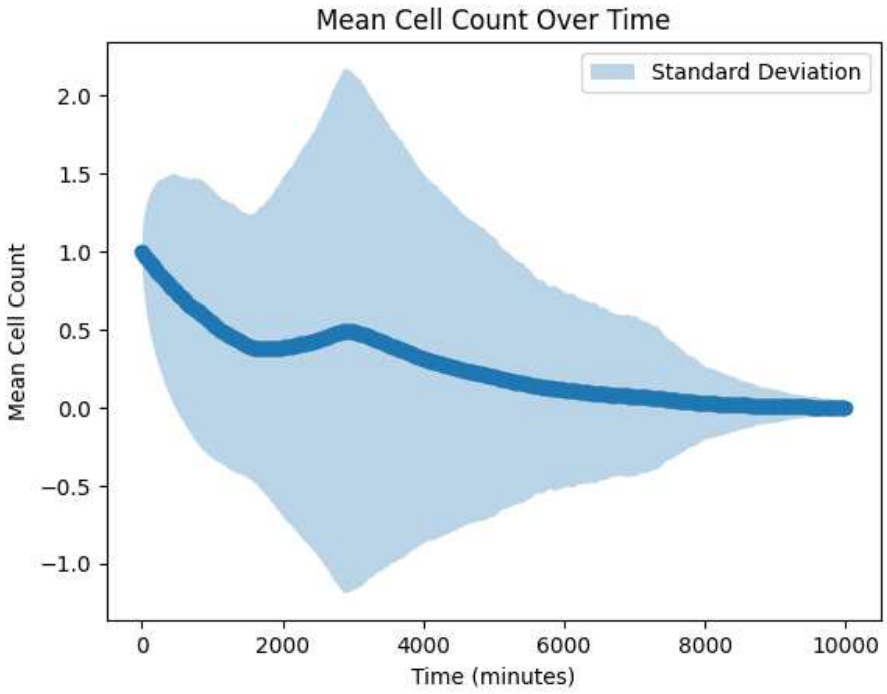


Figure 4.8: The mean cell count over time of the time-dependent rate model.

As demonstrated in Table 4.2, while the time-dependent model aligns more closely with the experimental data, both models effectively describe the central tendency and variability of the cell population(Figure 4.9).

Statistic	Constant Rate Model	Time-Dependent Model	Experimental Data
Mean	12.5	11.5	10.94
Median	12.5	11.5	9.0
Lower Quartile (25%)	6.5	6.0	7.0
Upper Quartile (75%)	18.5	17.0	14.5

Table 4.2: Statistics comparison of cell population mean, median, and quartiles between constant rate, time-dependent rate models, and experimental data

Another quantity that we can compare in the same manner is the first passage time. Although the simulations run longer than the duration of the experiment, the first passage times are still comparable since the system reaches an absorbing state after some time, and the comparison is between the distributions of these values.

The statistics for the first passage times are summarized in Table 4.3. These results suggest that while the models provide a reasonable approximation, the experimental data exhibit longer

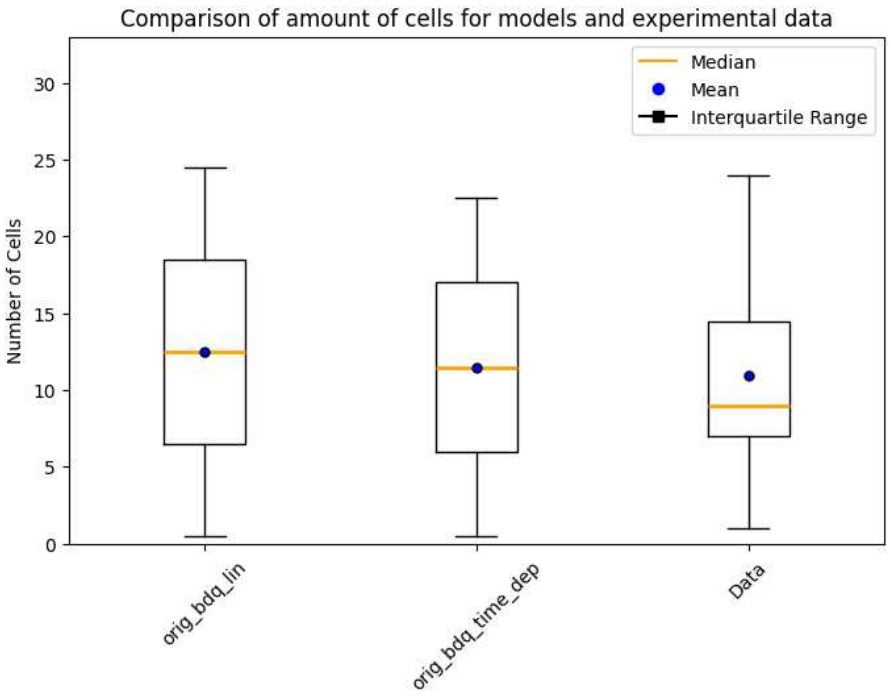


Figure 4.9: Boxplots of the distribution of the amount of cells in the lineages of the constant rate and time-dependent rate models with the experimental data

first passage times. This discrepancy indicates that there may be some lag in the death process in the actual data, as the first passage times are higher in the data compared to the models (Figure 4.10).

Statistic	Constant Rate Model	Time-Dependent Model	Experimental Data
Mean (minutes)	1763.72	1370.89	2380.82
Median (minutes)	1005.0	660.0	2340.1
Lower Quartile (25%) (minutes)	375.0	260.0	1995.05
Upper Quartile (75%) (minutes)	2355.0	1480.0	2865.05

Table 4.3: Statistics comparison of first passage time statistics between constant rate, time-dependent rate models, and experimental data

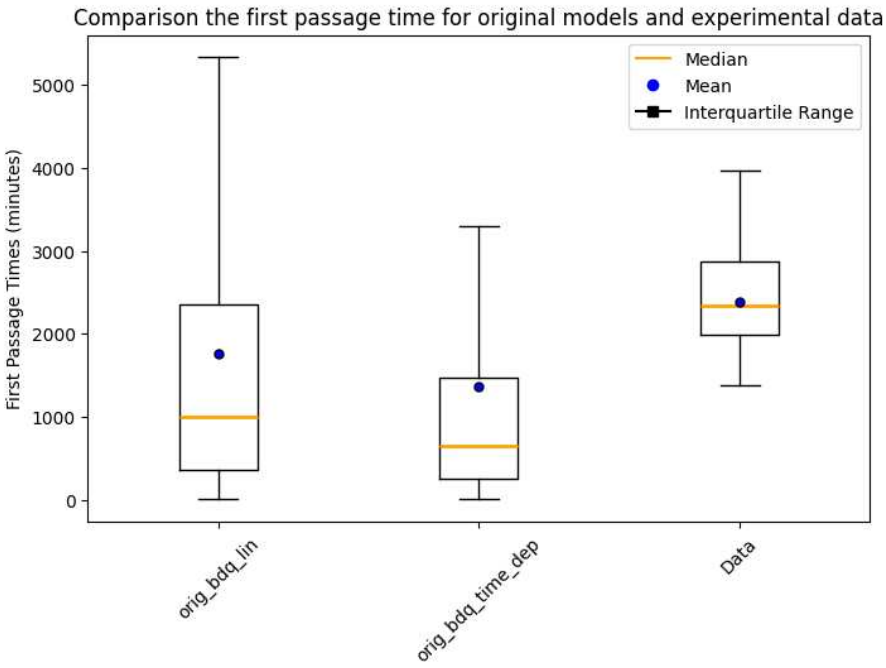


Figure 4.10: Boxplots of the distribution of the First Passage Times of the constant rate and time-dependent rate models with the experimental data

5 Discussion and Conclusion

In this work, we analyzed the dynamics of mouse ESCs by examining experimental data and building a minimal stochastic model to study the differentiation process. Our focus was on exploring the mechanisms through which these cells dynamically transition during differentiation, a process characterized by complex, non-linear and stochastic behaviors. Understanding these processes is crucial, as they play a central role in the development and maintenance of tissues, with disruptions potentially leading to diseases or developmental disorders.

The interdisciplinary nature of this study, which combined systems biology with methods from statistical mechanics, proved to be effective. Stochastic modeling, particularly through the Gillespie algorithm, allowed us to simulate the inherent randomness and variability in biological systems. This approach provided insights that deterministic models might fail to capture, especially in representing the stochastic nature of cellular events.

By building and exploring a simple birth and death model, we test the agreement between the simulation data and the theoretical predictions of the lifetime distribution of the cells for different birth and death rates. Additionally, this approach allowed us to investigate the effect of the "no change" rate on the simulation, which was found to be minimal. Although variations in the peak and spread of lifetime distributions were observed, no clear trend emerged, indicating that this process introduces variability without fundamentally altering the overall behavior of the system.

The simple model paved the way for simulating the behavior of the cells observed in the experiment by implementing the Gillespie algorithm for the processes identified during the ex-

periment and by inferring the model parameters from the experimental data. We tested both a model with constant rates and a model with time-dependent rates. The time-dependent rate model performed better in capturing the non-monotonic nature of the mean cell count over time, the signature of the non-linearity of the underlying dynamics. Furthermore, a comparison of the distribution of cell counts and the FPTs from the simulations with the experimental data revealed that while both models provided a reasonable representation of the data, they could be both improved, as the FPTs were higher in the data compared to the models.

We encountered several limitations during our work that impacted our ability to fully explore and validate our models. One significant limitation was the shorter duration of the experiment compared to the simulation time. This discrepancy made it difficult to directly compare time-dependent quantities, as the shorter experimental timeframe was not enough to generate significant statistics to then compare with the model. Put differently, for such short times series, the stochastic variability is extremely high, and it is difficult to estimate robust transient properties.

In other words, the limited amount of data posed challenges for our analysis. While the available data were sufficient for drawing biological conclusions, it fell short when used for discrimination of different stochastic models, where a larger dataset would have allowed for more robust validation and a finer resolution of the underlying processes.

Another challenge was the computational power required for our simulations. Stochastically simulating a large number of elements over extended periods is computationally intensive, and at times, our resources were insufficient to run these simulations as efficiently or as extensively as desired. This limitation restricted our ability to explore alternative scenarios or conduct more in-depth sensitivity analyses, which could have provided additional insights into the system's behavior.

Building on the limitations encountered in this study, there are several avenues for future work that could further enhance the modeling of stem cell dynamics.

Addressing the limitation of data quantity by obtaining a larger and more diverse dataset would significantly improve the robustness of model inference and validation. A more extensive dataset would in fact enable a finer resolution of the stochastic processes and provide a more reliable basis for parameter inference and model comparison.

Moreover, improving computational efficiency is essential for fast and more extensive simulations. With enhanced computational power or more efficient algorithms, it would be possible to explore a wider range of scenarios, conduct more detailed sensitivity analyses, and potentially simulate larger systems or longer time periods, thereby gaining deeper insights into the

behavior of the model.

Exploring alternative stochastic algorithms could potentially offer better modeling for simulating stem cell differentiation dynamics by enhancing accuracy and efficiency. For instance, the *population-based Monte Carlo (PBMC)* method is well-suited for handling large systems with significant population variability, potentially reducing computational time while maintaining accuracy[42]. The *moment closure technique*, which approximates the moments of the probability distribution, could also be explored for more efficient simulations without fully solving the Chemical Master Equation[43].

Despite the challenges encountered, this study yielded several positive outcomes. The time-dependent rate model successfully captured the non-monotonic behavior observed in the experimental data, providing a closer representation of the stem cell differentiation process. The comparison of simulation results with experimental data demonstrated that our models could effectively reproduce key aspects of stem cell dynamics, such as the distribution of cell counts and FPTs. These successes highlight the potential of stochastic modeling to uncover insights into complex biological systems that deterministic approaches might miss.

This work emphasizes the crucial role of integrating experimental data with stochastic modeling to gain deeper insights into the complexities of stem cell dynamics. By developing these models, we gain a fundamental understanding of the inherent randomness and variability in biological processes, which are essential to accurately capture how stem cells differentiate and develop.

As stem cells are central to tissue repair and regeneration, the ability to predict and manipulate their behavior through precise modeling has the potential to transform therapeutic approaches. The developments made in this study can help develop more refined models that could eventually guide the creation of targeted treatments for a variety of conditions, including degenerative diseases and tissue repair. Ultimately, this work brings us closer to realizing the full potential of stem cell research, paving the way for more effective treatments and a profound understanding of human biology.

References

[1] S. E. Strawbridge, G. B. Blanchard, A. Smith, H. Kugler, and G. Martello, "Embryonic stem cells commit to differentiation by symmetric divisions following a variable lag period," *bioRxiv*, 2020. [Online]. Available: https://www.biorxiv.org/content/early/2020/06/18/2020.06.17.157578

[2] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *Journal of Physical Chemistry*, vol. 81, no. 25, pp. 2340–2361, 1977.

[3] U. Alon, *An Introduction to Systems Biology: Design Principles of Biological Circuits*. CRC Press, 2006.

[4] H. Kitano, "Systems biology: A brief overview," *Science*, vol. 295, no. 5560, pp. 1662–1664, 2002.

[5] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, "From molecular to modular cell biology," *Nature*, vol. 402, no. 6761, pp. C47–C52, 1999.

[6] N. E. Radde and M.-T. Hütt, "The physics behind systems biology," *EPJ Nonlinear Biomedical Physics*, vol. 4, no. 1, p. 7, 2016. [Online]. Available: https://doi.org/10.1140/epjnbp/s40366-016-0034-8

[7] A. Kimura, *Introduction to Quantitative Biology*. Singapore: Springer Nature Singapore, 2022, pp. 1–10. [Online]. Available: https://doi.org/10.1007/978-981-16-5018-5_1

[8] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *The Journal of Physiology*, vol. 117, no. 4, pp. 500–544, 1952.

[9] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[10] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, "The large-scale organization of metabolic networks," *Nature*, vol. 407, no. 6804, pp. 651–654, 2000.

[11] T. Ideker, T. Galitski, and L. Hood, "Integrated genomic and proteomic analyses of a systematically perturbed metabolic network," *Science*, vol. 292, no. 5518, pp. 929–934, 2001.

[12] A. Ashkin, "Optical trapping and manipulation of neutral particles using lasers," *Proceedings of the National Academy of Sciences*, vol. 94, no. 10, pp. 4853–4860, 1997.

[13] L. Hood, J. R. Heath, M. E. Phelps, and B. Lin, "Systems biology and new technologies enable predictive and preventative medicine," *Science*, vol. 306, no. 5696, pp. 640–643, 2004.

[14] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry*. Amsterdam: North Holland, 1992.

[15] J. Paulsson, "Summing up the noise in gene networks," *Nature*, vol. 427, no. 6973, pp. 415–418, 2004. [Online]. Available: https://doi.org/10.1038/nature02257

[16] M. B. Elowitz and S. Leibler, "A synthetic oscillatory network of transcriptional regulators," *Nature*, vol. 403, no. 6767, pp. 335–338, 2000. [Online]. Available: https://doi.org/10.1038/35002125

[17] P. S. Swain, M. B. Elowitz, and E. D. Siggia, "Intrinsic and extrinsic contributions to stochasticity in gene expression," *PNAS*, vol. 99, no. 20, pp. 12 795–12 800, 2002. [Online]. Available: https://doi.org/10.1073/pnas.162041399

[18] G. Tkačik, C. G. Callan Jr, and W. Bialek, "Information flow and optimization in transcriptional regulation," *PNAS*, vol. 105, no. 34, pp. 12 265–12 270, 2008. [Online]. Available: https://doi.org/10.1073/pnas.0806077105

[19] T. Mora and W. Bialek, "Are biological systems poised at criticality?" *Journal of Statistical Physics*, vol. 144, no. 2, pp. 268–302, 2011. [Online]. Available: https://doi.org/10.1007/s10955-011-0229-4

[20] D. Krotov, J. O. Dubuis, T. Gregor, and W. Bialek, "Morphogenesis at criticality," *PNAS*, vol. 111, no. 10, pp. 3683–3688, 2014. [Online]. Available: https://doi.org/10.1073/pnas.1324186111

[21] C. E. Murry and G. Keller, "Differentiation of embryonic stem cells to clinically relevant populations: Lessons from embryonic development," *Cell*, vol. 132, no. 4, pp. 661–680, 2008.

[22] J. Nichols and A. Smith, "Naive and primed pluripotent states," *Cell Stem Cell*, vol. 4, no. 6, pp. 487–492, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1934590909002240

[23] P. D. Tonge, X. Corso-Díaz, C. Monetti, M. C. Puri, N. Cloonan, A. Lo Nigro, M.-L. Li, D. Lee, and A. Nagy, "Divergent reprogramming routes lead to alternative stem-cell states," *Nature*, vol. 516, pp. 192–197, 2014.

[24] R. A. Young, "Control of the embryonic stem cell state," *Cell*, vol. 144, no. 6, pp. 940–954, 2011.

[25] G. Martello and A. Smith, "The nature of embryonic stem cells," *Annual Review of Cell and Developmental Biology*, vol. 30, pp. 647–675, 2014.

[26] M. G. Y. B. E. S. S. A. G. Dunn, S.-J., "Defining an essential transcription factor program for naïve pluripotency," *Science*, vol. 344, pp. 1156–1160, 2014.

[27] Y. Shavit, B. Yordanov, S.-J. Dunn, C. M. Wintersteiger, T. Otani, Y. Hamadi, F. J. Livesey, and H. Kugler, "Automated synthesis and analysis of switching gene regulatory networks," *Biosystems*, vol. 146, pp. 26–34, 2016, information Processing in Cells and Tissues. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0303264716300338

[28] B. D. MacArthur, A. Ma'ayan, and I. R. Lemischka, "Systems biology of stem cell fate and cellular reprogramming," *Nature Reviews Molecular Cell Biology*, vol. 10, pp. 672–684, 2012.

[29] C. Li and J. Wang, "Quantifying waddington landscapes and paths of non-adiabatic cell fate decision transitions," *Journal of The Royal Society Interface*, vol. 10, p. 20130787, 2013.

[30] C. Mulas, T. Kalkan, and A. Smith, "Nodal secures pluripotency upon embryonic stem cell progression from the ground state," *Stem Cell Reports*, vol. 9, no. 1, pp. 77–91, Jul 2017.

[31] H. G. Leitch, K. Blair, W. Mansfield, H. Ayetey, P. Humphreys, J. Nichols, M. A. Surani, and A. Smith, "Embryonic germ cells from mice and rats exhibit properties consistent with a generic pluripotent ground state," *Development*, vol. 137, no. 14, pp. 2279–2287, Jul 2010.

[32] A. Domogatskaya, S. Rodin, and K. Tryggvason, "Laminin, key component of the extracellular matrix, is crucial for cell adhesion and growth," *Cell Adhesion Migration*, vol. 6, no. 3, pp. 203–207, 2012.

[33] W. Wongpaiboonwattana and M. P. Stavridis, "Neural differentiation of mouse embryonic stem cells in serum-free monolayer culture," *Journal of Visualized Experiments*, no. 99, p. e52823, May 2015.

[34] Q.-L. Ying, J. Wray, J. Nichols, L. Batlle-Morera, B. Doble, J. Woodgett, P. Cohen, and A. Smith, "The ground state of embryonic stem cell self-renewal," *Nature*, vol. 453, no. 7194, pp. 519–523, 2008.

[35] P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: Appropriate use and interpretation," *Anesthesia & Analgesia*, vol. 126, no. 5, pp. 1763–1768, May 2018.

[36] F. S. Nahm, "What the p values really tell us," *Korean Journal of Pain*, vol. 30, no. 4, pp. 241–242, Oct 2017. [Online]. Available: https://doi.org/10.3344/kjp.2017.30.4.241

[37] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures: Third Edition*, 3rd ed. Chapman and Hall/CRC, 2003.

[38] A. Arkin, J. Ross, and H. H. McAdams, "Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected escherichia coli cells," *Genetics*, vol. 149, no. 4, pp. 1633–1648, 1998.

[39] S. Suweis, *Physical Model of Living System*, G. Cataldi, F. Manzali, and L. Rosset, Eds., Università degli Studi di Padova, Astronomy and Physics Department, 2024, master Degree in Physics of Data, 2023-2024.

[40] S. Redner, "A first look at first-passage processes," 2023. [Online]. Available: https://arxiv.org/abs/2201.10048

[41] S. Azaele, J. R. Banavar, A. Maritan, and J. M. Stacey, "Universal dynamics of human pathogen incidence and prevalence," *Nature*, vol. 444, no. 7121, pp. 743–746, 2006. [Online]. Available: https://doi.org/10.1038/nature05320

[42] D. M. Bortz, P. W. Nelson, and V. A. Jansen, "Population-based stochastic simulation algorithm for chemical kinetics," *SIAM Journal on Scientific Computing*, vol. 27, no. 2, pp. 674–693, 2006. [Online]. Available: https://doi.org/10.1137/040605330X

[43] R. Grima, "A study of the accuracy of moment-closure approximations for stochastic chemical kinetics," *The Journal of Chemical Physics*, vol. 136, no. 15, p. 154105, 2012. [Online]. Available: https://doi.org/10.1063/1.3701840

Acknowledgments

I would like to express my gratitude to Prof. Samir Suweis for his invaluable supervision and guidance throughout my thesis.

I am thankful to Prof. Graziano Martello for supervising my internship at the Martello laboratory and for expanding my perspective through the interdisciplinary experience in his lab.

A special thanks to Dr. Clelia Corridori for her assistance and support throught every step of my internship and thesis.

I am also grateful to Prof. Sandro Azaele for sharing his work, which has been crucial for my thesis.

I also extend my thanks to all the professors who have taught me during my master's program for their knowledge and mentorship.

Lastly, I am grateful to my family and friends for their constant support and encouragement, which have been a source of strength throughout this journey.