ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Rejection Gillespie Algorithm for non-Markovian Stochastic Processes

Master Thesis

Miroslav Phan

February, 2021

Advisors: Prof. Dr. Niko Beerenwinkel, Dr. María Rodríguez Martínez, Aurélien Pelissier
Department of Biosystems Science and Engineering, ETH Zürich

Abstract

The Gillespie algorithm is commonly applied for simulating Poisson point processes, which follow an exponential waiting-time distribution, and are completely memoryless. However, stochastic processes governing biochemical interactions are empirically known to exhibit properties of memory, an inherently non-Markovian feature. The presence of such non-Markovian processes can significantly influence the course of a simulation. In order to handle processes with memory, several extensions to the Gillespie algorithm have been proposed. However, these extensions are either limited by a high computational cost or are only applicable to a narrow selection of probability distributions found in biology. To challenge the aforementioned issues, a non-Markovian Rejection Gillespie algorithm is proposed in this thesis. The new algorithm is capable of generating simulations with non-exponential waiting-times, while remaining highly computationally efficient. It also incorporates the flexible Weibull distribution, allowing it to interpolate between a range of distributions relevant for biological processes. The proposed method was subsequently evaluated against an experimental dataset, for which it was able to correctly capture the underlying non-Markovian dynamics.

Contents

Chapter 1

Introduction

Stochastic processes involved in biochemical systems have commonly been modelled as independent Poisson point processes [12]. Reactants in such systems exhibit a Markovian (memoryless) property, whereby their fate is influenced only by their current state, and their reaction waiting-times are exponentially distributed [21]. However, in the case of many real world processes, the Markovian assumption does not hold. Examples within the realm of biology include cell proliferation [29], cell apoptosis [7], epidemic spreading dynamics [8], and retroviral budding events [11]. Further evidence suggests non-Markovian dynamics beyond biochemical processes, such as patterns of humans of activity [2], financial markets [9], and earthquake occurrences [5].

The presence of non-Markovian processes can significantly alter the course of a simulation [26], and thus they need to be modelled correctly. The Direct Gillespie method [10] has been extensively used for simulating large stochastic systems, yet it relies on the Markovian assumption in order to remain stochastically exact and computationally efficient. To overcome the issue, Boguñá and colleagues introduced an extension to the Gillespie algorithm that allows for time-dependent reaction rates [3]. However, the extension is limited by a high computational cost and low accuracy for simulations of small reactant population sizes. Another solution provided by Masuda and Rocha [17] uses the Laplace transform, which can only efficiently simulate long-tailed distributions, while being heavily constrained for other distributions more commonly found in biology, such as the log-normal.

Several methods have been developed to reduce the computational cost of the Direct Gillespie algorithm [31][28][25] by exploiting the Poisson thinning operation [15]. They work by discarding unfavourable reactions, which is equivalent to adding an additional empty Poisson point process, making the approach consistent with assumptions stipulated by the Direct Gillespie.

The non-Markovian Rejection Gillespie algorithm (NMRG) proposed in this thesis builds on the approaches listed above in order to create a computationally efficient Gillespie algorithm with individual reactant properties. In addition, NMRG utilises a reparametrised Weibull distribution [16] to allow the modulation of the shape of reactant waiting-time distributions. It also seeks to reduce the approximation error for simulations of small population sizes. The model was subsequently evaluated against a mouse embryonic stem cell dataset with known non-Markovian dynamics [24].

Chapter 2

Methods

2.1 Direct Gillespie Algorithm

The Direct Gillespie Algorithm (DG) [10] applies a Monte Carlo method to numerically simulate the time evolution of a system, generating statistically correct trajectories of its reactants. It assumes N independent Poisson point processes with propensity $\lambda_i \, (1 \leq i \leq N)$. According to the point process superposition theorem, the union of these processes also forms a Poisson point process $\lambda_0 = \sum_{i=1}^{N} \lambda_i$. The next time increment is determined by

$$\Delta t = \frac{\ln(1/u)}{\sum_{i=1}^{N} \lambda_i}, \tag{2.1}$$

where $u \in U(0,1)$ is drawn from a uniform distribution. We will see in Section 2.2.3 that Δt follows an exponential distribution. The probability of a process i undergoing a reaction at the determined time increment is given by

$$p_i = \frac{\lambda_i}{\sum_{i=1}^{N} \lambda_i}. \tag{2.2}$$

The time is advanced by Δt and the number of reactant molecules is updated according to the stoichiometry of the selected reaction i. For first-order reaction kinetics, the propensity λ_i for reaction i depends on the number of reactant molecules N_i and their reaction rate r_i: $\lambda_i = N_i \cdot r_i$. The pseudo-code for DG is described in Algorithm 1.

2.2 Non-Markovian Gillespie Algorithm

The Direct Gillespie described above can be modified for time-dependent reaction rates of reactants, and thus allow for non-Markovian inter-event times. We consider again N independent stochastic processes, now with

Algorithm 1 Direct Gillespie Algorithm

Time $t \leftarrow 0$
Initialise list of reactants
while $t < t_{end}$ **do**
  Compute total propensity $\lambda_0 = \sum_{i=1}^{N} \lambda_i$
  Generate uniform random number $u \in U(0,1)$
  Compute time increment $\Delta t = \frac{\ln(1/u)}{\lambda_0}$
  Select the firing of reaction i with probability $p_i = \frac{\lambda_i}{\lambda_0}$
  Update list of reactants according to stoichiometry of reaction i
  $t \leftarrow t + \Delta t$
end while

each reactant treated as an individual process with *rate* $\lambda_i(t_i)$ $(1 \leq i \leq N)$, where t_i is the time elapsed since the last firing of process i. *Note that in this section, each reactant is considered as a separate reaction channel. Therefore, λ_i will now correspond to the instantaneous reaction rate, as opposed to reaction propensity in Section 2.1.* The equations previously described in the section on DG for determining Δt (Eq. (2.1)) and p_i (Eq. (2.2)) can be rewritten as follows, with the time increment

$$\Delta t = \frac{\ln(1/u)}{\sum_{i=1}^{N} \lambda_i(t_i)} \tag{2.3}$$

and with the probability of a process i generating an event over all processes j $(1 \leq j \leq N)$ as

$$p_i = \frac{\lambda_i(t_i + \Delta t)}{\sum_{j=1}^{N} \lambda_j(t_j + \Delta t)}. \tag{2.4}$$

2.2.1 First order Taylor series approximation

Non-Markovian Gillespie follows the correct distribution

This non-Markovian extension is still stochastically exact and consistent with the Direct Gillespie algorithm. Boguñá and colleagues took advantage of the *first order Taylor series approximation* to propose the following proof [3]:

Theorem 2.1 *The non-Markovian extension follows a stochastically exact Gillespie algorithm*

Proof Let t_i be the time elapsed since the last firing of process i, with its probability density given by $\psi(t_i)$. We can then write the survival function for i (the probability of an inter-event time being larger than t_i) as

$$\Psi_i(t_i) = \int_{t_i}^{\infty} \psi_i(\tau) d\tau. \tag{2.5}$$

The probability that no process generates an event within the time step Δt is given by

$$\Phi\left(\Delta t \mid \{t_j\}\right) = \prod_{j=1}^{N} \frac{\Psi_j\left(t_j + \Delta t\right)}{\Psi_j\left(t_j\right)}. \tag{2.6}$$

Thus, by solving $\Phi\left(\Delta t \mid \{t_j\}\right) = u$, where $u \in U(0,1)$ is drawn from a uniform distribution, the time Δt until the next event can be determined. However, finding this solution incurs a high computational cost, and thus Eq. (2.6) can be approximated in the limit of $N \to \infty$, at which Δt is close to 0. By rewriting $\Phi\left(\Delta t \mid \{t_j\}\right)$, we obtain

$$\Phi\left(\Delta t \mid \{t_j\}\right) = \exp\left[-\sum_{j=1}^{N} \ln \frac{\Psi_j\left(t_j\right)}{\Psi_j\left(t_j + \Delta t\right)}\right], \tag{2.7}$$

where $\Psi_j\left(t_j + \Delta t\right)$ can be expanded using the Taylor series, such that for small enough Δt, we get $O(\Delta t^2) \approx 0$, which leads to the **first order Taylor series approximation**:

$$\begin{aligned} &= \exp\left[-\sum_{j=1}^{N} \ln \frac{\Psi_j\left(t_j\right)}{\Psi_j\left(t_j\right) - \psi_j\left(t_j\right)\Delta t + O\left(\Delta t^2\right)}\right] \\ &\approx \exp\left[-\Delta t\left(\sum_{j=1}^{N} \lambda_j(t_j)\right)\right], \end{aligned} \tag{2.8}$$

where the instantaneous rate $\lambda_j(t_j)$ is defined as

$$\lambda_j(t_j) = \frac{\psi_j(t_j)}{\Psi_j(t_j)}.$$

With this approximation, $\Phi\left(\Delta t \mid \{t_j\}\right) = u$ has the solution

$$\Delta t = \frac{\ln(1/u)}{\sum_{j=1}^{N} \lambda_j(t_j)}.$$

□

2.2.2 Second order Taylor series approximation

By considering the quadratic term of the Taylor expansion in the denominator of Eq. (2.7)

$$\Psi_j\left(t_j + \Delta t\right) = \Psi_j\left(t_j\right) - \psi_j\left(t_j\right)\Delta t + \frac{1}{2}\psi_j'\left(t_j\right)\Delta t^2 + O\left(\Delta t^3\right),$$

we obtain an additional term for the approximation

$$\Phi\left(\Delta t \mid \{t_j\}\right) \approx \exp\left[-\Delta t \sum_{j=1}^{N} \lambda_j(t_j) - \frac{1}{2}\Delta t^2 \sum_{j=1}^{N} \lambda_j'(t_j)\right]. \tag{2.9}$$

Solving $\Phi\left(\Delta t \mid \{t_j\}\right) = u$ to determine the next time increment Δt yields

$$\Delta t = \frac{-\sum_{j=1}^{N} \lambda(t_j) + \sqrt{(\sum_{j=1}^{N} \lambda(t_j))^2 - 2\sum_{j=1}^{N} \lambda'(t_j) \cdot \ln(u)}}{\sum_{j=1}^{N} \lambda'(t_j)}. \tag{2.10}$$

Note that in Eq. (2.10) above, we only took the positive solution of the quadratic formula.

2.2.3 Applying different distributions to the non-Markovian Gillespie algorithm

Given the probability density function $\varphi_j(t_i) = \lambda_i(t_i) \cdot \exp\left(-\int_0^{t_i} \lambda_i(\tau) d\tau\right)$ [20], one can recover different wait-time distributions depending on the choice of $\lambda_i(t_i)$.

Exponential distribution

By setting a constant rate $\lambda_i(t_i) = \lambda_0$, we obtain an exponential distribution: $\varphi_i(t) = \lambda_0 \cdot \exp\left(-\lambda_i(t)\right)$, and thus recover the Direct Gillespie algorithm.

Weibull distribution

With $\lambda_i(t) = \beta t^\alpha$ we obtain a Weibull distribution: $\varphi_i(t) = \beta t^\alpha \cdot \exp\left(-\frac{\beta t^{\alpha+1}}{\alpha+1}\right)$. Details on how the Weibull distribution is parametrised for the purposes of this thesis are provided in Section 2.4.

2.3 Non-Markovian Rejection Gillespie Algorithm (NMRG) with individual reactant properties

The presence of a high number of processes involved in the Gillespie simulation leads to a high computational cost, particularly as the propensity needs to be recalculated for each individual reactant at every iteration, since individual properties are being tracked. To minimise the number of propensity updates, we set an upper propensity bound for each reaction of the same type, and postpone updating the propensities until the rejection step. We can then accept the reaction based on its individual reaction rate, or reject it by exploiting the Poisson thinning process.

More concretely, for each reaction channel i, let there be:

i A set of reactants $\{\mu_{i,1} \dots \mu_{i,N}\} \in \mathcal{A}_i$, each with an individual 'time since last firing'

ii A set of time-dependent reaction rates $\{r_{i,1}(\mu_{i,1}) \dots r_{i,N}(\mu_{i,N})\} \in \mathcal{R}_i$ based on the individual properties of the reactants in $\mathcal{A}_j$

iii The number of reactants involved in reaction i, $N_i = |\mathcal{A}_i|$

We then set the upper bound for each reaction channel based on the reactant with the highest individual rate

$$r_{i,\max} = \max_{\{\mu_{i,1} \dots \mu_{i,N}\} \in \mathcal{A}_i} r_i(\mu_{i,1} \dots \mu_{i,N}), \tag{2.11}$$

compute the reaction propensity with the upper bound

$$\lambda_{i,\max} = N_i \cdot r_{i,\max}, \tag{2.12}$$

compute the total propensity

$$\lambda_{0,\max} = \sum_{i=1}^{N} \lambda_{i,\max}, \tag{2.13}$$

select the next time increment

$$\Delta t = \frac{\ln(1/u_1)}{\lambda_{0,\max}}, \tag{2.14}$$

choose reaction channel i with probability p_i

$$p_i = \frac{N_i \cdot r_{i,\max}}{\lambda_{0,\max}}, \tag{2.15}$$

and finally, we decide whether to accept the individual selected reactant μ_i (or otherwise reject) with a probability of

$$p_{\text{accept}} = \frac{\lambda_i}{\lambda_{i,\max}}. \tag{2.16}$$

Rejecting a reaction is equivalent to adding an additional empty Poisson reaction channel, which is still consistent with DG (see Theorem 2.2 and Theorem 2.3).

The pseudo-code for implementation of the NMRG model used in this thesis is described in Algorithm 2.

Tracking individual properties of reactants

The NMRG model relies on tracking t_i - the time since the last firing of reactant μ_i, which will affect the individual reaction rates based on the choice of the function for $r_i(t_i)$. Here we have chosen the instantaneous rate to be proportional to the power of time,

$$r_i(t_i) = \beta t_i^{\alpha} \ (0 \leq \alpha),$$

such that its probability density function corresponds to the Weibull distribution (described in Section 2.4)

$$\varphi_i(t) = \beta t^{\alpha} \cdot \exp\left(-\frac{\beta t^{\alpha+1}}{\alpha+1}\right).$$

Hence, the longer the reactant μ_i had spent without reacting, the higher its individual rate.

Non-Markovian Rejection Gillespie follows the correct distribution

The following proof has been adapted from Thanh and Priami [25]. The NMRG algorithm follows the same distribution as the Direct Gillespie algorithm (DG) if we can show that:

i Reaction R_j is accepted with probability $p_{\text{accept}}(R_j \mid R) = \lambda_j/\lambda_0$

ii The time increment Δt follows the same exponential distribution as in DG, $f_{\Delta t}(x) = \lambda_0 \cdot \exp(-\lambda_0 x)$

Theorem 2.2 *The reaction R_j in NMRG is accepted with the same probability as in the Direct Gillespie*

Proof We define $p_{\text{accept}}(R_j)$ as the joint probability between R_j being first selected and then accepted, $\lambda_{j,\max}$ as the upper propensity bound for reaction R_j , and $\lambda_{0,\max} = \sum_{j=1}^{N} \lambda_{j,\max}$:

$$p_{\text{accept}}(R_j) = \frac{\lambda_{j,\max}}{\lambda_{0,\max}} \times \frac{\lambda_j}{\lambda_{j,\max}} = \frac{\lambda_j}{\lambda_{0,\max}}$$

We then denote by $p_{\text{accept}}(R)$ the probability of some reaction being accepted

$$p_{\text{accept}}(R) = \frac{\lambda_0}{\lambda_{0,\max}}.$$

The conditional probability $p_{\text{accept}}(R_j \mid R)$ can be exploited to show the probability of R_j being accepted given that some reaction had been accepted:

$$p_{\text{accept}}(R_j \mid R) = \left(\frac{\lambda_j}{\lambda_{0,\max}}\right) / \left(\frac{\lambda_0}{\lambda_{0,\max}}\right) = \frac{\lambda_j}{\lambda_0} \qquad \square$$

Theorem 2.3 *The time increment Δt in NMRG has an exponential probability density function*

Proof We denote by k the number of rejections until a reaction is accepted, with time being advanced by $\Delta t = -\frac{\ln(u)}{\lambda_{0,\max}}$ after each rejection. It follows that for k attempts,

$$\Delta t = -\frac{1}{\lambda_{0,\max}} \ln\left(\prod_{i=1}^{k} u_i\right), \tag{2.17}$$

which corresponds to the Erlang distribution. In addition, k is geometrically distributed with probability $p_{\text{accept}}(R)$, i.e. $P(X = k) = \big(1 - p_{\text{accept}}(R)\big)^{k-1} \times p_{\text{accept}}(R)$.

The PDF for Δt can be expressed as the derivative of its CDF, using the Fundamental theorem of calculus

$$\begin{aligned} f_{\Delta t}(x) &= \frac{d}{dx} F_{\Delta t}(x) \\ &= \frac{d}{dx} P(\Delta t \leq x), \end{aligned}$$

where $P(\Delta t \leq x)$ can be partitioned for values of k:

$$\begin{aligned} &= \frac{d}{dx} \sum_{k=1}^{\infty} P(\Delta t \leq x \mid k = k) P(k = k) \\ &= \frac{d}{dx} \sum_{k=1}^{\infty} P(\Delta t \leq x \mid k = k) \left(1 - \frac{\lambda_0}{\lambda_{0,\max}}\right)^{k-1} \cdot \frac{\lambda_0}{\lambda_{0,\max}}, \end{aligned}$$

and as shown in Eq. (2.17), the distribution of Δt parametrised by k follows an Erlang distribution:

$$\begin{aligned} &= \sum_{k=1}^{\infty} \frac{d}{dx} F_{\text{Erlang}(k,\lambda_{0,\max})} \left(1 - \frac{\lambda_0}{\lambda_{0,\max}}\right)^{k-1} \cdot \frac{\lambda_0}{\lambda_{0,\max}} \\ &= \sum_{k=1}^{\infty} f_{\text{Erlang}(k,\lambda_{0,\max})} \left(1 - \frac{\lambda_0}{\lambda_{0,\max}}\right)^{k-1} \cdot \frac{\lambda_0}{\lambda_{0,\max}} \\ &= \sum_{k=1}^{\infty} \frac{\lambda_{0,\max}^k \cdot x^{k-1} \cdot \exp(-\lambda_{0,\max} x)}{(k-1)!} \cdot \left(\frac{\lambda_{0,\max} - \lambda_0}{\lambda_{0,\max}}\right)^{k-1} \cdot \frac{\lambda_0}{\lambda_{0,\max}} \\ &= \lambda_0 \exp(-\lambda_{0,\max} x) \sum_{k=1}^{\infty} \frac{(\lambda_{0,\max} - \lambda_0)^{k-1} \cdot x^{k-1}}{(k-1)!} \\ &= \lambda_0 \exp(-\lambda_{0,\max} x) \cdot \exp(x(\lambda_{0,\max} - \lambda_0)) = \lambda_0 \cdot \exp(-\lambda_0 x). \end{aligned}$$

Hence, in the non-Markovian Rejection Gillespie algorithm, Δt follows an exponential distribution. □

Algorithm 2 Non-Markovian Rejection Gillespie with individual reactant properties

Time $t \leftarrow 0$
Initialise list of reactants
while $t < t_{end}$ **do**
  Calculate upper bound for each reaction channel $r_{i,\max}$
  Compute total propensity $\lambda_{0,\max} = \sum_{i=1}^{N} N_i \cdot r_{i,\max}$
  Generate uniform random numbers $\{u_1, u_2, u_3\} \in U(0,1)$
  Compute time increment $\Delta t = \frac{\ln(1/u_1)}{\lambda_{0,\max}}$
  $t \leftarrow t + \Delta t$
  Select the firing of reaction i with probability $p_i = \frac{\lambda_{i,\max}}{\lambda_{0,\max}}$
  Randomly select reactant μ_i from $\mathcal{A}_i$
  if $r_i(\mu_i) > r_{i,\max} \cdot u_2$ **then**
    Update list of reactants according to stoichiometry of reaction i
  end if
end while

2.4 Weibull Distribution

The Weibull distribution has the ability to assume the characteristics of different types of commonly used distributions. As such, it can interpolate between the exponential and Rayleigh distributions, as well as model symmetrical (normal), log-normal, left-skewed, and right-skewed data. Its probability density function is commonly parametrised as

$$f(x;\lambda,k) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \times \exp^{-(x/\lambda)^k}, \tag{2.18}$$

where $k > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter. Its mean is given by

$$\mathrm{E}[X] = \lambda\, \Gamma\left(1 + \frac{1}{k}\right), \tag{2.19}$$

and variance given by

$$\mathrm{Var}(X) = \lambda^2 \left[\Gamma\left(1 + \frac{2}{k}\right) - \Gamma^2\left(1 + \frac{1}{k}\right)\right]. \tag{2.20}$$

By setting $k = 1$ we recover the exponential distribution.

Weibull distribution for time-dependent reaction rates

For the purposes of this thesis, we aim to parametrise the Weibull distribution for the time-dependent reaction rates, such that its mean remains constant over different values of the shape parameter k. Thus, we define

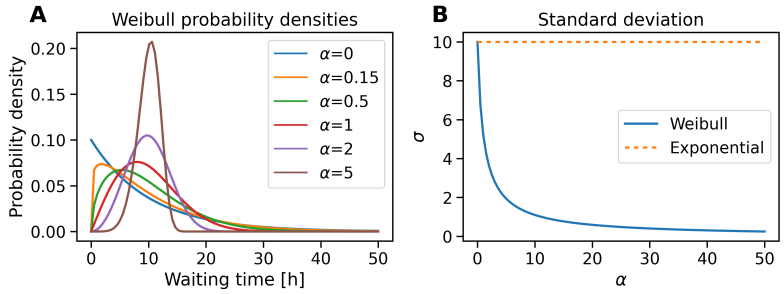

Figure 2.1: **Weibull distribution for different values of α.** (**A**) The probability density of the Weibull distribution shifts away from exponential as α increases. (**B**) The standard deviation decreases as α increases. The orange dashed line shows the constant standard deviation for an exponential distribution ($\alpha = 0$) used in the Direct Gillespie model. Both plots were generated for $r_0 = 0.1\ \mathrm{h}^{-1}$.

$\lambda = \left(\frac{\alpha+1}{\beta}\right)^{\frac{1}{\alpha+1}}$, $\beta = (\alpha+1)\left[r_0\Gamma\left(\frac{\alpha+2}{\alpha+1}\right)\right]^{\alpha+1}$, and $k = \alpha + 1$ with $(0 \leq \alpha)$. This results in the following probability density function with the independent time variable t,

$$f\left(t; \left(\frac{\alpha+1}{\beta}\right)^{\frac{1}{\alpha+1}}, \alpha+1\right) = \beta t^{\alpha} \times \exp\left(-\frac{\beta t^{\alpha+1}}{\alpha+1}\right),$$

with a mean

$$\mathrm{E}[T] = \frac{1}{r_0},$$

where r_0 is the rate hyperparameter. The variance is given by

$$\mathrm{Var}(T) = \frac{1}{r_0^2}\left(\frac{\Gamma\left(\frac{\alpha+3}{\alpha+1}\right)}{\Gamma^2\left(\frac{\alpha+2}{\alpha+1}\right)} - 1\right).$$

As visualised in Figure 2.1, we can tune the shape and standard deviation of the wait-time distribution with the new shape parameter α and the mean with the rate parameter r_0.

2.5 Ordered heap for algorithm speedup

Starting from Python 3.7, its dictionary implementation is guaranteed to keep the insertion order. This feature can be utilised to maintain a max-heap-like data structure for the purposes of the NMRG method. This serves to speed up the $r_{i,\max}$ calculation step in Eq. (2.11).

Recall that we have defined the time-dependent rate as $r_i(t_i) = \beta t_i^{\alpha}$. Hence, $r_{i,\max}$ depends on the reactant with highest wait-time (time spent without reacting), $\max_{\{t_1 \ldots t_N\}} t_i$. Finding a maximum value in a list or dictionary

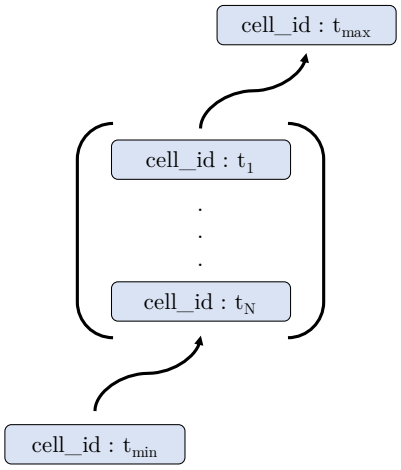

Figure 2.2: **Diagram of a max-heap data structure implemented in NMRG.** New cells added to the bottom of the heap are guaranteed to have the smallest t_i, while cells at the top of the heap will have the highest t_i. Lookup, insertion, and deletion are all operations with a constant cost $O(1)$

incurs a cost of $O(N)$, whereas a max-heap maintains a constant cost of $O(1)$. Adding a new entry to the max-heap is also constant in $O(1)$.

As illustrated in Figure 2.2, each entry in the max-heap consists of a unique ID for each cell, coupled with its value of t_i. When a new cell enters the reaction channel, it is added to the bottom of the heap. As the algorithm updates only a single reactant at a time, its t_i is guaranteed to be the lowest in the system, as it had just reacted. Conversely, the cell at the top of the heap is guaranteed to have the highest t_i. Once a cell in the reaction channel had reacted, it is removed from the max-heap by looking up its unique ID with a constant cost of $O(1)$.

2.6 Hyperparameter optimisation

The α and r_0 hyperparameters of the inter-event time distribution for each reaction channel were tuned using the MaxLIPO method from the Dlib C++ library [13]. The parameter search was constrained to the space of $(0.0100 \leq r_0 \leq 0.0800) \in \mathbb{R}$ and $(0 \leq \alpha \leq 6) \in \mathbb{R}$. Fitting of the data was evaluated by minimising the root mean square error (RMSE). Tuning was run for 10 hours in an HPC environment, corresponding to a minimum of 3600 iterations.

2.7 Code availability

The model was implemented in Python and is publicly available on the GitHub repository: https://github.com/mirophan/nmrg.

2.8 Dataset used for model evaluation

The mouse embryonic stem cell dataset was obtained from Stumpf and colleagues [24]. It consists of two biological replicates, each derived from mice with a different genetic background, R1 [19] and E14 (E14tg2a) [6][14], respectively. Using a well established protocol [30], the cells were directed to undergo 3 distinct cell states for neuroectoderm differentiation. Individual cells were FACS sorted and their gene expression changes recorded using a high-throughput RT-PCR array over the course of 168 hours in order to obtain a time-series data of cell states. The cell state labels were classified via k-means clustering with 3 clusters. The data were personally communicated by the authors.

Chapter 3

Results

This section presents the performance of the proposed non-Markovian Rejection Gillespie model (NMRG). The simulations are benchmarked against the Direct Gillespie model (DG) as a baseline.

As ground truth, an experimental dataset from Stumpf and colleagues [24] was obtained. It consists of a time series for the differentiation of mouse embryonic stem cells with known non-Markovian dynamics. The cells were directed to undergo 3 distinct cell states in the early stages of neuroectodermal development [1][4], initially starting as an embryonic stem cell (ESC) and passing through an epiblast (EPI) state before finally committing to a neuronal stem cell (NPC) as shown in Figure 3.1. In order to capture general rather than cell-line specific trends, the experiment was conducted for two biological replicates, with ESCs derived from mice with two distinct genetic backgrounds, R1 and E14.

A hyperparameter search (detailed in Section 2.6) was performed for both models in order to optimise their fit to the experimental data. The values found by the optimiser are listed in Tables A.1 and A.2.

3.1 Fitting Direct Gillespie to data

The Direct Gillespie model (DG) (detailed in Section 2.1) assumes that individual cells are not influenced by their past, stochastically undergo differentiation at a constant rate, and thus their reaction waiting-times follow an exponential distribution. This Markovian assumption is necessary for the simplification and mathematical tractability of the model.

As shown in Figure 3.2, the DG model does not describe the experimental data well. The poor fit is evidenced by the root mean squared error (RMSE) of 0.243 for R1 and 0.236 for E14 cell lines (averaged over all three cell types), which is higher when compared to the non-Markovian approach

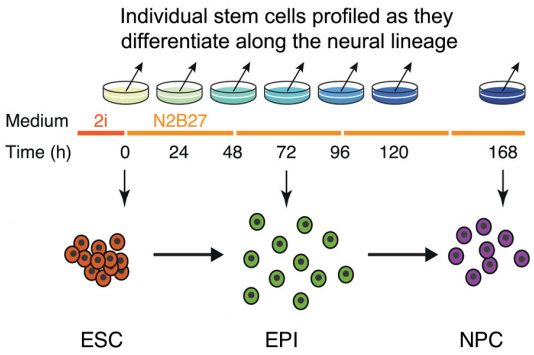

Figure 3.1: **Experimental design of the ground truth dataset.** Mouse embryonic stem cells were directed to undergo differentiation towards the neuroectoderm. Single cell gene expression was profiled at each time step. *Figure adapted from Stumpf and colleagues [24] under the Creative Commons Attribution (CC BY 4.0). Copyright 2007 by Stumpf and colleagues*

presented in Figure 3.3. The histograms in Figure 3.2B show that the exponential waiting-time distributions were correctly recovered for both differentiation processes (ESC $\rightarrow$ EPI, EPI $\rightarrow$ NPC). As input, the DG model only required two parameters, the average reaction rates r_{esc} and r_{epi} (listed in Table A.1), for each process.

3.2 Fitting non-Markovian Rejection Gillespie to data

The non-Markovian Rejection Gillespie model (NMRG) (Section 2.3) accounts for individual cell properties by tracking their time spent without reacting, t_i. In contrast to the DG model, the individual reaction rates are now proportional to the power of the waiting-time, $r_i(t_i) = \beta t_i^{\alpha}$ $(0 \leq \alpha)$, such that their probability density function follows a Weibull distribution (described in Section 2.4).

The NMRG model was able to better capture the differentiation dynamics of the experimental data, as shown in Figure 3.3, with the average RMSE of 0.062 and 0.110 for the R1 and E14 cell lines, respectively. The Weibull wait-time distributions were correctly recovered for both differentiation processes (ESC $\rightarrow$ EPI, EPI $\rightarrow$ NPC), as evidenced in Figure 3.3B, whereby the simulation probability density (PDF) matched the theoretical PDF calculated by inputting parameters from A.2 into 2.18. The distributions gave a mean residence of 31.1h in the ESC and 59.7h in the EPI states.

The NMRG model requires a rate and shape parameter for each differentiation process, for a total of 4 parameters: $r_{0,\text{esc}}$, $r_{0,\text{epi}}$, α_{esc}, α_{epi}, listed in Table A.2. The addition of the α parameters effectively allowed the modu-

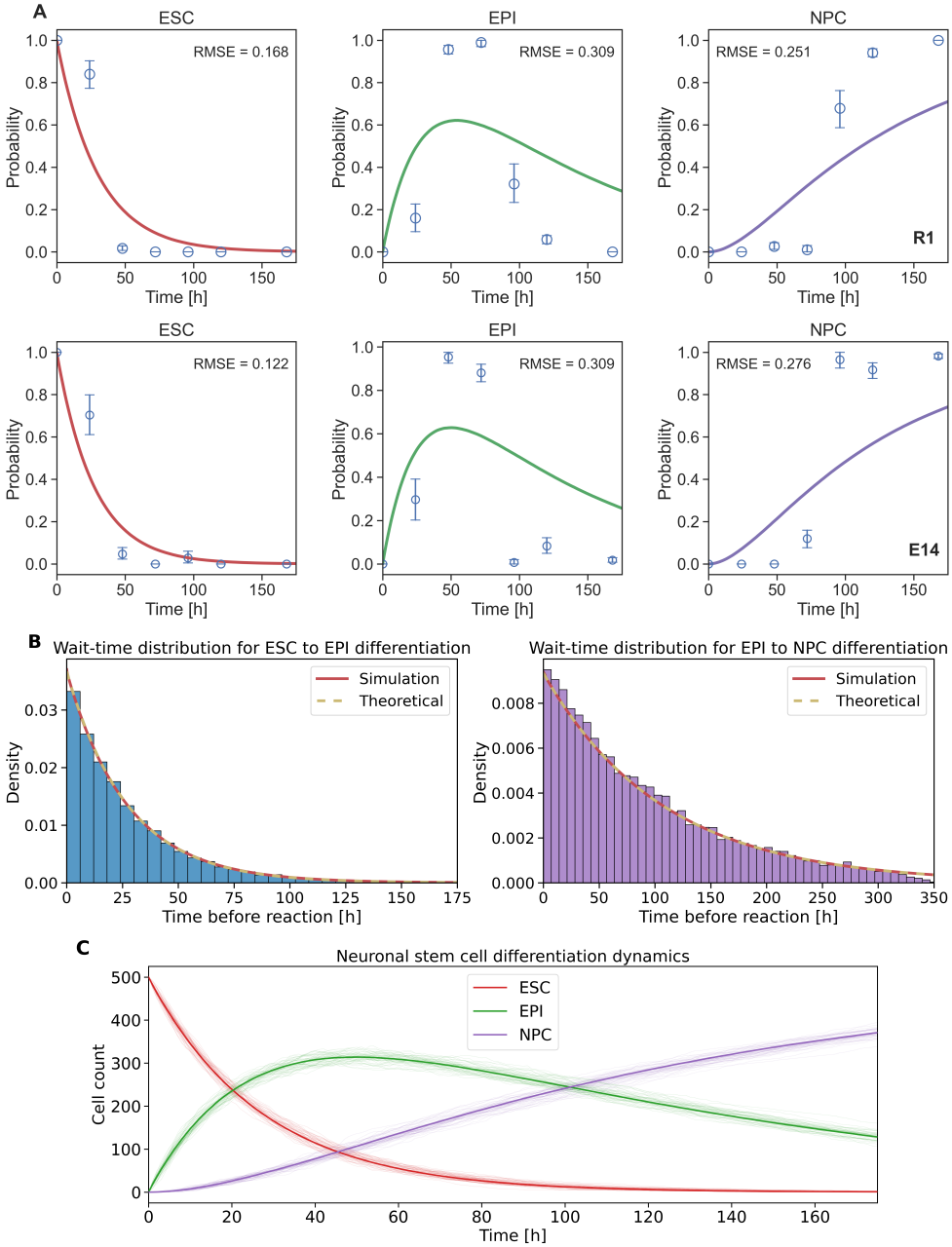

Figure 3.2: **Fitting the Direct Gillespie model to data.** (**A**) Experimental data represented by blue circles with 95% confidence intervals calculated from bootstrapped k-means clustering. Cell counts from the simulation were L1 normalised. The model was fit to two separate of mouse embryonic stem cells lineages, R1 and E14. Root mean squared error (RMSE) reported for simulation fit to data. (**B**) Histograms for the waiting-time distributions of ESC and EPI cell firing of E14 simulation. The resulting probability density function (PDF) corresponds to a Weibull distribution with a mean residence of 27.2h in ESC and 90.9h in EPI states. The theoretical PDF was calculated according to the initial simulation parameters α and r_0 in Table A.1. (**C**) Population dynamics averaged over all simulations (bold lines), as well as individual simulation trajectories (thin lines).

lation of the shape and variance of the reaction waiting-times (Figure 2.1).

3.3 Algorithm runtime improvement

The Gillespie algorithm in general is a powerful approach for modelling the dynamics of large stochastic systems. However, in the context of modelling non-Markovian systems, the Direct Gillespie (DG), as well as other proposed methods [3][27] are hindered by a high computational cost. In the Markovian implementation, the runtime complexity for the DG method follows the order of $O(M)$ [23], where M is the number of reaction channels. For a non-Markovian system with a population size of N, where each reactant has its own reaction rate, DG needs to maintain $M = N$ reaction channels, and thus the complexity increases to $O(N)$.

In contrast, the non-Markovian Rejection Gillespie (NMRG) implementation is highly computationally efficient as a result of two key implementations; first, the rejection step outlined in Section 2.3 reduces the number of necessary propensity updates, and second, an ordered heap data structure (Section 2.5) is maintained for storing the $r_{i,\max}$ values for Eq. (2.11), thus minimising the number of reaction rate updates. The runtime complexity is in the order of $O(M + R)$, where R corresponds to the number of rejections. The individual reactants are pooled into M reaction channels, and thus for general simulations it can be assumed that $M \ll N$. As the population size increases, the number of rejections R increases. In practice, a rejection step only involves drawing a random number, whereas simulating a whole reaction would involve updating multiple lists, incurring a higher computational cost.

Figure 3.4 concretely shows the differences in runtimes between the DG (modified to allow for non-Markovian properties according to the work done by Boguñá and colleagues [3]) and NMRG methods, where the number of reaction channels $M = 2$ (ESC $\rightarrow$ EPI, EPI $\rightarrow$ NPC) is fixed and the population size is gradually increased on a $\log_e$ scale. The NMRG method is evidently more efficient and scalable to larger systems.

3.4 Reducing wait-time approximation error

The non-Markovian extension to the Gillespie proposed by Boguñá and colleagues [3] does not hold for systems with small population sizes N. As described in Section 2.2.1, the Taylor series in Eq. (2.7) is approximated for a small Δt. In order to improve on the approximation, an additional quadratic term was added to the expansion, as described in Section 2.2.2. In the case of the Direct Gillespie with constant reaction rates, the quadratic term in

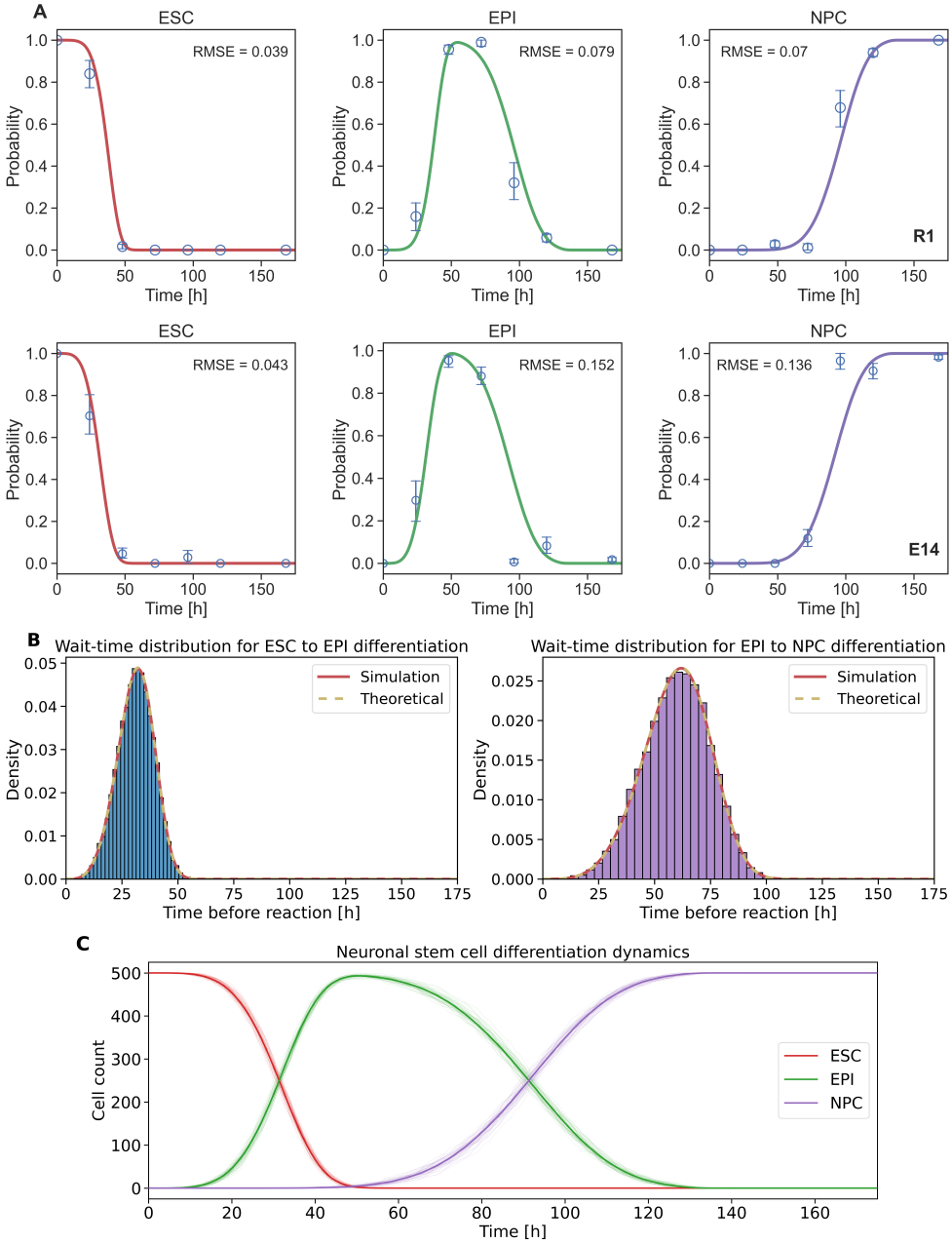

Figure 3.3: **Fitting the non-Markovian Rejection Gillespie model to data.** (**A**) Experimental data represented by blue circles with 95% confidence intervals calculated from bootstrapped k-means clustering. Cell counts from the simulation were L1 normalised. The model was fit to two separate of mouse embryonic stem cells lineages, R1 and E14. Root mean squared error (RMSE) reported for simulation fit to data. (**B**) Histograms for the waiting-time distributions of ESC and EPI cell firing (for E14). The resulting probability density function (PDF) corresponds to a Weibull distribution with a mean residence of 31.1h in ESC and 59.7h in EPI states. The theoretical PDF was calculated according to the initial simulation parameters α and r_0 in Table A.2. (**C**) Population dynamics averaged over all simulations (bold lines), as well as individual simulation trajectories (thin lines).

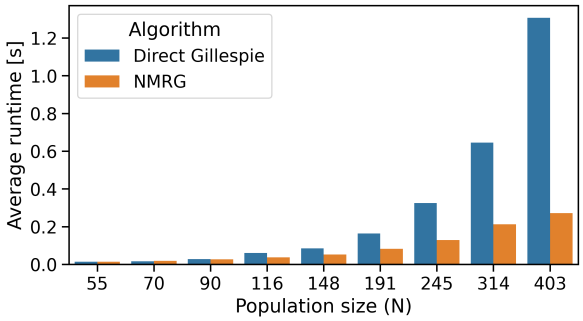

Figure 3.4: **Runtime comparison between the DG and NMRG algorithms.** The Direct Gillespie was modified to allow for non-Markovian properties according to the work done by Boguñá and colleagues [3]. The average runtime corresponds to one simulation episode, averaged over 100 simulations. Population sizes (N) for the simulations were selected on a $\log_e$ scale.

Eq. (2.9) is reduced to zero, whereas implementations with time-dependent reaction rates will have a non-zero term, thus motivating the proposed improvement.

Histograms from Figure 3.3B were used for the error evaluation, whereby the differences between the theoretical and simulation distributions were compared. Earth mover's distance (EMD) [22] was used as a metric to compare the distance between the two distributions.

Figure 3.5 shows the approximation error for simulations over increasing population sizes (N). The first order Taylor approximation (T1) corresponds to Eq. (2.8) (only linear term considered), whereas the second order Taylor approximation (T2) corresponds to Eq. (2.10) (both linear and quadratic terms considered). In the cases of very low population sizes, T2 error is indeed lower than T1 error. However, the gap is quickly closed as the population size increases past $N = 7$, with T1 error dipping below T2. Eventually past the $N = 148$ mark, both T1 and T2 errors converge and tend towards 0.

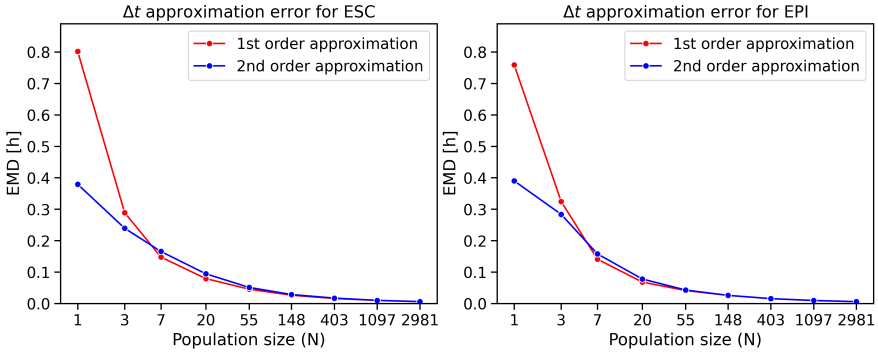

Figure 3.5: **Wait-time approximation error decreases with increasing population size.** The first order Taylor approximation (T1) corresponds to Eq. (2.8) (only linear term considered). The second order Taylor approximation (T2) corresponds to Eq. (2.10) (both linear and quadratic terms considered). The Earth mover's distance (EMD) metric was averaged over 500 simulations for each population size. Population sizes (N) were selected on a $\log_e$ scale.

Chapter 4

Discussion

Stochastic processes found both within and beyond biochemical systems are known to be governed by non-Markovian dynamics as described in Chapter 1. In order to quantitatively understand these processes, a non-Markovian Rejection Gillespie (NMRG) model is proposed.

Due to the inherent complexities of non-Markovian dynamics governing each individual reactant in a system, their analytical solutions are difficult to formulate, and are often not applicable to general systems [32]. Therefore, the model was evaluated against a ground truth dataset of mouse embryonic stem cell time series with known non-Markovian characteristics [24].

The NMRG model improves on the Direct Gillespie (DG) model in correctly recapitulating the experimental data shown in Figure 3.3. By modulating the reaction waiting-time distribution, NMRG could force a peak accumulation of EPI cells between the 48h and 72h marks, followed by a rise in NPC cells at 100h, in congruence with *in vivo* neuroectoderm development at embryonic day 7.5 [18]. The reported RMSE values are lower when fitting the NMRG model to the R1 cell line data compared to the E14 data. This could be explained by E14 containing biologically unrealistic data points where $N_{\text{ESC}}(t = 72) = 0$ and $N_{\text{ESC}}(t = 96) = 0$ are both followed by an increase in cell count in the next time point. The directed differentiation is strictly linear according to the experimental design in Figure 3.1 and does not allow for spontaneous birth events. These zero counts are an artefact of data pre-processing, where cells missing house-keeping gene reads had been removed.

The method is computationally efficient and scalable as seen in Figure 3.4. Hence, the reported average runtime of 0.2 seconds per simulation of size $N = 400$ reactants scales up to 18 000 simulations per hour. For cases involving simulations with a low number of reactants, a more accurate second order Taylor series approximation was developed. However, the improve-

ment is only useful for very low population counts of $(1 \leq N < 7)$, after which the improvement is slightly worse or negligible when compared to the first order approximation.

NMRG only requires two parameters for each reaction channel, namely the rate r_0 and the shape α, allowing for easy hyperparameter optimisation. For modelling unknown systems, the parameters will need to be inferred from existing literature. To that end, $1/r_0$ reflects the mean, while α directly reflects the standard deviation of the event waiting-times, which are measured regularly when studying biological processes. Having a distribution with two parameters makes it interpretable and intuitive, and thus unlike Hidden Markov models, NMRG avoids overfitting with too many parameters.

Although the NMRG model performed well on the small embryonic stem cell dataset with a simple differentiation model, it would be valuable to see the model applied to a larger scale system with reactants involved in multiple reaction channels at once. An ideal dataset could be obtained by performing a time series of single-cell RNA sequencing on a biological system with non-Markovian dynamics.

The Weibull distribution is highly flexible and can modulate its shape. However, an alternative probability distribution, such as the Gaussian, may be more appropriate in the context of different systems. In fact, the time-dependent rate function $r_i(t_i)$ can be modified to change its probability density function. However, specifically in the case of the Gaussian, the rate function is difficult to formulate. Future efforts could start by further extending the NMRG model to be exact for normally distributed waiting-times.

An attempt has been made to reduce the wait-time approximation error as seen in Figure 3.5. However, the improvement is limited to extremely small population sizes, and in some cases even leads to worse performance. To further investigate the potential for improvement, one would need to calculate the exact Δt in Eq. (2.6) and compare it to the Δt output by the second order Taylor series approximation in Eq. (2.9). It would also be useful to explore why the second order approximation sometimes performs worse than the supposedly less accurate first order approximation.

In addition to the aforementioned applications to biological systems, non-Markovian modelling can also be useful for studying human behavioural patterns. These inherently follow non-exponential distributions, which can potentially be well modelled by NMRG. Thus, the ongoing efforts to model epidemic spreading dynamics can be enhanced by factoring in differences in human behaviour, such as compliance to public measures, movement within one's social network, or threat perception.

In summary, the non-Markovian Rejection Gillespie is a versatile model with the potential to be applicable to many real-world systems governed by

non-Markovian dynamic processes. The promising results presented on the small stem cell model build a path for further development of the approach towards better understanding of systems governing living matter.

Chapter 5

Acknowledgements

I am extremely thankful to my parents for their unrelenting support throughout this Master's degree and beyond.

Over the course of this thesis, I received a great deal of support and assistance. I would like to thank Professor Niko Beerenwinkel for his supervision and advice. I am also grateful to my supervisor Aurélien Pélissier, whose guidance was instrumental to the success of this thesis. Equally, I would also like to extend my graditute to Dr. María Rodríguez Martínez for her expert advice and supervision. Finally, I would like to thank Dr. Patrick Stumpf and his colleagues for kindly providing the dataset used in this thesis [24].

Appendix A

Appendix

A.1 Hyperparameters used in simulations

Table A.1: Hypeparameters for Standard Gillespie

	Cell lineage	
Parameters	**R1**	**E14**
$r_{0,\text{esc}}$	0.03325	0.03668
$r_{0,\text{epi}}$	0.00873	0.00933
α_{esc}	0	0
α_{epi}	0	0

Table A.2: Hypeparameters for non-Markovian Rejection Gillespie

	Cell lineage	
Parameters	**R1**	**E14**
$r_{0,\text{esc}}$	0.02749	0.03217
$r_{0,\text{epi}}$	0.01708	0.01675
α_{esc}	4.53829	3.41100
α_{epi}	3.59621	3.61365

Bibliography

[1] Elsa Abranches, Margarida Silva, Laurent Pradier, Herbert Schulz, Oliver Hummel, Domingos Henrique, and Evguenia Bekman. Neural Differentiation of Embryonic Stem Cells In Vitro: A Road Map to Neurogenesis in the Embryo. *PLOS ONE*, 4(7):e6286, July 2009. Publisher: Public Library of Science.

[2] Albert-László Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, May 2005. Number: 7039 Publisher: Nature Publishing Group.

[3] Marian Boguñá, Luis F. Lafuerza, Raúl Toral, and M. Ángeles Serrano. Simulating non-Markovian stochastic processes. *Physical Review E*, 90(4):042108, October 2014. Publisher: American Physical Society.

[4] Thorsten Boroviak, Remco Loos, Paul Bertone, Austin Smith, and Jennifer Nichols. The ability of inner-cell-mass cells to self-renew as embryonic stem cells is acquired following epiblast specification. *Nature Cell Biology*, 16(6):516–528, June 2014.

[5] Alvaro Corral. Long-term clustering, scaling, and universality in the temporal occurrence of earthquakes. *Physical Review Letters*, 92(10):108501, March 2004.

[6] Thomas Doetschman, Ronald G. Gregg, Nobuyo Maeda, Martin L. Hooper, David W. Melton, Simon Thompson, and Oliver Smithies. Targetted correction of a mutant HPRT gene in mouse embryonic stem cells. *Nature*, 330(6148):576–578, December 1987. Number: 6148 Publisher: Nature Publishing Group.

[7] Mark R Dowling, Dejan Milutinović, and Philip D Hodgkin. Modelling cell lifespan and proliferation: is likelihood to die or to divide indepen-

dent of age? *Journal of the Royal Society Interface*, 2(5):517–526, December 2005.

[8] Mi Feng, Shi-Min Cai, Ming Tang, and Ying-Cheng Lai. Equivalence and its invalidation between non-Markovian and Markovian spreading dynamics on complex networks. *Nature Communications*, 10(1):3748, August 2019. Number: 1 Publisher: Nature Publishing Group.

[9] Vladimir Filimonov and Didier Sornette. Quantifying reflexivity in financial markets: Toward a prediction of flash crashes. *Physical Review E*, 85(5):056108, May 2012.

[10] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, December 1977. Publisher: American Chemical Society.

[11] Micha Gladnikoff and Itay Rousso. Directly Monitoring Individual Retrovirus Budding Events Using Atomic Force Microscopy. *Biophysical Journal*, 94(1):320–326, January 2008.

[12] N. G. Van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier, November 1992. Google-Books-ID: 3e7XbMoJzmoC.

[13] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(60):1755–1758, 2009.

[14] Michael R. Kuehn, Allan Bradley, Elizabeth J. Robertson, and Martin J. Evans. A potential animal model for Lesch–Nyhan syndrome through introduction of HPRT mutations into mice. *Nature*, 326(6110):295–298, March 1987. Number: 6110 Publisher: Nature Publishing Group.

[15] P. a. W. Lewis and G. S. Shedler. Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics Quarterly*, 26(3):403–413, 1979. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/nav.3800260304.

[16] Nancy R. Mann, Nozer D. Singpurwalla, and Ray E. Schafer. Methods for statistical analysis of reliability and life data, 1974. Publisher: Wiley.

[17] Naoki Masuda and Luis E. C. Rocha. A Gillespie Algorithm for Non-Markovian Stochastic Processes. *SIAM Review*, 60(1):95–115, January 2018.

[18] Sophie Morgani, Jennifer Nichols, and Anna-Katerina Hadjantonakis. The many faces of Pluripotency: in vitro adaptations of a continuum of in vivo states. *BMC Developmental Biology*, 17(1):7, June 2017.

[19] A. Nagy, J. Rossant, R. Nagy, W. Abramow-Newerly, and J. C. Roder. Derivation of completely cell culture-derived mice from early-passage embryonic stem cells. *Proceedings of the National Academy of Sciences*, 90(18):8424–8428, September 1993. Publisher: National Academy of Sciences Section: Research Article.

[20] A. Papoulis and H. Saunders. Probability, Random Variables and Stochastic Processes (2nd Edition). *Journal of Vibration and Acoustics*, 111(1):123–125, January 1989.

[21] Etienne Pardoux. *Markov Processes and Applications: Algorithms, Networks, Genome and Finance*. John Wiley & Sons, November 2008. Google-Books-ID: 4Kcmibtb2dUC.

[22] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2):99–121, November 2000.

[23] Kevin R. Sanft and Hans G. Othmer. Constant-complexity stochastic simulation algorithm with optimal binning. *The Journal of Chemical Physics*, 143(7), August 2015.

[24] Patrick S. Stumpf, Rosanna C. G. Smith, Michael Lenz, Andreas Schuppert, Franz-Josef Müller, Ann Babtie, Thalia E. Chan, Michael P. H. Stumpf, Colin P. Please, Sam D. Howison, Fumio Arai, and Ben D. MacArthur. Stem Cell Differentiation as a Non-Markov Stochastic Process. *Cell Systems*, 5(3):268–282.e7, September 2017.

[25] Vo Hong Thanh, Corrado Priami, and Roberto Zunino. Efficient rejection-based simulation of biochemical reactions with stochastic noise and delays. *The Journal of Chemical Physics*, 141(13):134116, October 2014. Publisher: American Institute of Physics.

[26] P. Van Mieghem and R. van de Bovenkamp. Non-Markovian Infection Spread Dramatically Alters the Susceptible-Infected-Susceptible Epidemic Threshold in Networks. *Physical Review Letters*, 110(10):108701, March 2013. Publisher: American Physical Society.

[27] Christian L. Vestergaard and Mathieu Génois. Temporal Gillespie Algorithm: Fast Simulation of Contagion Processes on Time-Varying Networks. *PLOS Computational Biology*, 11(10):e1004579, October 2015. Publisher: Public Library of Science.

[28] Margaritis Voliotis, Philipp Thomas, Ramon Grima, and Clive G. Bowsher. Stochastic Simulation of Biomolecular Networks in Dynamic Environments. *PLOS Computational Biology*, 12(6):e1004923, June 2016. Publisher: Public Library of Science.

[29] Christian A. Yates, Matthew J. Ford, and Richard L. Mort. A Multi-stage Representation of Cell Proliferation as a Markov Process. *Bulletin of Mathematical Biology*, 79(12):2905–2928, December 2017.

[30] Qi-Long Ying, Marios Stavridis, Dean Griffiths, Meng Li, and Austin Smith. Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture. *Nature Biotechnology*, 21(2):183–186, February 2003.

[31] Christoph Zechner and Heinz Koeppl. Uncoupled Analysis of Stochastic Reaction Networks in Fluctuating Environments. *PLOS Computational Biology*, 10(12):e1003942, December 2014. Publisher: Public Library of Science.

[32] Jiajun Zhang and Tianshou Zhou. Markovian approaches to modeling intracellular reaction processes with molecular memory. *Proceedings of the National Academy of Sciences*, 116(47):23542–23550, November 2019. Publisher: National Academy of Sciences Section: Biological Sciences.