# Approximation of Probabilistic Reachability for Chemical Reaction Networks using the Linear Noise Approximation[*]

Luca Bortolussi[3], Luca Cardelli[1,2], Marta Kwiatkowska[2], and Luca Laurenti[2]

[1] Microsoft Research
[2] Department of Computer Science, University of Oxford
[3] Department of Mathematics and Geosciences, University of Trieste

**Abstract.** We study time-bounded probabilistic reachability for Chemical Reaction Networks (CRNs) using the Linear Noise Approximation (LNA). The LNA approximates the discrete stochastic semantics of a CRN in terms of a continuous space Gaussian process. We consider reachability regions expressed as intersections of finitely many linear inequalities over the species of a CRN. This restriction allows us to derive an abstraction of the original Gaussian process as a time-inhomogeneous discrete-time Markov chain (DTMC), such that the dimensionality of its state space is independent of the number of species of the CRN, ameliorating the state space explosion problem. We formulate an algorithm for approximate computation of time-bounded reachability probabilities on the resulting DTMC and show how to extend it to more complex temporal properties. We implement the algorithm and demonstrate on two case studies that it permits fast and scalable computation of reachability properties with controlled accuracy.

## 1 Introduction

It is well known that a biochemical system evolving in a spatially homogeneous environment, at constant volume and temperature, can be modelled as a continuous-time Markov chain (CTMC) [18]. Stochastic modelling is necessary to describe stochastic fluctuations for low molecular counts [14,16], when deterministic models are not accurate [15]. Computing the probability distributions of the species over time is achieved by solving the Chemical Master Equation (CME) [25]. Unfortunately, numerical solution methods based on uniformisation [4] are often infeasible because of the state space explosion problem. A more scalable transient analysis can be achieved by employing statistical model checking based on the Stochastic Simulation Algorithm (SSA) [17], but to obtain good accuracy large numbers of simulations are needed, which for some systems can be very time consuming.

[*]This research is supported by a Royal Society Research Professorship and ERC AdG VERIWARE. LB is supported by EU-FET project QUANTICOL (nr 600708).

A promising approach is to instead approximate the CTMC induced by a biochemical system as a *continuous state space* stochastic process by means of the *Linear Noise Approximation (LNA)*, a Gaussian process derived as an approximation of the CME [25]. Its solution requires solving a number of differential equations that is quadratic in the number of species and independent of the molecular population. As a consequence, the LNA is generally much more scalable than a discrete state stochastic representation and has been successfully used for model checking of large biochemical systems [12,7]. However, none of these approaches enables the computation of global *probabilistic reachability* properties, that is, the probability of reaching a particular region of the state space in a particular time interval. This property is important not only to analyse biochemical systems, for example to quantify the probability that a particular protein or gene is ever expressed in Gene Regulatory Networks, but is also fundamental for the verification of more complex temporal logic properties, since model checking for CSL [2] or LTL [24] is reduced to the computation of reachability probabilities.

**Contributions.** We derive an algorithm to compute a fast and scalable approximation of probabilistic reachability using the LNA, where the target region of the state space is given by a polytope, i.e. an intersection of a set of linear inequalities over the species of a CRN. More specifically, we compute the probability that the system falls in the target region during a specified time interval. Given a set of $k$ linear inequalities, and relying on the fact that a linear combination of the components of a Gaussian distribution is still Gaussian, we discretize time and space for the $k$-dimensional stochastic process defined by the particular linear combinations. This permits the derivation of an abstraction in terms of a time-inhomogeneous *discrete-time Markov chain* (DTMC), whose dimension is independent of the number of species, since a linear combination is always uni-dimensional, and ensures scalability, as in general we are interested in one or at most two linear inequalities. This abstraction can then be used for model checking of complex temporal properties [21,2,4]. In order to compute such an abstraction, the most delicate aspect is to derive equations for the transition kernel of the resulting DTMC. This is given by the conditional probability at the next discrete time step given the system in a particular state. Reachability probabilities are then computed by making the target set absorbing. We use our algorithm to extend the Stochastic Evolution Logic (SEL) introduced in [12] to enable model checking of probabilistic reachability of linear combinations of the species of a CRN. We show the effectiveness of our approach on two case studies, also in cases where existing numerical model checking techniques are infeasible.

**Related work.** Algorithms to compute the reachability probabilities over discrete state space Markov processes are well understood [4]. They require computation of transient probabilities in a modified Markov chain, where states in the target region are made absorbing. Unfortunately, their practical use is severely hindered by state space explosion, which in a CRN grows exponentially with the number of molecules when finite, and may be infinite, in which case finite projection methods have to be used [23]. As a consequence, approximate but faster algorithms are appealing, in particular for CRNs, where it is not necessary to provide certified guarantees on reachability probabilities. The mainstream solution is to rely on simulations combined with statistical inference to obtain estimates [9]. These methods, however, are still computationally expensive. A recent trend of works explored as an alternative whether estimates could be obtained by relying on approximations of the stochastic process based on mean-field [6] or linear noise [8,7,12]. However, reachability properties, like those considered here, are very challenging. In fact, most approaches consider either local properties of individual molecules [6], or properties obtained by observing the behaviour of individual molecules and restricting the target region to an absorbing subspace of the (modified) model [7]. The only approach dealing with more general subsets, [8], imposes restrictions on the behaviour of the mean-field approximation, whose trajectory has to enter the reachability region in a finite time.

Our approach differs in that it is based on the LNA and considers regions defined by polytopes, which encompasses most properties of practical interest. The simplest idea would be to consider the LNA and compute reachability probabilities for this stochastic process, invoking convergence theorems for the LNA to prove the asymptotic correctness. Unfortunately, there is no straightforward way to do this, since dealing with a continuous space and continuous time diffusion process, e.g., Gaussian, is computationally hard, and computing reachability is challenging (see [10]). As a consequence, discrete abstractions are appealing.

## 2 Background

**Chemical Reaction Networks.** A *chemical reaction network (CRN)* $C = (\Lambda, R)$ is a pair of finite sets, where $\Lambda$ is a set of *chemical species*, $|\Lambda|$ denotes its size, and $R$ is a set of reactions. Species in $\Lambda$ interact according to the reactions in $R$. A *reaction* $\tau \in R$ is a triple $\tau = (r_\tau, p_\tau, k_\tau)$, where $r_\tau \in \mathbb{N}^{|\Lambda|}$ is the *reactant complex*, $p_\tau \in \mathbb{N}^{|\Lambda|}$ is the *product complex* and $k_\tau \in \mathbb{R}_{>0}$ is the coefficient associated with the rate of the reaction. $r_\tau$ and $p_\tau$ represent the stoichiometry of reactants and products. Given a reaction $\tau_1 = ([1,1,0]^T, [0,0,2]^T, k_1)$, where $\cdot^T$ is the transpose of a vector, we often refer to it as $\tau_1 : \lambda_1 + \lambda_2 \rightarrow^{k_1} 2\lambda_3$. The *state change* associated to a reaction $\tau$ is defined by $\upsilon_\tau = p_\tau - r_\tau$. For example, for $\tau_1$ as above, we have $\upsilon_{\tau_1} = [-1,-1,2]^T$. Assuming well mixed environment, constant volume $V$ and temperature, a *configuration* or *state* $x \in \mathbb{N}^{|\Lambda|}$ of the system is given by a vector of the number of molecules of each species. Given a configuration $x$ then $x(\lambda_i)$ represents the number of molecules of $\lambda_i$ in the configuration and $\frac{x(\lambda_i)}{N}$ is the concentration of $\lambda_i$ in the same configuration, where $N = V \cdot N_A$ is the volumetric factor or system size, $V$ is the volume and $N_A$ Avogadro's number. The *deterministic* semantics approximates the concentrations of species over time as the solution $\Phi(t)$ of the rate equations [11], a set of differential equations of the form:

$$\frac{d\Phi(t)}{dt} = F(\Phi(t)) = \sum_{\tau \in R} \upsilon_\tau \cdot (k_\tau \prod_{i=1}^{|\Lambda|} \Phi_i^{r_{i,\tau}}(t)) \qquad (1)$$

where $\Phi_i^{r_{i,\tau}}(t)$ is the $i$th component of vector $\Phi(t)$ raised to the power of $r_{i,\tau}$, $i$th component of vector $r_\tau$. The initial condition is $\Phi(0) = \frac{x_0}{N}$. It is known that Eqn (1) is accurate in the limit of high population [15].

**Stochastic Semantics.** The propensity rate $\alpha_\tau$ of a reaction $\tau$ is a function of the current configuration $x$ of the system such that $\alpha_\tau(x)dt$ is the probability that a reaction event occurs in the next infinitesimal interval $dt$. We assume mass action kinetics, therefore $\alpha_\tau(x) = k_\tau \frac{\prod_{i=1}^{|\Lambda|} r_{i,\tau}!}{N^{|r_\tau|-1}} \prod_{i=1}^{|\Lambda|} \binom{x(\lambda_i)}{r_{i,\tau}}$, where $r_{i,\tau}!$ is the factorial of $r_{i,\tau}$, and $|r_\tau| = \sum_{i=1}^{|\Lambda|} r_{i,\tau}$ [1]. To simplify the notation, $N$ is considered embedded inside the coefficient $k$ for any reaction. The stochastic semantics of the CRN $C = (\Lambda, R)$ is represented by a *time-homogeneous continuous-time Markov chain* (CTMC) [15] $(X^N(t), t \in \mathbb{R}_{\geq 0})$ with state space $S$, where in $X^N$ we made explicit the dependence by $N$. $X^N(t)$ is a random vector describing the molecular population of each species at time $t$. Let $x_0 \in \mathbb{N}^{|\Lambda|}$ be the initial condition of $X^N$ then $P(X^N(0) = x_0) = 1$. For $x \in S$, we define $P(x,t) = P(X^N(t) = x \,|\, X^N(0) = x_0)$. The transient evolution of $X^N$ is described by the Chemical Master Equation (CME), a set of differential equations

$$\frac{\mathrm{d}}{\mathrm{d}t} \left(P(x,t)\ \right) = \sum_{\tau \in R} \{\alpha_\tau(x - \upsilon_\tau) P(x - \upsilon_\tau, t) - \alpha_\tau(x) P(x,t)\}. \qquad (2)$$

Solving Eqn (2) requires computing the solution of a differential equation for each reachable state. The size of the reachable state space depends on the number of species and molecular populations and can be huge or even infinite. As a consequence, solving the CME is generally feasible only for CRNs with very few species and small molecular populations.

**Linear Noise Approximation.** The *Linear Noise Approximation* (LNA) is a continuous state space approximation of the CME, which approximates the CTMC induced by a CRN as a Gaussian process [25]. In [26], the LNA has been derived as a linearized solution of the Chemical Langevin Equation (CLE) [19]. This derivation shows that the LNA is accurate if the two *leap conditions* on the reactions are satisfied. The leap conditions are satisfied at time $t$ if (i) there exists an infinitesimal time interval $dt$ such that the propensity rate of each reaction is approximately constant during $dt$ and if (ii) each reaction fires many times during $dt$. It is possible to show that, assuming mass action kinetics, in the limit of high volume these conditions are always satisfied. The LNA at time $t$ approximates the distribution of $X^N(t)$ with the distribution of the random vector $Y^N(t)$ such that

$$X^N(t) \approx Y^N(t) = N\Phi(t) + N^{\frac{1}{2}} G(t) \qquad (3)$$

where $G(t) = (G_1(t), G_2(t), ..., G_{|\Lambda|})$ is a random vector, independent of $N$, representing the stochastic fluctuations at time $t$ and $\Phi(t)$ is the solution of Eqn (1). The probability distribution of $G(t)$ is then given by the solution of a linear Fokker-Planck equation [26]. As a consequence, for every time instant $t$, $G(t)$ has a multivariate normal distribution whose expected value $E[G(t)]$ and covariance matrix $C[G(t)]$ are the solution of the following differential equations:

$$\frac{\mathrm{d}E[G(t)]}{\mathrm{d}t} = J_F(\Phi(t))E[G(t)] \tag{4}$$

$$\frac{\mathrm{d}C[G(t)]}{\mathrm{d}t} = J_F(\Phi(t))C[G(t)] + C[G(t)]J_F^T(\Phi(t)) + W(\Phi(t)) \tag{5}$$

where $J_F(\Phi(t))$ is the Jacobian of $F(\Phi(t))$, $J_F^T(\Phi(t))$ its transpose, $W(\Phi(t)) = \sum_{\tau \in R} \upsilon_\tau {\upsilon_\tau}^T \alpha_{c,\tau}(\Phi(t))$ and $F_j(\Phi(t))$ the $j$th component of $F(\Phi(t))$. We assume $X^N(0) = x_0$ with probability 1; as a consequence $E[G(0)] = 0$ and $C[G(0)] = 0$, which implies $E[G(t)] = 0$ for every $t$. The following theorem illustrates the nature of the approximation using the LNA.

**Theorem 1.** *[15] Let $C = (\Lambda, R)$ be a CRN and $X^N$ the discrete state space Markov process induced by $C$. Let $\Phi(t)$ be the solution of rate equations with initial condition $\Phi(0) = \frac{x_0}{N}$ and $G$ be the Gaussian process with expected value and variance given by Eqns* (4) *and* (5)*. Then, for any $t < \infty$ and $N \to \infty$,*

$$N^{\frac{1}{2}} \left| \frac{X^N(t)}{N} - \Phi(t) \right| \Rightarrow_N G(t). \tag{6}$$

In the above, ⇒ indicates convergence in distribution [5]. The LNA is exact in the limit of high populations, but can also be used in different scenarios if the leap conditions are satisfied [20,26]. To compute the LNA it is necessary to solve $O(|\Lambda|^2)$ first order differential equations, and the complexity is independent of the initial number of molecules of each species. Therefore, one can avoid the exploration of the state space that methods based on uniformization rely upon.

## 3 Linear Noise Approximation of Reachability Probabilities

We are interested in computing the probability that the CTMC induced by a biochemical network enters a region of the state space at some time instant between $t_1$ and $t_2$. In order to exploit the LNA, we will first discretize time for the Gaussian process given by the LNA, with a fixed (or adaptive) step size $h$, which we can do effectively owing to the Markov property and the knowledge of its mean and covariance. As a result, we obtain a *discrete-time, continuous space*, Markov process with a Gaussian transition kernel. Then, by resorting to state space discretization, we compute the reachability probability on this new process, obtaining an approximation converging to the LNA approximation as $h$ tends to zero.

At first sight, there seems to be little gain, as we now have to deal with a $|\Lambda|$-dimensional continuous state space. Indeed, for general regions this can be the case. However, if we restrict to regions defined by linear inequalities, we can exploit properties of Gaussian distributions (i.e. their closure wrt linear combinations), reducing the dimension of the continuous space to the number of different linear combinations used in the definition of the linear inequalities (in fact, the same hyperplane can be used to fix both an upper and a lower bound). As typically we are interested in regions defined by one or two inequalities, the complexity will then be dramatically reduced.

### 3.1 Reachability Problem: Formal Definition

Recall that, given a CRN $C = (\Lambda, R)$ with initial configuration $x_0$, its stochastic behaviour is described by the CTMC $X^N$. A path of $X^N$ is a sequence $\omega = x_0 t_1 x_1 t_1 x_2 ...$ where $x_i \in \mathbb{N}^{|\Lambda|}$ is a state and $t_i \in \mathbb{R}_{>0}$ is the time spent in the state $x_i$. A path is finite if there is a state $x_k$ that is absorbing. $\omega(t)$ is the state of the path at time $t$. $Path(X^N, x_0)$ is the set of all (finite and infinite) paths of the CTMC starting in $x_0$. We work with the standard probability measure $Prob$ over paths $Path(X^N, x_0)$ defined using cylinder sets [21].

We now formalize the reachability problem we want to solve. For a simpler presentation, we restrict to a single linear inequality over the species. This still covers many practical scenarios, in particular in systems biology. Next, we show how to generalise the method to regions specified by the intersection of more than one hyperplane, though the complexity of our method will grow exponentially with the number of different hyperplanes, unless additional approximations are introduced.

**Definition 1.** *Let $C = (\Lambda, R)$ be a CRN with initial state $x_0$, fix vector of weights $B \in \mathbb{Z}^{|\Lambda|}$, finite set of disjoint intervals $I = [l_1, u_1] \cup ... \cup [l_k, u_k], k \geq 1$, such that, for $i \in [1, k]$, $[l_i, u_i] \subseteq \mathbb{R} \cup [-\infty, +\infty]$, and an interval $[t_1, t_2] \subset \mathbb{R}_{\geq 0}$. The reachability probability of $B$-weighted linear combination of species falling in the target set $I$ in time interval $[t_1, t_2]$, for initial condition $x_0$, is*

$$P_{reach}(B, x_0, I, [t_1, t_2]) = Prob\{\omega \in Path(X^N, x_0) | B \cdot \omega(t) \in I, t \in [t_1, t_2]\}. \quad (7)$$

### 3.2 LNA and Dimensionality Reduction

In order to approximate the reachability probability in Eqn (7), we rely on the LNA $Y^N(t)$ of $X^N(t)$ (Eqns (4) (5)). By Eqn (3), we have that the distribution of $Y^N(t)$ is Gaussian with expected value and covariance matrix given by

$$E[Y^N(t)] = N\Phi(t)$$
$$C[Y^N(t)] = N^{\frac{1}{2}} C[G(t)] N^{\frac{1}{2}} = NC[G(t)].$$

Let $B \in \mathbb{Z}^{|\Lambda|}$, then $Z^N = B \cdot Y^N$ is a uni-dimensional process and for any $t$ it represents the time evolution of the linear combination of the species defined by $B$ over time. Furthermore, $Z^N(t)$ is also Gaussian distributed, being the linear combination of Gaussian variables. In particular, $Z^N(t)$ is characterised by the following mean and covariance:

$$E[Z^N(t)] = BE[Y^N(t)] \quad (8)$$
$$C[Z^N(t)] = BC[Y^N(t)]B^T \quad (9)$$

Note also that the distribution of $Z^N$ depends on $Y^N$ *only via its mean and covariance*, which are obtained by solving ODEs (4) and (5). This is the key feature that enables an effective dimensionality reduction.

### 3.3 Time Discretization Scheme

We now introduce an exact time discretization scheme for $Z^N$. Fix a small time step $h > 0$. By sampling $Y^N$ at step $h$ and invoking the Markov property,[4] we obtain a *discrete-time Markov process* (DTMP) $\bar{Y}^N(k) = Y^N(kh)$ on continuous space. Applying the linear projection mapping $Z^N$ to $\bar{Y}^N(k)$, and leveraging its Gaussian nature, we obtain a process $\bar{Z}^N(k) = Z^N(kh)$ which is also a DTMP, though with a kernel depending on time through the mean and variance of $Y^N$.

**Definition 2.** *A (time-inhomogeneous) discrete-time Markov process (DTMP) $(\bar{Z}^N(k), k \in \mathbb{N})$ is uniquely defined by a triple $(S, \sigma, T)$, where $(S, \sigma)$ is a measurable space and $T : \sigma \times S \times \mathbb{N} \to [0, 1]$ is a transition kernel such that, for any $z \in S$, $A \in \sigma$ and $k \in \mathbb{N}$, $T(A, z, k)$ is the probability that $\bar{Z}^N(k + 1) \in A$ conditioned on $\bar{Z}^N(k) = z$. $S$ is the state space of $\bar{Z}^N$.*

In order to characterise $\bar{Z}^N$, we need to compute its transition kernel. This can be done by computing $f_{Z^N(t+h)|Z^N(t)=\bar{z}}(z)$, i.e. the density function of $Z^N(t+h)$ given the event $Z^N(t) = \bar{z}$.

Consider the joint distribution $Y^N(t), Y^N(t+h)$, which is Gaussian. Its projected counterpart $Z^N(t), Z^N(t+h)$ is thus also Gaussian, with covariance function $cov(Z^N(t), Z^N(t+h)) = Bcov(Y^N(t), Y^N(t+h))B^T$, where $cov(Y^N(t), Y^N(t+h))$ is the covariance function of $Y^N$ at times $t$ and $t + h$. It follows by the linearity of $B$ that $f_{Z^N(t+h)|Z^N(t)=\bar{z}}$ is Gaussian too, and to fully characterize it we need to compute $E[Z^N(t + h)|Z^N(t) = \bar{z}]$ and $C[Z^N(t + h)|Z^N(t) = \bar{z}]$. To this end, we need to derive $cov(Y^N(t), Y^N(t + h))$. From now on, we denote $cov(Y^N(t+h), Y^N(t)) = C_Y(t+h, t)$ and $cov(Z^N(t+h), Z^N(t)) = C_Z(t+h, t)$. Following [15], we introduce the following matrix differential equation

$$\frac{d\Omega(t, s)}{dt} = J_F(\Phi(t))\Omega(t, s) \tag{10}$$

with $t \geq s$ and initial condition $\Omega(s, s) = Id$, where $Id$ is the identity matrix of dimension $|\Lambda|$. Then, as illustrated in [15], we have

$$C_Y(t, t + h) = \int_0^t \Omega(t, s)J_F(\Phi(s))[\Omega(t + h, s)]^T ds. \tag{11}$$

This is an integral equation, which has to be computed numerically. To simplify this task, we derive an equivalent representation in terms of differential equations. This is given by the following lemma.

**Lemma 1.** *Solution of Eqn* (11) *is given by the solution of the following differential equations*

$$\frac{dC_Y(t, t + h)}{dt} = W(\Phi(t))\Omega^T(t+h, t)+J_F(\Phi(t))C_Y(t, t+h)+C_Y(t, t+h)J_F^T(\Phi(t+h)) \tag{12}$$

[4] The Gaussian process obtained by linear noise approximation is Markov, as it is the solution of a linear Fokker-Planck equation (stochastic differential equation) [25].

*with initial condition $C_Y(0,h)$ computed as the solution of*

$$\frac{dC_Y(0,s)}{ds} = C_Y(0,0+s)J_F^T(\Phi(s)).$$

$\Omega(t+h,t)$ can be computed by solving Eqn (10). Knowledge of $C_Y(t,t+h)$ allows us to directly compute $C_Z(t,t+h) = BC_Y(t,t+h)B^T$. Then, by using the law for conditional expectation of a Gaussian distribution, we finally have

$$\begin{aligned} &E[Z^N(t+h)|Z^N(t)=\bar{z}] = \\ &\qquad E[Z^N(t+h)] + C_Z(Z(t+h),Z(t))C[Z(t)]^{-1}(\bar{z}-E[Z^N(t)]) \end{aligned} \tag{13}$$

$$C[Z^N(t+h)|Z^N(t)=\bar{z}] = C[Z^N(t+h)] - C_Z(t,t+h)C[Z^N(t)]^{-1}C_Z(t,t+h). \tag{14}$$

Note that the resulting kernel is time-inhomogeneous. The dependence on $t$ is via the mean and covariance of $Y^N$, which are functions of time and define completely the distribution of $Y^N$.

### 3.4 Computation of Reachability Probabilities

In order to compute the reachability probability for the DTMP $\bar{Z}^N(k)$, we discretize its continuous state space, obtaining an abstraction in terms of a discrete-time Markov chain (DTMC) $Z^{N,D}(k)$ with state space $S$. That is, the states of the original Markov process are partitioned into a countable set of non-overlapping sets. We assume an order relation between elements of each set and, for each set, we consider a representative point, given by the median of the set. $S$ is given by the set of representative points. In particular, we partition the state space of $\bar{Z}^N$ in intervals of length $2\Delta z$, where $\Delta z$ defines how fine our space discretization is. A possible choice is $\Delta z = 0.5$, which basically means $S \subseteq \mathbb{Z}$. For $\Delta z \to 0$ the error introduced by the space discretization goes to zero. However, when the molecular population is of the order of hundreds or thousands, it can be beneficial to consider $\Delta z > 0.5$, since a coarser state space aggregation is reasonable.

Then, we solve the reachability problem on the resulting DTMC. For $z', z \in S$, the transition kernel of $Z^{N,D}(k)$ is defined as

$$T(z',z,k) = \int_{z'-\Delta z}^{z'+\Delta z} f_{Z^N(hk+h)|Z^N(hk)=z}(x)dx, \tag{15}$$

where $h$ is the discrete time step, assumed to be fixed for a simpler notation. Finally, in order to compute the reachability of the target set $I$ we make all the states $z \in I$ absorbing. That is, for $z \in I$

$$T(z',z,k) = \begin{cases} 1 & \text{if } z' = z \\ 0 & \text{otherwise} \end{cases}$$

Algorithm 1 illustrates our approach for computing reachability probabilities.

**Algorithm 1** Compute Time-Bounded Probabilistic Reachability

```
Require: A CRN $C = (\Lambda, R)$ with initial condition $x_0$, $B \in \mathbb{Z}^{|\Lambda|}$, a finite time interval
    $[t_1, t_2]$, a target set $I$ and a threshold $\mathcal{TH}$.
 1: function ComputeReach($C$, $B$, $x_0, I, [t_1, t_2], \mathcal{TH}$)
 2:     Set $t = 0$, $S = \{B \cdot x_0\}$ and $P(Z^{N,D}(0) = B \cdot x_0) = 1$
 3:     while $t < t_1$ do
 4:         Compute time step $h$
 5:         for each $z \in S$ do
 6:             Propagate probability at time $t + h$ and update $S$
 7:         for each $z \in S$ do
 8:             if $P(Z^{N,D}(t + h) = z) < \mathcal{TH}$ then
 9:                 $S \leftarrow S - \{z\}$
10:         $t \leftarrow t + h$
11:     while $t < t_2$ do
12:         Compute time step $h$
13:         for each $z \in S/I$ do
14:             Propagate probability at time $t + h$ and update $S$
15:         for each $z \in S/I$ do
16:             if $P(Z^{N,D}(t + h) = z) < \mathcal{TH}$ then
17:                 $S \leftarrow S - \{z\}$
18:         $t \leftarrow t + h$
19:     return $P_{reach}(B, x_0, I, [t_1, t_2]) = \sum_{z \in I} P(Z^{N,D}(t) = z)$
```

In Line 1, we initialize the system at time 0. In the context of the algorithm, $S$ is a set containing the reachable states at a particular time with probability mass greater than the threshold $\mathcal{TH}$. $\mathcal{TH}$ equals $10^{-14}$ in all our experiments. This guarantees that the algorithm always terminates in finite time even if the state space is not finite. Initially, we have that $S$ contains only one state $B \cdot x_0$. Then, in Lines $3 - 10$, we propagate the probability for any discrete step until $t < t_1$, as illustrated in [21]. For generality, we assume that the time step $h$ is chosen adaptively, according to the system dynamics. Propagating probability is possible, as for any $z' \in S$, $T(z', z, k)$, which has a Gaussian nature, defines the probability of being in $z'$ in the next discrete time step by $Z^{N,D}(k) = z$. From Line 12 to 20, we compute probabilistic reachability $P_{reach}(B, x_0, I, [t_1, t_2])$ by propagating the probability only for states that are not in $I$. When we reach $t \geq t_2$, we have that $P_{reach}(B, x_0, I, [t_1, t_2]) \approx \sum_{z \in I} P(Z^{N,D}(t) = z | Z^{N,D}(0) = B \cdot x_0)$.

### 3.5 Correctness

The method we present is approximate. In particular, errors are introduced in two ways: by resorting to the LNA and by discretisation of time and space of the LNA. The quality of these two approximations is controlled by three parameters: (a) $N$, the system size, which influences the accuracy of LNA, (b) $h$, the step size, and (c) $\Delta z$, the discretization step, which influences the quality of the approximation of the reachability probability of the LNA.

Recall that $X^N$ and $Z^{N,D}$ are, respectively, the CTMC induced by a CRN and the DTMC obtained by discretization of the LNA of $X^N$ for a particular $N$. Fix $B \in \mathbb{Z}^{|\Lambda|}$ and $I$, a set of disjoint closed intervals of reals, and denote by $P_{X^N}(B, t_1, t_2)$ and $P_{Z^{N,D}}(B, t_1, t_2)$, $t_1 < t_2$, the reachability probabilities for $Z^{N,D}$ and $X^N$. Then, we have the following result

**Theorem 2.** *With the notation above, for $t_1 \leq t_2 < \infty$:*

$$\lim_{N \to \infty} \lim_{h \to 0} \lim_{\Delta z \to 0} \{|P_{X^N}(B, t_1, t_2) - P_{Z^{N,D}}(B, t_1, t_2)|\} = 0.$$

The convergence stated in Theorem 2 means that, since $N$ is fixed for a given CRN, that even if we have control over $h$ and $\Delta z$, the quality of the approximation depends on how well the LNA approximated the CRN. Error bounds would be a viable companion to estimate the committed error, but we are not aware of any explicit formulation of them for the convergence of the LNA. However, experimental results in Section 5 show that the error committed is generally limited also for moderately small $N$ and quite large $h$.

### 3.6 Complexity

Complexity of the method depends on the following: (a) the equations we need to solve, (b) the step size $h$, and (c) the space discretization step $\Delta z$. Algorithm 1 requires solving Eqns (12) and Eqns (5), that is, a set of differential equations quadratic in the number of species. In fact, solving these equations requires computing $J_F$, Jacobian of $F$. However, the number of equations we need to solve is independent of the number of molecules in the system. This guarantees the scalability of our approach. An important point is that Eqn (12) requires solving Eqn (11) once for each sampling point of the numerical solution of Eqn (12). A possible way to avoid this is to consider approximate solutions of Eqn (11), which are accurate in the limit of $h \to 0$. However, to keep this approximation under control, $h$ has to be chosen really small, slowing down the computation. Moreover, for any sample point, Eqn (11) is solved only for a small time interval (between $t$ and $t + h$). As a consequence, in practice, the computational cost introduced in solving Eqn (11) is under control.

A smaller value of $h$ implies that, for a given time interval, we have a greater number of discrete time steps, which can slow down the computation in some cases. The value of $\Delta z$ determines the number of states of the resulting DTMC. However, we stress that we discretize $Z^N(t)$, a uni-dimensional distribution (or $m$-dimensional in the case we have $m > 1$ linear inequalities). As a consequence, the number of reachable states with probability mass is generally limited and under control. Obviously, if the number of molecules is large and $\Delta z$ extremely small, then this is detrimental on performance.

### 3.7 Extensions

*Remark 1.* Our approach can be easily extended to target regions defined by intersections of finitely many linear inequalities over species. That is, we consider a set of linear predicates $Z_j^N = B_j \cdot X^N(t) \in I_j$, $j = 1 \ldots, m$ with

$m > 1$, and ask what is the probability that during a finite time interval we are in a state where each predicate is verified. In order to do that, we can define $B = (B_1, ..., B_m)^T \in \mathbb{Z}^{m \times |\Lambda|}$, a matrix where each row is a vector specifying a different linear combination. As a consequence, $Z^N = B \cdot Y^N$ is an $m$ dimensional Gaussian process and all the properties we used for the unidimensional case remain valid in this extended scenario. The resulting DTMC $Z^{N,D}$ is $m-$dimensional. However, note that $m$ is generally equal to 1 or 2 in practical applications (see Remark 2).

*Remark 2.* The method presented here can be extended to compute the probability of a non-nested until formula of $CSL$ [3], that is, a formula of the type

$$P_{\sim p}[\Psi_1 U^{[t_1,t_2]} \Psi_2].$$

This formula is satisfied if the probability of a path such that there exists $t \in [t_1, t_2]$ for which $\Psi_2$ is satisfied and, for all $t' \in [0, t]$, $\Psi_1$ is satisfied meets the bound $p$. We restrict $\Psi_1, \Psi_2$ to linear inequalities over species. Computing this probability, as explained in [21], requires computing two terms: (a) the probability of reaching a state between $[0, t_1)$ such that $\neg\Psi_1$ is satisfied, and (b) the probability of reaching a state during $[t_1, t_2]$ where $\neg\Psi_1 \wedge \Psi_2$ is satisfied. The former is simply reachability on $\neg\Psi_1$. The latter can be computed by considering reachability over the bi-dimensional system given by the joint distribution of the linear combinations associated to $\neg\Psi_1$ and $\Psi_2$.

## 4 Stochastic Evolution Logic (SEL)

The method presented here permits an extension of the Stochastic Evolution Logic (SEL) introduced in [12] for approximate model checking of CRNs based on the LNA. Here, we extend the original formulation of the logic with an operator for computing (time-bounded) probabilistic reachability. However, as explained in Remark 2, more complex temporal behaviours could be introduced as well.

Let $C = (\Lambda, R)$ be a CRN with initial state $x_0$, then SEL enables evaluation of the probability, reachability, variance and expectation of linear combinations of populations of the species of $C$. The syntax of SEL is given by

$$\eta := P_{\sim p}[B, I]_{[t_1,t_2]} \quad | \quad F_{\sim p}[B, I]_{[t_1,t_2]} \quad | \quad Q_{\sim v}[B]_{[t_1,t_2]} \quad | \quad \eta_1 \wedge \eta_2 \quad | \quad \eta_1 \vee \eta_2$$

where $Q = \{supV, infV, supE, infE\}$, $\sim = \{<, >\}$, $p \in [0, 1]$, $v \in \mathbb{R}$, $B \in \mathbb{Z}^{|\Lambda|}$, $I = [l_1, u_1] \cup ... \cup [l_k, u_k], k \geq 1$, such that, for $i \in [1, k]$, $[l_i, u_i] \subseteq \mathbb{R} \cup [-\infty, +\infty]$ is a finite set of disjoint intervals and $[t_1, t_2]$ is a closed time interval, with the constraint that $t_1 \leq t_2$ and $t_1, t_2 \in \mathbb{R}_{\geq 0}$. If $t_1 = t_2$ the interval reduces to a singleton.

Formulae $\eta$ describe global properties of the stochastic evolution of the system. $(B, I)$ specifies a linear combination of the species, where $B \in \mathbb{Z}^{|\Lambda|}$ is a vector of weights defining the linear combination and $I$ is a set of disjoint closed real intervals. $P_{\sim p}[B, I]_{[t_1,t_2]}$ is the probabilistic operator, which specifies the average value of the probability that the linear combination defined by $B$ falls within the range $I$ over the time interval $[t_1, t_2]$. Given $Pr_{B,I}^{X^N}(t) = Prob\{\omega \in Path(X^N, x_0) \mid B \cdot \omega(t) \in I\}$, then, for $t_1 = t_2$, its semantics is defined as

$$X^N, x_0 \models P_{\sim p}[B, I]_{[t_1,t_1]} \quad \leftrightarrow \quad Pr_{B,I}^{X^N}(t_1) \sim p.$$

Instead, for $t_1 < t_2$ we have

$$X^N, x_0 \models P_{\sim p}[B, I]_{[t_1,t_2]} \quad \leftrightarrow \quad \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} Pr_{B,I}^{X^N}(t)\,\mathrm{d}t \sim p.$$

$F_{\sim p}[B, I]_{[t_1,t_2]}$ is the new probabilistic reachability operator, which specifies the probability that the linear combination of species defined by $B$ reaches $I$ during $[t_1, t_2]$. Its semantics can be defined as

$$X^N, x_0 \models F_{\sim p}[B, I]_{[t_1,t_2]} \leftrightarrow Prob(\omega \in Path(X^N, x_0) | B \cdot \omega(t) \in I, t \in [t_1, t_2]) \sim p$$

The operators $supE, infE, infV, supV$, see [12], respectively, yield the supremum and infimum of expected value and variance of the random variables associated to $B$ within the specified time interval. The quantitative value associated to a formula can be computed by writing =? instead of $\sim p$ or $\sim v$. For instance, $F_{=?}[B, I]_{[t_1,t_2]}$ gives the probability value associated to the reachability property. The following example illustrates that the $P$ and $F$ operators differ.

*Example 1.* Consider the following CRN, taken from [13], modelling a phosphorelay network

$$\tau_1 : L1 + ATP \rightarrow^{0.01} L1p + ATP; \quad \tau_2 : L1p + L2 \rightarrow^{0.01} L2p + L2;$$
$$\tau_3 : L2p + L3 \rightarrow^{0.01} L3p + L2; \quad \tau_4 : L3p \rightarrow^{0.1} L3;$$

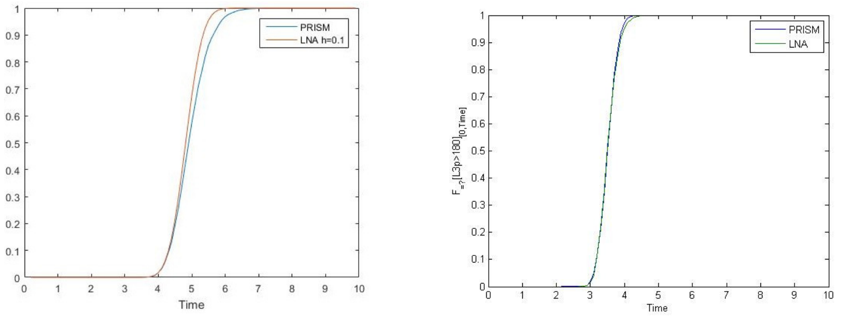
with initial conditions $x_0(L1) = x_0(L2) = x_0(L3) = 50$, $x_0(ATP) = 150$ and all other species equal 0. Then, if we consider $P_{>0.3}[L3p, [40, \infty]]_{[0,10]}$, which is true if the average probability that $L3p > 40$ is greater that 0.3. Then, this is evaluated to false. Instead, $F_{>0.3}[L3p, [40, \infty]]_{[0,10]}$, which models the probability of being in a state where $L3p > 40$ during the first 10 seconds, is evaluated as true.

## 5 Experimental Results

We implemented Algorithm 1 in Matlab and evaluated it on two case studies. All the experiments were run on an Intel Dual Core i7 machine with 8 GB of RAM. The first case study is a Phospohorelay Network with 7 species. We use this example to show the trade-off between the different parameters and the molecular population. More precisely, we show that the accuracy of our approach increases as the number of molecules grows, but can still give fast and accurate results when the molecular population is not large. The second example is a Gene Regulatory network. We use this example to show how our approach is more powerful than existing approximate techniques, and is able to accurately handle properties where existing techniques fail. We validate our results by comparing our method with statistical model checking (SMC) as implemented in PRISM [22]. In fact, for both examples, exact numerical computation of the reachability probabilities on the CTMC is infeasible because of state space explosion.

### 5.1 Phosphorelay Network

The first case study is a three-layer phosphorelay network as shown in Example 1. There are 3 layers, $(L1, L2, L3)$, which can be found in phosphorylate form $(L1p, L2p, L3p)$, and the ligand $B$. We consider the initial condition $x_0 \in \mathbb{N}^7$ such that $x_0(L1) = x_0(L2) = x_0(L3) = L \in \mathbb{N}$, $x_0(L1p) = x_0(L2p) = x_0(L3p) = 0$ and $x_0(B) = 150$. In Figure 1, we compare the estimates obtained by our approach for two different initial conditions ($L = 100$ and $L = 200$) with statistical model checking as implemented in PRISM [22], with 30000 simulations and confidence interval equal to 0.01. In both experiments we consider $\Delta z = 0.5$.



(a)

(b)

Fig. 1: Comparison of the evaluation of $F_{[0,Time]}[L3p > 80]$ (Fig 1a) and $F_{[0,Time]}[L3p > 180]$ (Fig 1b) using statistical model checking as implemented in PRISM and our approach. In Fig 1a, we used $h = 0.1$, $\Delta z = 0.5$, and $L = 100$. In Fig 1b, we considered $h = 0.1$, $\Delta z = 0.5$ and $L = 200$.

In Figure 1a we can see that, if we increase the time interval of interest, the error tends to increase. This is because, for $L = 100$, the LNA and CME do not have perfect convergence. As a consequence, at every step of the discretized DTMC, a small error is introduced. This source of error is present until we enter the target region with probability 1. If we increase $L$ this error disappears, and the inaccuracies are due to the finiteness of $h$ and $\Delta z$. However, already for $h = 0.1$ and $L = 100$, the LNA produces a fast and reasonably accurate approximation. In the following table we compare our approach and PRISM evaluations for different values of $L$ and $h$ and $\Delta z = 0.5$. In order to compare the accuracy we consider the absolute average error, $||\epsilon||_1$, and the maximum absolute error, $||\epsilon||_\infty$. $||\epsilon||_1 = \frac{1}{|\Sigma|}\sum_{n\in\Sigma}|F^Y_{[0,n]} - F^X_{[0,n]}]|$ and $||\epsilon||_\infty = max_{n\in\Sigma}\{|F^Y_{[0,n]} - F^X_{[0,n]}]|\}$, where $\Sigma$ is the set of discrete times between 0 and 10, and $F^Y_{[0,n]}$ and $F^X_{[0,n]}$ are the evaluation of the particular reachability formula in the interval $[0, n]$ according to the LNA and PRISM.

| Property | Ex. Time | h | L | $\|\|\epsilon\|\|_1$ | $\|\|\epsilon\|\|_\infty$ |
|---|---|---|---|---|---|
| $F_{=?}[L3p > 80]_{[0,Time]}$, $Time \in [0,10]$ | 97 sec | 0.1 | 100 | 0.0088 | 0.11 |
| $F_{=?}[L3p > 180]_{[0,Time]}$, $Time \in [0,10]$ | 130 sec | 0.1 | 200 | 0.0015 | 0.0217 |
| $F_{=?}[L3p > 80]_{[0,Time]}$, $Time \in [0,10]$ | 28 sec | 0.5 | 100 | 0.0381 | 0.24 |
| $F_{=?}[L3p > 180]_{[0,Time]}$, $Time \in [0,10]$ | 39 sec | 0.5 | 200 | 0.0289 | 0.14 |

The results show that the best accuracy is obtained for $h = 0.1$ and $L = 200$, where $h = 0.1$ induces a finer time discretization, whereas the worst are for $h = 0.5$ and $L = 100$. We comment that the numerical solution of the CME using PRISM is not feasible for this model, and our method is several orders of magnitude faster than statistical model checking with PRISM (30000 simulations for each time point).

### 5.2 Gene Expression

We consider the following gene expression model, as introduced in [27]:

$$\tau_1 :\rightarrow^{0.5} mRNA;\;\; \tau_2 : mRNA \rightarrow^{0.0058} mRNA + P;$$

$$\tau_3 : mRNA \rightarrow^{0.0029} W;\;\; \tau_4 : P \rightarrow^{0.0001} W;$$

with initial condition $x_0$ such that all the species have initial concentrations equal to 0. We consider the property $F_{=?}[\geq 175]_{[0,Time]}$, which quantifies the probability that the $mRNA$ is produced for at least 175 molecules during the first $Time$ seconds, for $Time \in [0, 1000]$. This is a particularly difficult property because the trajectory of the mean-field of the model, and so the expected value of the LNA, does not enter the target region. As a consequence, approximate approaches introduced in [15] and [8], which are based on the hitting times of the mean-field model, fail and evaluate the probability as always equal to 0.

Conversely, our approach is able to evaluate correctly such a property. Figure 2 compares the value computed by our approach with statistical model checking of the same property as implemented in PRISM over 30000 simulations for each time point and confidence interval 0.01. In Figure 2 we consider $h = 1.8$ and $\Delta z = 0.5$ and demonstrate that our approach is able to correctly estimate such a difficult property. Note that, as the mean-field does not enter the target region, for each time point the probability to enter the target region depends on a portion of the tail of the Gaussian given by the LNA. As a consequence, the accuracy of our results strictly depends on how

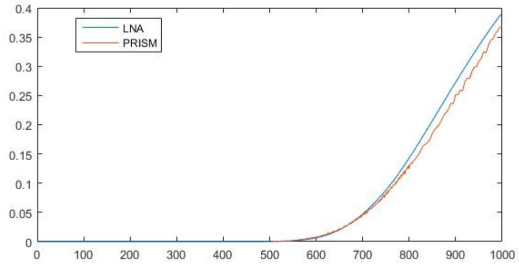


Fig. 2: The figure plots $F_{=?}[mRNA \geq 174]_{[0,Time]}$ for $h = 1.85$ and $\Delta z = 0.5$. The x-axis represents the value of $Time$ and the y-axis the quantitative value of the formula for that value of $Time$.

well the LNA approximates the original CTMC, much more than for properties where the mean-field enters the target region. In the following table, we evaluate our results for two different values of $h$ and $\Delta z = 0.5$.

| Property | Ex. Time | h | $\|\|\epsilon\|\|_1$ | $\|\|\epsilon\|\|_\infty$ |
|---|---|---|---|---|
| $F_{=?}[mRNA \geq 174]_{[0,Time]}$, $Time \in [0, 100]$ | 298 sec | 1.85 | 0.0075 | 0.022 |
| $F_{=?}[mRNA \geq 174]_{[0,Time]}$, $Time \in [0, 100]$ | 152 sec | 5 | 0.0147 | 0.13 |

## 6 Conclusion

We presented a method for computing (time-bounded) probabilistic reachability for CRNs based on the LNA, which is challenging because the LNA yields a continuous time and uncountable state space stochastic process. As a consequence, existing methods that rely on finite state spaces cannot be used directly and discretizing the uncountable state space defined by the LNA will lead to state space explosion. In order to overcome these issues, we considered reachability regions defined as polytopes. Using the fact that the LNA is a solution of a linear Fokker-Planck equation, and so a Gaussian Markov process, for a given linear combination of the species of a CRN, we are able to project the original, multi-dimensional Gaussian process onto a uni-dimensional stochastic process. We then derived an abstraction in terms of a time-inhomogeneous DTMC, whose state space is independent of the number of the species of a CRN, as it is derived by discretizing linear combinations of the species. This ensures scalability. Finally, we used our approach to extend the Stochastic Evolution Logic in order to verify complex temporal properties. On two case studies, we showed that our approach permits fast and scalable probabilistic analysis of CRNs. The accuracy depends on parameters controlling space and time discretization, as well as the accuracy of the LNA.

## References

1. D. F. Anderson and T. G. Kurtz. Continuous time Markov chain models for chemical reaction networks. In *Design and analysis of biomolecular circuits*, pages 3–42. Springer, 2011.
2. A. Aziz, K. Sanwal, V. Singhal, and R. Brayton. Verifying continuous time Markov chains. In *Computer Aided Verification*, pages 269–276. Springer, 1996.
3. A. Aziz, K. Sanwal, V. Singhal, and R. Brayton. Model-checking continuous-time Markov chains. *ACM Transactions on Computational Logic (TOCL)*, 1(1):162–170, 2000.
4. C. Baier, B. Haverkort, H. Hermanns, and J.-P. Katoen. Model-checking algorithms for continuous-time markov chains. *Software Engineering, IEEE Transactions on*, 29(6):524–541, 2003.
5. P. Billingsley. *Convergence of probability measures.* Wiley, 1999.
6. L. Bortolussi and J. Hillston. Fluid model checking. In *CONCUR 2012–Concurrency Theory*, pages 333–347. Springer, 2012.
7. L. Bortolussi and R. Lanciani. Model checking Markov population models by central limit approximation. In *Quantitative Evaluation of Systems*, pages 123–138. Springer, 2013.

8. L. Bortolussi and R. Lanciani. Stochastic approximation of global reachability probabilities of Markov population models. In *Computer Performance Engineering*, pages 224–239. Springer, 2014.
9. L. Bortolussi, D. Milios, and G. Sanguinetti. Smoothed model checking for uncertain continuous-time Markov chains. *Information and Computation*, 2016.
10. L. M. Bujorianu. *Stochastic reachability analysis of hybrid systems*. Springer Science & Business Media, 2012.
11. L. Cardelli. On process rate semantics. *Theoretical Computer Science*, 391(3):190–215, 2008.
12. L. Cardelli, M. Kwiatkowska, and L. Laurenti. Stochastic analysis of chemical reaction networks using linear noise approximation. In *Computational Methods in Systems Biology*, pages 64–76. Springer, 2015.
13. A. Csikász-Nagy, L. Cardelli, and O. S. Soyer. Response dynamics of phosphorelays suggest their potential utility in cell signalling. *Journal of The Royal Society Interface*, 8(57):480–488, 2011.
14. A. Eldar and M. B. Elowitz. Functional roles for noise in genetic circuits. *Nature*, 467(7312):167–173, 2010.
15. S. N. Ethier and T. G. Kurtz. *Markov processes: characterization and convergence*, volume 282. John Wiley & Sons, 2009.
16. N. Fedoroff and W. Fontana. Small numbers of big molecules. *Science*, 297(5584):1129–1131, 2002.
17. D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361, 1977.
18. D. T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, 188(1):404–425, 1992.
19. D. T. Gillespie. The chemical Langevin equation. *The Journal of Chemical Physics*, 113(1):297–306, 2000.
20. R. Grima. Linear-noise approximation and the chemical master equation agree up to second-order moments for a class of chemical systems. *Physical Review E*, 92(4):042124, 2015.
21. M. Kwiatkowska, G. Norman, and D. Parker. Stochastic model checking. In *Formal methods for performance evaluation*, pages 220–270. Springer, 2007.
22. M. Kwiatkowska, G. Norman, and D. Parker. Prism 4.0: Verification of probabilistic real-time systems. In *Computer aided verification*, pages 585–591. Springer, 2011.
23. B. Munsky and M. Khammash. The finite state projection algorithm for the solution of the chemical master equation. *The Journal of chemical physics*, 124(4):044104, 2006.
24. A. Pnueli. The temporal logic of programs. In *Foundations of Computer Science, 1977., 18th Annual Symposium on*, pages 46–57. IEEE, 1977.
25. N. G. Van Kampen. *Stochastic processes in physics and chemistry*, volume 1. Elsevier, 1992.
26. E. Wallace, D. Gillespie, K. Sanft, and L. Petzold. Linear noise approximation is valid over limited times for any chemical system that is sufficiently large. *IET systems biology*, 6(4):102–115, 2012.
27. V. Wolf, R. Goel, M. Mateescu, and T. A. Henzinger. Solving the chemical master equation using sliding windows. *BMC systems biology*, 4(1):1, 2010.