Statistical model checking based calibration and analysis of bio-pathway models

Sucheendra K. Palaniappan[1], Benjamin M. Gyori[2], Bing Liu[3], David Hsu[1,2], P.S. Thiagarajan[1,2], and Edmund M. Clarke[3]

[1] School of Computing, National University of Singapore, 117417, Singapore
[2] NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, 117417, Singapore
[3] Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Abstract. We present a statistical model checking (SMC) based framework for studying ordinary differential equation (ODE) models of bio-pathways. We address cell-to-cell variability explicitly by using probability distributions to model initial concentrations and kinetic rate values. The core component of our framework is an SMC procedure for verifying the dynamical properties of an ODE system accompanied by such prior distributions. As an important feature, our specification logic used to formalize properties can encode both qualitative properties and experimental data. This enables us to develop SMC based parameter estimation and sensitivity analysis procedures. We have evaluated our method on two large pathway models, namely, the segmentation clock network and the MLC phosphorylation pathway. The results show that our method scales well and yields good parameter estimates that are robust. Our sensitivity analysis framework leads to interesting insights about the underlying dynamics of these systems.

1 Introduction

Biochemical networks – often called bio-pathways – govern a variety of cellular functions. Their malfunctioning can lead to major diseases [1]. Thus it is important to understand their dynamics using mathematical models [2]. However, building and analyzing such models poses considerable challenges. In this paper, we address the particular challenge of accounting for variable behavior across individual cells. A natural way to cater for this is to use a probabilistic system model such as continuous time Markov chains (CTMCs) [3]. However, such models typically track the occurrences of individual reactions. Hence for pathways of realistic size, calibrating these models using experimental data and analyzing them using stochastic simulations is very difficult. The alternative is to use ordinary differential equations (ODEs) to capture the dynamics. This approach is often computationally more tractable, although it requires that the number of molecules of each type involved in the pathway be abundantly present [4]. In this paper our focus is on accounting for cell-to-cell variability in the setting of ODEs based models.

Variability in a population of cells has at least two major causes. First, as shown in [5], differences in the initial concentrations of proteins are the primary source of variability in the response to external stimuli. Second, due to differing internal and external conditions among cells, the values of kinetic rate constants also vary across cells [6]. In our ODEs setting the variables will represent the concentrations of the biochemical species (typically proteins) in the pathway, and hence the initial concentrations of these species will constitute the initial values of the variables. Further, the parameters appearing in the equations will consist of the kinetic rate constants governing the reactions.

Thus we can capture cell-to-cell variability in the behavior of the bio-pathway by studying the ODE dynamics across a range of values for the initial concentrations and kinetic rate constant values. We do this in a probabilistic setting by assuming initial probability distributions (usually uniform) over an interval of values for the initial concentrations and rate constants. We then show that the resulting space of trajectories can be used to construct a natural probability measure space if the vector field defined by the ODE system is continuously differentiable. In our setting this requirement is easily met. Consequently we can devise a statistical model checking (SMC) procedure to check if the set of trajectories that satisfy a given property specified in our specification logic passes a statistical test whose strength is chosen by the user. In this sense we construct a principled method for analyzing the dynamics of a bio-pathway in the presence of dynamic variability across a population of cells.

To demonstrate the applicability of our approach, we develop an SMC based parameter estimation method. The unknown model parameters usually consist of initial concentrations and kinetic rate constants. Here, for convenience, we shall assume all the initial concentrations are known but that their nominal values can vary over a cell population. The parameter estimation procedure searches through the value space of the unknown parameters to determine the "best" combination of values that can explain the given data and predict new behaviors [7]. The key step in this procedure is to determine the fitness-to-data of the current set of parameter values. We use our specification logic to encode both experimental time series data and known qualitative trends concerning the dynamics of the pathway. We then use our SMC procedure to determine the goodness of the given set of parameter values, while taking into account that these values can fluctuate across the population of cells that the data is based on. Subsequently, we use a global optimization strategy known as SRES [8] to choose a new set of candidate parameter values according to the SMC based score assigned to the current set.

An important analysis task to be performed on the model is quantifying the influence of different parameters on the model dynamics. The information gained from such a sensitivity analysis procedure can help in robustness analysis, optimal experimental design and drug target selection [9]. We show how SMC can be used to generate the statistics needed by the global sensitivity analysis method MPSA [10]. Consequently, one can incorporate a rich class of dynamic behaviors – encoded as formulas in our specification logic – to drive our sensitivity analysis method.

We evaluated our method on two pathway models taken from the BioModels database [11]. For both case studies, we assumed that noisy experimental data and qualitative dynamic traits of a few species were known. This data was separated into training and test components. A subset of the rate constants were assumed to be unknown and estimated using our parameter estimation procedure. The first model, the segmentation clock pathway, consists of 16 differential equations and 75 rate constants, out of which 39 were fixed to be unknown. The second model, the thrombin dependent MLC pathway consists of 105 differential equations and 197 rate constants, out of which 100 were fixed to be unknown. Our results (Section 5) show that our SMC based technique is efficient and scales well. We also applied our sensitivity analysis method to obtain interesting insights into the dynamics of these two bio-pathways.

1.1 Related work

Probabilistic model checking of stochastic models is an active field of research [12–15]. Of particular interest in our context are sampling based methods such as [16, 17], which verify probabilistic properties using a fixed number of sampled trajectories. In contrast, SMC based methods such as [12, 18] adaptively generate a sufficient number of trajectories to determine if the property is satisfied while meeting the strengths of the statistical test specified by the user.

Turning to parameter estimation, a brute force search of the parameter space is employed in [14] for Petri nets. In the ODE context, parameter estimation combined with model checking appears in [19] using again a brute force sampling based parameter search approach, and in [20], using an evolutionary strategy to guide the search. However, both these techniques only generate a single simulation trace of the ODE to evaluate a proposed set of parameters. A symbolic model checking approach is explored for the restricted class of multi-affine ODEs in [21,22]. The work reported in [17] deploys a genetic algorithm to search for the best set of parameters. A fixed number of samples – this number is fixed in an ad hoc manner – is generated, and the probability of satisfying a property is calculated to be the fraction of the samples which satisfy the property. In all these studies, the quality of the estimated parameters is not validated using test data (i.e. data that was not used as training data). While [17] does mention identifying critical parameters, we believe that our approach is the first systematic attempt to develop a property-based sensitivity analysis framework using statistical model checking.

In the next section, we introduce ODE models and their dynamics. In Section 3, we discuss our specification logic and the statistical model checking procedure. Subsequently, we present our parameter estimation and sensitivity analysis framework. Experimental results are reported in Section 5. Detailed proofs are available in the Appendix, and additional experimental results are reported in the supplementary material [23].

2 ODE based models and their behaviors

A popular formalism for describing the dynamics of a biochemical network is a system of ODEs. For each molecular species x_i in the pathway, there will be an equation of the form $dx_i/dt = f_i(\mathbf{x}, \Theta_i)$. Here f_i describes the kinetics of the reactions that produce and consume x_i, $\mathbf{x}$ denotes the concentrations of the molecular species taking part in these reactions, while the vector Θ_i gives the rate constants governing these reactions.

Each x_i is a real-valued function of $t \in \mathbb{R}_+$, where $\mathbb{R}_+$ denotes the set of non-negative reals. We shall realistically assume that $x_i(t)$ takes values in the interval $[L_i, U_i]$, where L_i and U_i are non-negative rationals with $L_i < U_i$. Hence the state space of the system is $\mathbf{V} = [L_1, U_1] \times \ldots \times [L_n, U_n]$, a bounded subset of $\mathbb{R}^n_+$. Let $\Theta = \bigcup_i \Theta_i = \{\theta_1, \theta_2, \ldots, \theta_m\}$ be the set of all rate constants. We again assume that the range of values for each θ_j is $[L^j, U^j]$ for $1 \leq j \leq m$. We shall present the SMC procedure while assuming that all the rate constants are known. In Section 4, it will become clear how unknown rate constants are handled.

An implicit assumption in what follows is that the value of a rate constant, when fixed initially, does not change during the time evolution of the dynamics, although this value can be different for different cells. To capture the cell-to-cell variability regarding the initial states, we define for each variable x_i an interval $[L_i^{init}, U_i^{init}]$ with $L_i \leq L_i^{init} < U_i^{init} \leq U_i$. The actual value of the initial concentration of x_i is assumed to fall in this interval. Similarly, we shall assume that the nominal value of the rate

constant θ_j falls in the interval $[L^j_{init}, U^j_{init}]$ with $L^j \leq L^j_{init} < U^j_{init} \leq U^j$. We set $INIT = (\prod_i [L^{init}_i, U^{init}_i]) \times (\prod_j [L^j_{init}, U^j_{init}])$. Thus $INIT$ captures the cell-to-cell variability in the initial concentration and the rate constant values. In what follows we let $\mathbf{v}$ to range over $\prod_i [L^{init}_i, U^{init}_i]$ and $\mathbf{w}$ to range over $\prod_j [L^j_{init}, U^j_{init}]$

We will represent our system of ODEs in vector form as $d\mathbf{x}/dt = F(\mathbf{x}, \Theta)$, with $F_i(\mathbf{x}, \Theta) := f_i$. Recall that a function $f : \mathbf{V} \to \mathbf{V}$ is a C^1 function if f', the derivative of f, exists at all $\mathbf{v} \in \mathbf{V}$ and is a continuous function. In the setting of biochemical networks, the expressions in f_i will model kinetic laws such as mass law and Michaelis-Menten [4]. Thus it is reasonable to assume that each f_i is composed out of rational functions, which would imply that $f_i \in C^1$ for each i, and hence $F : \mathbf{V} \to \mathbf{V}$ is also a C^1 function. As a result, for each $(\mathbf{v}, \mathbf{w}) \in INIT$, the system of ODEs will have a unique solution $\mathbf{X}_{\mathbf{v},\mathbf{w}}(t)$ [24], viewed as a function $\mathbf{X} : \mathbb{R} \to \mathbf{V}$. Further, it will satisfy: $\mathbf{X}_{\mathbf{v},\mathbf{w}}(0) = \mathbf{v}$ and $\mathbf{X}'_{\mathbf{v},\mathbf{w}}(t) = F(\mathbf{X}_{\mathbf{v},\mathbf{w}}(t))$. We are also guaranteed that $\mathbf{X}_{\mathbf{v},\mathbf{w}}(t)$ is a C^0-function (i.e. continuous function) [24], and hence measurable. This fact will be crucial for our SMC procedure.

In our application, the dynamics will be of interest only up to a maximal time point T. Fixing such a T, a *trajectory* starting from $\mathbf{v} \in \mathbf{V}$ at time 0 and with $\mathbf{w}$ as the parameter values is denoted $\sigma_{\mathbf{v},\mathbf{w}}$. It is the (continuous) function $\sigma_{\mathbf{v},\mathbf{w}} : [0, T] \to \mathbf{V}$ satisfying: $\sigma_{\mathbf{v},\mathbf{w}}(t) = X_{\mathbf{v},\mathbf{w}}(t)$. The behavior of our dynamical system is the set of trajectories given by $BEH = \{\sigma_{\mathbf{v},\mathbf{w}} \mid (\mathbf{v}, \mathbf{w}) \in INIT\}$. Our goal is to develop an SMC procedure to verify the dynamical properties of BEH.

3 Statistical model checking of ODE dynamics

3.1 Bounded linear time temporal logic

To formally express dynamical properties of BEH, we use formulas in a specification logic. We will use bounded linear time temporal logic (BLTL) since our trajectories will be of finite duration. An atomic proposition in our logic will be of the form (i, ℓ, u) with $L_i \leq \ell < u \leq U_i$. Such a proposition will be interpreted as "the current concentration level of x_i falls in the interval $[\ell, u]$", and we fix a finite set of such atomic propositions.

We first introduce the syntax and then the semantics of BLTL formulas. The formulas of BLTL are defined as: (i) Every atomic proposition as well as the constants *true*, *false* are BLTL formulas. (ii) If ψ, ψ' are BLTL formulas then $\neg\psi$ and $\psi \vee \psi'$ are BLTL formulas. (iii) If ψ, ψ' are BLTL formulas and $t \leq T$ is a *positive integer* then $\psi \mathbf{U}^{\leq t} \psi'$ and $\psi \mathbf{U}^{t} \psi'$ are BLTL formulas. We have mildly strengthened BLTL to be able to express that a certain property will hold exactly at t time units from now. This will enable us to encode experimental data in the specification. The derived propositional operators such as $\wedge$, $\supset$, $\equiv$, and the temporal operators $\mathbf{G}^{\leq t}$, $\mathbf{F}^{\leq t}$ are defined in the usual way.

We will interpret the formulas of our logic at the finite set of time points $\mathcal{T} = \{0, 1, \ldots, T\}$. Such a discretization is reasonable since experimental data will be available only at a finite number of discrete time points. Further, qualitative properties of interest are expressible in discrete time. We assume that $\mathcal{T}$ has been chosen appropriately and it includes all the relevant time points with respect to the specified properties.

The semantics of the logic is defined in terms of the relation $\sigma, t \models \varphi$, where σ is a trajectory in BEH and $t \in \mathcal{T}$.

- $\sigma, t \models (i, \ell, u)$ iff $\ell \leq \sigma(t)(i) \leq u$ where $\sigma(t)(i)$ is the i^{th} component of the n-dimensional vector $\sigma(t) \in \mathbf{V}$.

- $\neg$ and $\vee$ are interpreted in the usual way.
- $\sigma, t \models \psi \mathbf{U}^{\leq k} \psi'$ iff there exists k' such that $k' \leq k,\ t + k' \leq T$ and $\sigma, t + k' \models \psi'$. Further, $\sigma, t + k'' \models \psi$ for every $0 \leq k'' < k'$.
- $\sigma, t \models \psi \mathbf{U}^{k} \psi'$ iff $t + k \leq T$ and $\sigma, t + k \models \psi'$. Further, $\sigma, t + k' \models \psi$ for every $0 \leq k' < k$.

We now define $models(\psi) = \{\sigma \mid \sigma, 0 \models \psi,\ \sigma \in BEH\}$.

Next, we wish to make statements of the form $P_{\geq r}(\psi)$, where the intended meaning is that the probability that a trajectory in BEH belongs to $models(\psi)$ is at least r. To assign meaning to such statements, we need to define a probability measure over sets of trajectories. Note, however, that the trajectory $\sigma \in BEH$ is completely determined by $\sigma(0)$, the (vector) value it assumes at $t = 0$. Hence we will identify BEH with $INIT$, the set of initial states. To make this explicit, we define the set $Models(\psi) \subseteq INIT$ as: $(\mathbf{v}, \mathbf{w}) \in Models(\psi)$ iff $\sigma_{\mathbf{v},\mathbf{w}} \in models(\psi)$. We define the formulas of PBLTL as $P_{\geq r}\psi$ and $P_{\leq r'}\psi$ provided $r \in [0, 1)$, $r' \in (0, 1]$ and ψ is a BLTL formula. We shall say that $\mathcal{S}$, the system of ODEs, meets the specification $P_{\geq r}\psi$ – and this is denoted $\mathcal{S} \models P_{\geq r}\psi$ – iff $P(Models(\psi)) \geq r$, while $\mathcal{S} \models P_{\leq r'}\psi$ iff $P(Models(\psi)) \leq r'$. Here, and in what follows, P is the standard probability measure assigned to members of the σ-algebra generated by the open intervals contained in $INIT$. It is easy to show that $Models(\psi)$ is a member of this σ-algebra for every ψ. The only case that requires an argument is the one for atomic propositions, and here the measurability of the solution functions $\mathbf{X}_{\mathbf{v},\mathbf{w}}(t)$ is crucial. The details can be found in the Appendix.

3.2 Statistical model checking of PBLTL formulas

According to [25], whether $\mathcal{S} \models P_{\geq r}\psi$, can be formulated as a sequential hypothesis test between the null hypothesis H0 : $p \geq r + \delta$ and the alternative hypothesis H1 : $p \leq r - \delta$, where $p = P(Models(\psi))$. Here, δ signifies the indifference region supplied by the user. The *strength* of the test is decided by parameters α and β which bound the Type-I (false positive) and Type-II (false negative) errors respectively. Thus the verification is carried out approximately but with guaranteed confidence levels and error bounds. The test proceeds by generating a sequence of sample trajectories $\sigma_1, \sigma_2, \ldots$ by randomly sampling an initial state from $INIT$. One assumes a corresponding sequence of Bernoulli random variables $y_1, y_2 \ldots$, where each y_k is assigned the value 1 if $\sigma_k, 0 \models \psi$. Otherwise y_k is assigned the value 0. For each $m \geq 1$, after drawing m samples, we compute a quantity q_m as:

$$q_m = \frac{[r-\delta]^{(\sum_{i=1}^{m} y_i)}[1-[r-\delta]]^{(m-\sum_{i=1}^{m} y_i)}}{[r+\delta]^{(\sum_{i=1}^{m} y_i)}[1-[r+\delta]]^{(m-\sum_{i=1}^{m} y_i)}} \tag{1}$$

Hypothesis H0 is accepted if $q_m \geq \widehat{A}$, and hypothesis H1 is accepted if $q_m \leq \widehat{B}$. If neither is the case then another sample is drawn. The constants $\widehat{A}$ and $\widehat{B}$ are chosen such that it results in a test of strength (α, β). In practice, a good approximation is $\widehat{A} = \frac{1-\beta}{\alpha}$ and $\widehat{B} = \frac{\beta}{1-\alpha}$. A detailed account of our *on-line* model checking algorithm (used to verify each trajectory) can be found in the Appendix.

4 Analysis methods

Here we present our parameter estimation and sensitivity analysis methods. In doing so, we assume the terminology and notations developed in the previous sections. As

a first step, we describe how experimental data can be encoded as a BLTL formula. Assume, without loss of generality, that $O \subseteq \{x_1, x_2, \ldots, x_k\}$ is the set of variables for which experimental data is available, and which has been allotted as training data to be used for parameter estimation. Assume $\mathcal{T}_i = \{\tau_1^i, \tau_2^i, \ldots, \tau_{T_i}^i\}$ are the time points at which the concentration level of x_i has been measured and reported as $[\ell_t^i, u_t^i]$ for each $t \in \mathcal{T}_i$. The interval $[\ell_t^i, u_t^i]$ is chosen to reflect the noisiness, the limited precision and the cell-population based nature of the experimental data. For each $t \in \mathcal{T}_i$, we define the formula $\psi_i^t = \mathbf{F}^t(i, \ell_t^i, u_t^i)$. Then $\psi_{exp}^i = \bigwedge_{t \in \mathcal{T}_i} \psi_i^t$. We then set $\psi_{exp} = \bigwedge_{i \in O} \psi_{exp}^i$. In case the species x_i has been measured under multiple experimental conditions, the above encoding scheme is extended in the obvious way.

Often qualitative dynamic trends will be available – typically from the literature – for some of the molecular species in the pathway. For instance, we may know that a species shows transient activation, in which its level rises in the early time points, and later falls back to initial levels. Similarly, a species may be known to show oscillatory behavior with certain characteristics. Such information can be described as BLTL formulas that we term to be *trend* formulas. Examples of such formulas can be found in the Appendix. We let ψ_{qlty} to be the conjunction of all the trend formulas.

Finally, we fix the PBLTL formula $P_{\geq r}(\psi_{exp} \wedge \psi_{qlty})$, where r will capture the confidence level with which we wish to assess the goodness of the fit of the current set of parameters to experimental data and qualitative trends. We also fix an indifference region δ and the strength of the test (α, β). The constants r, δ, α and β are to be fixed by the user. In our application, it will be useful to exploit the fact that both ψ_{exp} and ψ_{qlty} are conjunctions, and hence can be evaluated separately. As shown in [25], one can choose the strength of each of these tests to be $(\frac{\alpha}{J}, \beta)$, where J is the total number of conjuncts in the specification. This will ensure that the overall strength of the test is (α, β). Further, the results for the individual statistical tests can be used to compute the objective function associated with the global search strategy to be described below.

4.1 Parameter estimation based on PBLTL specification

We assume $\Theta_u = \{\theta_1, \theta_2, \ldots, \theta_K\}$ is the set of unknown parameters. For convenience we will assume that the other parameter values are known and that their nominal values do not fluctuate across the cell population. We will also assume nominal values for the initial concentrations and the range of their fluctuations of the form $[L_i^{init}, U_i^{init}]$ for each variable x_i. Again, for convenience, we fix a constant δ'' so that if the current estimate of the values of the unknown parameters is $\mathbf{w} \in \prod_{1 \leq j \leq K} [L^j, U^j]$ then this value will fluctuate in the range $[\mathbf{w}(j) - \delta'', \mathbf{w}(j) + \delta'']$. Setting $L_{init,\mathbf{w}}^j = \mathbf{w}(j) - \delta''$ and $U_{init,\mathbf{w}}^j = \mathbf{w}(j) + \delta''$ we define $INIT_{\mathbf{w}} = (\prod_i [L_i^{init}, U_i^{init}]) \times (\prod_j [L_{init,\mathbf{w}}^j U_{init,\mathbf{w}}^j])$. The set of trajectories $BEH_{\mathbf{w}}$ is defined accordingly.

To estimate the quality of $\mathbf{w}$, we run our SMC procedure – using $INIT_{\mathbf{w}}$ instead of $INIT$ – to verify $P_{\geq r}(\psi_{exp} \wedge \psi_{qlty})$. Depending on the outcome of this test for the various conjuncts in the specification, we assign a score to $\mathbf{w}$ using an objective function detailed below. We then iterate this scheme for various values of $\mathbf{w}$ generated using a suitable search strategy.

The objective function is formed as follows. Let J_{exp}^i $(= T_i)$ be the number of conjuncts in ψ_{exp}^i, and J_{qlty} the number of conjuncts in ψ_{qlty}. Let $J_{exp}^{i,+}(\mathbf{w})$ be the number of formulas of the form ψ_i^t (a conjunct in ψ_{exp}^i) such that the statistical test for

$P_{\geq r}(\psi_i^t)$ accepts the null hypothesis (that is, $P_{\geq r}(\psi_i^t)$ holds) with the strength $(\frac{\alpha}{J}, \beta)$, where $J = \sum_{i \in O} J_{exp}^i + J_{qlty}$. Similarly, let $J_{qlty}^+(\mathbf{w})$ be the number of conjuncts in ψ_{qlty} of the form $\psi_{\ell,qlty}$ that pass the statistical test $P_{\geq r}(\psi_{\ell,qlty})$ with the strength $(\frac{\alpha}{J}, \beta)$. Then $\mathcal{G}(\mathbf{w})$ is computed via:

$$\mathcal{G}(\mathbf{w}) = J_{qlty}^+(\mathbf{w}) + \sum_{i \in O} \frac{J_{exp}^{i,+}}{J_{exp}^i} \tag{2}$$

Thus the goodness to fit of $\mathbf{w}$ is measured by how well it agrees with the qualitative properties as well as the number of experimental data points with which there is acceptable agreement. To avoid over-training the model, we do not insist that every qualitative property and every data point must fit well with the dynamics predicted by $\mathbf{w}$.

The search strategy to evolve candidate parameters will use the values $\mathcal{G}(\mathbf{w})$ to traverse the parameter value space. Global search methods such as Genetic Algorithms (GA) [26], and Stochastic Ranking Evolutionary Strategy (SRES) [8] are computationally more intensive than local methods, but are much better at avoiding local minima. The overall structure of our parameter estimation procedure is presented in Algorithm 1. In practice, one usually maintains a *population* of parameter value vectors in each round, and a round is usually called a *generation*. For convenience, we have assumed that each population is a singleton in the description of Algorithm 1. We use the SRES strategy in our work since it is known to perform well in the context of pathway models [7]. The particular choice of search algorithm, however, is orthogonal to our proposed method.

4.2 Sensitivity analysis based on PBLTL specification

As another application of our SMC procedure, we have constructed a property based sensitivity analysis method by coupling our SMC routine with the global sensitivity analysis technique called multi-parametric sensitivity analysis (MPSA) [10]. We assume we have specified a set of properties (encoded as PBLTL formulas), and are interested in knowing which parameters, when changed, affect these properties significantly. The MPSA procedure involves sampling a large number of parameter combinations from their valid ranges. For each sampled combination, one calculates the objective value with respect to the PBLTL properties according to Equation 2. The objective values allow us to assess the extent to which each parameter affects the model's behavior to the given formulas. Intuitively, if the objective value shows strong dependence on the value of a parameter (over its range) then the output is sensitive to that parameter. The MPSA method employs statistical tests to quantify this dependence, which can be directly interpreted as a measure of sensitivity. The sensitivity is based on computing the Kolmogorov-Smirnov (KS) test to compare the two profiles consisting of (a) the cumulative appearance of *good* intervals along the value space of the parameter and (b) the same for the *bad* intervals. If these profiles differ significantly then the system is more sensitive to this parameter, and the KS test will assign a higher score to this parameter. Our procedure is outlined in Algorithm 2.

5 Results

We applied our SMC based analysis framework to pathway models taken from the BioModels database [11]. These models have nominal point values for all the rate constants and initial concentrations. We first verified a few properties of the two pathways

input : ODE model; PBLTL formulas; SMC parameters; Number of generations k; Initial parameter guess $\mathbf{w}_0$;
output: The best parameter found $\mathbf{w}_{\max}$

initialization: $\ell = 0$; $\mathcal{G}_{\max} = 0$;

while $\ell < k$ **do**
  Run SMC on the trajectories defined by $BEH_{\mathbf{w}_\ell}$ with respect to the PBLTL formulas;
  Compute $\mathcal{G}(\mathbf{w}_\ell)$;
  if $\mathcal{G}(\mathbf{w}_\ell) \geq \mathcal{G}_{\max}$ **then**
    $\mathbf{w}_{\max} = \mathbf{w}_\ell$;
    $\mathcal{G}_{\max} = \mathcal{G}(\mathbf{w}_\ell)$
  end
  $\mathbf{w}_{\ell+1}$ = Picked by SRES / GA search procedure based on $\mathbf{w}_\ell$;
  $\ell = \ell + 1$;
end

Algorithm 1. Parameter estimation

input : ODE model; PBLTL formulas; SMC parameters; Number of discretization intervals N_d; Objective function $\mathcal{G}$; threshold
output: Sensitivity$[1 \ldots K]$

Discretize each parameter into N_d intervals to get $(N_d)^K$ hypercubes;
for $i \leftarrow 0$ **to** N_d **do**
  $\mathbf{w}_i$ = Sample one hypercube out the $(N_d)^K$ using LHS;
  Run SMC on $BEH_{\mathbf{w}_i}$; Calculate $\mathcal{G}(\mathbf{w}_i)$;
  if $\mathcal{G}(\mathbf{w}_i) >$ threshold **then**
    Add $\mathbf{w}_i$ to *good* set;
  else
    Add $\mathbf{w}_i$ to *bad* set;
  end
end
for $j \leftarrow 0$ **to** K **do**
  Construct cumulative distribution of *good* and *bad* intervals in the range of parameter j;
  Sensitivity$[j]$ = KS statistic of difference of the two distributions;
end

Algorithm 2. Sensitivity analysis

using SMC. Then, for parameter estimation, we formulated qualitative trends for some species, and generated synthetic experimental data for some other species as follows. We set a $\pm 5\%$ range around the nominal value for the initial concentration of each species and assumed a uniform distribution over the resulting set of initial states. To mimic western blot data, which is cell population based, we averaged 10^4 random trajectories generated by sampling these initial concentration intervals. We then added noise to the data and used a major portion of it for training, and reserved the rest as test data. Finally, we fixed a subset of rate constants to be unknown, and ran our parameter estimation procedure. We let the variability in parameters (δ'') to be 0.5% of the proposed value.

We implemented our method using MATLAB and C++ on a PC with a 3.4Ghz Intel Core i7 processor with 8GB RAM. ODE systems were numerically solved using the SUNDIALS CVODE package [27–29]. The source code is available at [23]. The code has been optimized to take advantage of the multi-core architecture; all experimental results were run on 8 threads. The parameters used for the statistical model checking algorithm were $r = 0.9$, $\alpha = \beta = \delta = 0.05$ for all our experiments. To show the goodness of our estimated parameters (taking into account the variability concerning the initial states and reaction rates), we generated 1000 trajectories and plotted these to show that the estimated parameters result in a good fit to the data. In each case, experimental data is plotted along with the tolerance interval used in constructing the specification.

For the experiments reported in this section, we used an SRES based global strategy to guide the search. Here we present only the highlights of our experimental results. Many further details including the results obtained using a Genetic Algorithm based search can be found in the supplementary material [23].

5.1 The case studies

The segmentation clock network An oscillating segmentation clock governs the segmentation pattern of the spine in developing vertebrate embryos. It couples signaling pathways of FGF, Notch and Wnt, whose periodic behaviors are produced by negative feedback loops. The ODE model consists of 16 differential equations and 75 kinetic

rate parameters. Simulation time (T) was fixed at 200 minutes divided into 40 equally spaced time points.

The thrombin-dependent MLC phosphorylation pathway Endothelial cells form a dynamic barrier between blood/lymph and the underlying connective tissue, and their contraction plays a crucial role in physiological and pathological processes. Agonists such as thrombin play an important role in the contraction function through phosphorylation of MLC, while Rho-kinase is crucial for the sustained contraction of endothelial cells. The pathway model with 105 differential equations and 197 kinetic parameters is considerably large. Simulation time was fixed at 1000 seconds divided into 20 equally spaced time points.

5.2 Statistical model checking based verification

First, we used our SMC framework to verify pathway properties expressed in PBLTL. We used the nominal models (all rate parameter values known, taken from the BioModels database) to verify if they conformed to properties expressed in our logic. For instance, for the MLC phosphorylation pathway, it is known experimentally that the concentration of phosphorylated MLC starts at a low level, and then reaches a high steady state value. Our SMC method shows that the nominal model does not satisfy the property, instead, phosphorylated MLC exhibits a transient profile. This discrepancy has been studied in [30], and attributed to missing components and links in the proposed model. Details of these properties and their verification is presented in the Appendix.

5.3 Parameter estimation

For the segmentation clock pathway, we assumed 39 of the rate parameters as unknown. We used a combination of dynamic trends and quantitative experimental data. Specifically, we synthesized population based experimental time series data for Axin2 mRNA measured at 14 time points up to 165 minutes. For 5 other species {Notch protein, nuclear NicD, Lunatic fringe mRNA, active ERK and Dusp6 mRNA}, we encoded the dynamic trends as properties in our logic. The dynamic trend of 2 species (cytosolic NicD and Dusp6 protein) were used as test data. Parameter estimation was done with a population of 200 per generation and for 300 generations. The time taken by SRES based search was 2.3 hours. Figure 1 shows simulation profiles with the estimated parameters. Figure 1(a) shows that the model fits training data consisting of the experimental data of Axin2 mRNA and qualitative trends for 3 other species. Figure 1(b) shows dynamic trends of cytosolic NicD used for testing. The simulated time profiles fit the specified test properties (see [23]).

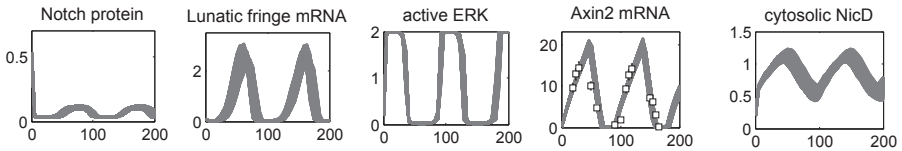

Fig. 1. Parameter estimation results of the segmentation clock pathway. (a) Training data including the experimental data for Axin2 mRNA and the dynamic trends for 3 species), and (b) the test data for one of the species.

To illustrate the scalability of our approach, for the thrombin pathway, we assumed 100 of the kinetic parameters to be unknown. We synthesized population based experimental time series data for 10 species including RGS_2, Rho.GTP, PKC.DAG, MLC_2, CPI-17, Ca-super-2-plus, p115RhoGEF-GTP-alpha, MYPT1-PPase, Rho-kinase.MLC, MYPT1.Rho-$kinase_2$. For thrombinR-active and 3IP3.IP3R we assumed that only the dynamic trend is known. The data of Rho-kinase.MLC and MYPT1.Rho-$kinase_2$ were reserved as test data to evaluate the quality of our parameter estimates, while the data of all other species was used to calibrate the model. Parameter estimation was done with a population of 100 per generation and for 1000 generations. The time taken by SRES based search was 48.8 hours. Figure 2 shows the fit to data of the simulation profiles with the best predicted parameter values for both the training data (Figure 2(a)) and the test data (Figure 2(b)).

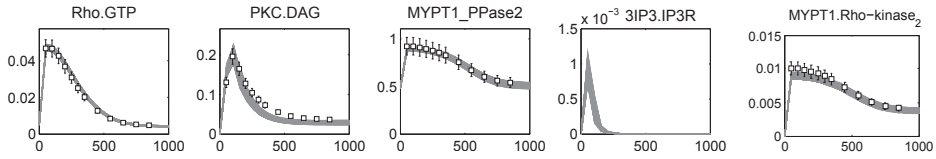

Fig. 2. Parameter estimation results of Thrombin-dependent MLC phosphorylation pathway. (a) Training data, including experimental data of 3 species and dynamic trends of one species, and (b) the test data for one of the species.

5.4 Property based sensitivity analysis

Here we report results just for the segmentation clock pathway. We evaluated the sensitivity of parameters against all properties used for parameter estimation. The results are shown in Figure 3(a). It can be seen that the most sensitive parameters are *ksDusp*, *kcDusp*, *VMsMDusp*, *VMdMDusp*, *VMaX*, *VMdX*. This also indicates that the reactions involving Dusp6 degradation and transcription affect the overall dynamics most. Since all these parameters belong to the FGF pathway, we hypothesize that FGF pathway is the most crucial component that drives the behavior of the system.

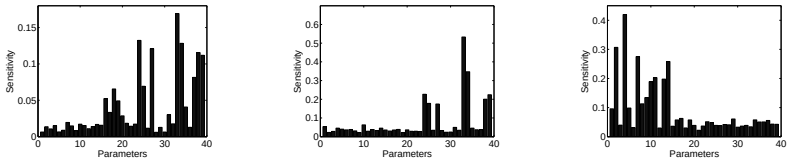

Fig. 3. Sensitivity analysis results. (a-c) Parameter sensitivities of the segmentation clock pathway with respect to (a) all properties, (b) Dusp6mRNA profile, and (c) nuclear nicD profile.

We next searched for parameters affecting the oscillatory property of *Dusp6 mRNA* alone. We found that the same set of parameters as above are the most crucial (see Figure 3(b)). However, when evaluating the oscillatory property of nuclear NicD (Figure 3(c)), we find that the parameters *vsN*, *kt1*, *VdNan* are the most significant. This

suggests that although the Notch synthesis (*vsN*), and nuclear NicD transportation (*kt1*) and degradation (*VdNan*) do not significantly affect the overall dynamics, they play a dominant role in segmentation patterning.

6 Conclusion

We have proposed an SMC based approach for studying ODEs based bio-pathway models. We have used the temporal logic BLTL to encode both quantitative experimental data and qualitative properties of pathway dynamics. To cater for variability among cells, we assume a uniform distribution over a set of initial states and kinetic rate constants – and impose a reasonable continuity restriction – and show how the probability of the property being met by the behavior of the model can be assessed using an SMC procedure. By combining this method with a global search strategy, we arrive at a parameter estimation procedure as well as a sensitivity analysis technique.

We have demonstrated the applicability of our method with the help of two ODEs based bio-pathway models: the segmentation clock network and the thrombin-dependent MLC phosphorylation pathway. Our method successfully obtained good parameter estimates using noisy cell-population data and qualitative knowledge. The results show that our method scales well and can cope with large biological networks. We also show results for performing property based sensitivity analysis, and thereby gain interesting insights about the pathway dynamics that would be difficult to obtain using conventional approaches.

Our parameter estimation method is a generic one and has the potential to be applied to model classes such as continuous time Markov chain (CTMC) models and stochastic differential equation (SDE) models [3]. We plan to explore this in our future work. Another interesting direction will be to develop a GPU-based implementation of our method to exploit the inherent massive parallelism in generating trajectories through numerical integration. In this connection, the platform-aware implementation of a related systems biology application presented in [15] promises to offer helpful pointers.

References

1. De Ferrari, G.V., Inestrosa, N.C.: Wnt signaling function in Alzheimer's disease. Brain Res. Rev. **33** (2000) 1–12
2. Aldridge, B.B., Burke, J.M., Lauffenburger, D.A., Sorger, P.K.: Physicochemical modelling of cell signalling pathways. Nat. Cell Biol. **8**(11) (2006) 1195–1203
3. Wilkinson, D.: Stochastic modelling for systems biology. CRC Press (2011)
4. Klipp, E., Herwig, R., Kowald, A., Wierling, C., Lehrach, H.: Systems biology in practice: concepts, implementation and application. Wiley-VCH, Weinheim (2005)
5. Spencer, S., Gaudet, S., Albeck, J., Burke, J., Sorger, P.: Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. Nature **459**(7245) (2009) 428–432
6. Gunawardena, J. Models in systems biology: the parameter problem and the meanings of robustness. In: Elements of computational systems biology. John Wiley & Sons (2010)
7. Moles, C.G., Mendes, P., Banga, J.R.: Parameter estimation in biochemical pathways: A comparison of global optimization methods. Genome Res. **13**(11) (2003) 2467 –2474
8. Runarsson, T., Yao, X.: Stochastic ranking for constrained evolutionary optimization. IEEE T. Evolut. Comput. **4** (2000) 284–294
9. Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S.: Global sensitivity analysis: the primer. Wiley-Interscience (2008)
10. Cho, K.H., Shin, S.Y., Kolch, W., Wolkenhauer, O.: Experimental design in systems biology, based on parameter sensitivity analysis using a Monte Carlo method: A case study for the TNFα-mediated NF-κB signal transduction pathway. Simulation **79**(12) (2003) 726–739

11. Le Novere, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., Snoep, J., Hucka, M.: BioModels Database: A free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic Acids Res. **34** (2006) D689–D691
12. Jha, S.K., Clarke, E.M., Langmead, C.J., Legay, A., Platzer, A., Zuliani, P.: A Bayesian approach to model checking biological systems. In: CMSB. Springer Berlin / Heidelberg (2009) 218–234
13. Heath, J., Kwiatkowska, M., Norman, G., Parker, D., Tymchyshyn, O.: Probabilistic model checking of complex biological pathways. Theor Comput Sci **391**(3) (2008) 239 – 257
14. Li, C., Nagasaki, M., Koh, C.H., Miyano, S.: Online model checking approach based parameter estimation to a neuronal fate decision simulation model in Caenorhabditis elegans with hybrid functional Petri net with extension. Mol. Biosyst. **7**(5) (2011) 1576–92
15. Liu, B., Hagiescu, A., Palaniappan, S.K., Chattopadhyay, B., Cui, Z., Wong, W., Thiagarajan, P.S.: Approximate probabilistic analysis of biopathway dynamics. Bioinformatics **28**(11) (2012) 1508–1516
16. Donaldson, R., Gilbert, D.: A monte carlo model checker for probabilistic ltl with numerical constraints. University of Glasgow, Dep. of CS, Tech. Rep (2008)
17. Donaldson, R., Gilbert, D.: A model checking approach to the parameter estimation of biochemical pathways. In: CMSB. Springer Berlin / Heidelberg (2008) 269–287
18. Clarke, E.M., Faeder, J.R., Langmead, C.J., Harris, L.A., Jha, S.K., Legay, A.: Statistical model checking in BioLab: Applications to the automated analysis of T-cell receptor signaling pathway. In: CMSB, Springer Berlin / Heidelberg (2008) 231–250
19. Calzone, L., Chabrier-Rivier, N., Fages, F., Soliman, S.: Machine learning biochemical networks from temporal logic properties. T. Comput. Syst. Biol. VI (2006) 68–94
20. Rizk, A., Batt, G., Fages, F., Soliman, S.: On a continuous degree of satisfaction of temporal logic formulae with applications to systems biology. In: CMSB, Springer Berlin / Heidelberg (2008) 251–268
21. Batt, G., Page, M., Cantone, I., Goessler, G., Monteiro, P., de Jong, H.: Efficient parameter search for qualitative models of regulatory networks using symbolic model checking. Bioinformatics **26**(18) (2010) i603 –i610
22. Barnat, J., Brim, L., Krejci, A., Streck, A., Safranek, D., Vejnar, M., Vejpustek, T.: On parameter synthesis by parallel model checking. IEEE/ACM T. Comput. Bi. **9**(3) (2012) 693–705
23. Supplementary information and source code. `http://www.comp.nus.edu.sg/~rpsysbio/SMC/`
24. Hirsch, M., Smale, S., Devaney, R.: Differential equations, dynamical systems, and an introduction to chaos. Academic Press (2012)
25. Younes, H.L.S., Simmons, R.G.: Statistical probabilistic model checking with a focus on time-bounded properties. Inform. Comput. **204** (2006) 1368–1409
26. Goldberg, D.: Genetic algorithms in search, optimization, and machine learning. Addison-Wesley (1989)
27. Hindmarsh, A., Brown, P., Grant, K., Lee, S., Serban, R., Shumaker, D., Woodward, C.: SUNDIALS: Suite of nonlinear and differential/algebraic equation solvers. ACM T. Math. Software **31**(3) (2005) 363–396
28. Vanlier, J., Tiemann, C., Hilbers, P., van Riel, N.: An integrated strategy for prediction uncertainty analysis. Bioinformatics **28**(8) (2012) 1130–1135
29. van Riel, N.: Speeding up simulations of ODE models in Matlab using CVode and MEX files (2012)
30. Maedo, A., Ozaki, Y., Sivakumaran, S., Akiyama, T., Urakubo, H., Usami, A., Sato, M., Kaibuchi, K., Kuroda, S.: Ca^{2+}-independent phospholipase A2-dependent sustained Rho-kinase activation exhibits all-or-none response. Genes Cells **11** (2006) 1071–1083
31. Wheeden, R., Zygmund, A.: Measure and integral: an introduction to real analysis. CRC Press (1977)

Appendix

ODE dynamics

As described in the main text, we represent our system of ODEs in vector form as $d\mathbf{x}/dt = F(\mathbf{x}, \Theta)$ with $F_i(\mathbf{x}, \Theta) := f_i$. A function $f : \mathbf{V} \to \mathbf{V}$ is a C^1 function if f', the derivative of f, exists at all $\mathbf{v} \in \mathbf{V}$, and is a continuous function. In the setting of biochemical networks, the expressions in f_i will model kinetic laws such as mass law and Michaelis-Menten [4]. Thus it is reasonable to assume each f_i is composed out of rational functions which would imply that $f_i \in C^1$ for each i and hence $F : \mathbf{V} \to \mathbf{V}$ is also a C^1 function. As a result, for each $(\mathbf{v}, \mathbf{w}) \in INIT$ the system of ODEs will have a unique solution $\mathbf{X}_{\mathbf{v},\mathbf{w}}(t)$ [24]. Further, it will satisfy: $\mathbf{X}_{\mathbf{v},\mathbf{w}}(0) = \mathbf{v}$ and $\mathbf{X}'_{\mathbf{v},\mathbf{w}}(t) = F(\mathbf{X}_{\mathbf{v},\mathbf{w}}(t))$. We are also guaranteed that $\mathbf{X}_{\mathbf{v},\mathbf{w}}(t)$ is a C^0-function (i.e. continuous function) [24].

It will be convenient to define the flow $\Phi_\theta : \mathbb{R}_+ \times \mathbf{V} \to \mathbf{V}$ for arbitrary initial vectors $\mathbf{v}$. Intuitively, $\Phi_{\mathbf{w}}(t, \mathbf{v})$ is the state reached under the ODE dynamics if the system starts at $\mathbf{v}$ at time 0. The flow will be the C^0-function given by: $\Phi_{\mathbf{w}}(t, \mathbf{v}) = \mathbf{X}_{\mathbf{v},\mathbf{w}}(t)$. Thus $\Phi_{\mathbf{w}}(0, \mathbf{v}) = \mathbf{X}_{\mathbf{v},\mathbf{w}(0)=\mathbf{v}}$ and $\partial(\Phi_{\mathbf{w}}(t, \mathbf{v}))/\partial t = F(\Phi_{\mathbf{w}}(t, \mathbf{v}))$ for all t. We will, in fact, work with $\Phi_{\mathbf{w},t} : \mathbf{V} \to \mathbf{V}$ where $\Phi_{\mathbf{w},t}(\mathbf{v}) = \Phi_w(t, \mathbf{v})$ for every t and every $\mathbf{v} \in \mathbf{V}$. again, $\Phi_{\mathbf{w},t}$ is guaranteed to be a C^0 function.

In our application the dynamics will be of interest only up to a maximal time point T. Fixing such a T, we define a *trajectory* starting from $\mathbf{v} \in \mathbf{V}$ denoted $\sigma_{\mathbf{v},\mathbf{w}}$ to be the (continuous) function $\sigma_{\mathbf{v},\mathbf{w}} : [0, T] \to \mathbf{V}$ satisfying: $\sigma_{\mathbf{v},\mathbf{w}}(t) = \Phi_{\mathbf{w},t}(\mathbf{v})$. The behavior of our dynamical system is the set of trajectories given by $BEH = \{\sigma_{\mathbf{v},\mathbf{w}} \mid (\mathbf{v}, \mathbf{w}) \in INIT\}$.

Probability measure on a family of trajectories with respect to BLTL formulas

To be able to assign a probability to $Models(\psi)$, we construct a probability measure over the standard σ-algebra generated by the open intervals contained in $INIT$. More precisely, recall that $INIT = (\prod_i [L_i^{init}, U_i^{init}]) \times (\prod_j [L_{init}^j, U_{init}^j])$. Then $\mathcal{B}(INIT)$ – written for convenience as just $\mathcal{B}$ below – is the smallest subset of 2^{INIT} satisfying (i) if $L_i^{init} \leq \ell_i < u_i \leq U_i^{init}$ for each i, and if $L_{init}^j \leq \ell^j < u^j \leq U_{init}^j$ for each j then $\prod_i (\ell_i, u_i) \times \prod_j (\ell^j, u^j) \in \mathcal{B}$; (ii) if $B \in \mathcal{B}$ then $\overline{B} = INIT - B \in \mathcal{B}$; (iii) if $\{B_k\}$ is a countable family of sets in $\mathcal{B}$ then $\bigcup_k B_k \in \mathcal{B}$.

The probability measure we define over $\mathcal{B}$ will be based on the assumption that each initial state in $INIT$ is equally likely. This so called uniform distribution assumption is made when there is no prior knowledge. Sometimes, however, valuable prior knowledge may be available. For instance, in [5] the initial distribution of a protein's concentration follows a certain log-normal distribution. Such information can be easily incorporated in the prior distribution of initial states. Here we shall work with a uniform distribution mainly for technical convenience.

Now suppose $\prod_i (\ell_i, u_i) \times \prod_j (\ell^j, u^j) \in \mathcal{B}$. We define $P(\prod_i (\ell_i, u_i) \times \prod_j (\ell^j, u^j)) = \prod_i \frac{(u_i - \ell_i)}{(U_i^{init} - L_i^{init})} \times \prod_j \frac{(u^j - \ell^j)}{(U_{init}^j - L_{init}^j)}$. It is a standard fact that P extends in a unique way to the probability measure $P : \mathcal{B} \to [0, 1]$ such that $P(INIT) = 1$ and $P(\emptyset) = 0$. Our goal now is to show that $Models(\psi) \in \mathcal{B}$ for every formula ψ. This will then ensure that $P(Models(\psi))$ is well-defined.

Let ψ be a formula and $t \in \mathcal{T}$. Then $\|\psi\|_t \subseteq INIT$ is defined inductively as follows.

- $\|(i,\ell,u)\|_t = \{(\mathbf{v},\mathbf{w}) \mid \sigma_{(\mathbf{v},\mathbf{w})}, t \models (i,\ell,u)\}$, where $\sigma_{(\mathbf{v},\mathbf{w})}$ is the trajectory in BEH with $\sigma_{(\mathbf{v},\mathbf{w})}(0) = (\mathbf{v},\mathbf{w})$.
- $\|\neg\psi\|_t = INIT - \|\psi\|_t$ and $\|\psi \vee \psi'\|_t = \|\psi\|_t \cup \|\psi'\|_t$
- $\|\psi \mathbf{U}^{\leq k} \psi'\|_t = \bigcup_{k' \leq k, t+k' \leq T} (\|\psi'\|_{t+k'} \cap (\bigcap_{0 \leq k'' < k'} \|\psi\|_{t+k''}))$
- $\|\psi \mathbf{U}^{k} \psi'\|_t = \|\psi'\|_{t+k} \cap (\bigcap_{0 \leq k' < k} \|\psi\|_{t+k'})$ if $t+k \leq T$. Otherwise $\|\psi \mathbf{U}^{k} \psi'\|_t = \emptyset$.

We now recall that due to the fact that each f_i is a C^1 function, $\Phi_{t,\theta} : \mathbf{V} \to \mathbf{V}$ is also a continuous function for every $t \in [0,T]$. This in turn implies Φ_t is in fact a *measurable* function [31] in the sense that if $B \in \mathcal{B}$ then $\Phi^{-1}_{t,\theta}(B) = \{\mathbf{v} \mid \Phi_\theta(\mathbf{v},t) \in B\}$ is a member of $\mathcal{B}$. This fact will play a crucial role in establishing the following result.

Theorem 1. *Let ψ be a BLTL formula and $t \in \mathcal{T}$. Then the following statements hold.*

1. $\|\psi\|_t \in \mathcal{B}$.
2. $Models(\psi) = \|\psi\|_0$.
3. $Models(\psi) \in \mathcal{B}$.

Proof. To prove the first part of the theorem by structural induction, assume that $\psi = (i,l,u)$ is an atomic proposition. We note that $\{\mathbf{v} \,|\, \ell \leq \mathbf{v}(i) \leq u\} = \prod_{j=1}^{n} (\ell_j, u_j)$ where $\ell_j = L_j$ and $u_j = U_j$ if $j \neq i$ and $\ell_j = \ell$ and $u_j = u$ if $j = i$ and hence $B \in \mathcal{B}$ where for convenience we set $B = \{\mathbf{v} \,|\, \ell \leq \mathbf{v}(i) \leq u\}$. From the definitions it follows that $\mathbf{v}' \in \|(i,\ell,u)\|_t$ iff $\sigma_{(\mathbf{v}',\theta)}, t \models (i,\ell,u)$ iff $\ell \leq \Phi_\theta(\mathbf{v}',t) \leq u$ iff $\Phi_\theta(\mathbf{v}',t) \in B$. This shows that $\|(i,\ell,u)\|_t = \Phi^{-1}_{t,\theta}(B)$, and since $\Phi_\theta(t)$ is measurable, we are assured that $\Phi^{-1}_{t,\theta}(B) \in \mathcal{B}$.

Next we note that if $\|\psi\|_t, \|\psi'\|_t \in \mathcal{B}$ then $\|\neg\psi\|_t \in \mathcal{B}$ and $\|\psi \vee \psi'\|_t \in \mathcal{B}$ since $\mathcal{B}$ is closed under complementation and (countable) union. Similarly, from $\|\psi\|_t, \|\psi'\|_t \in \mathcal{B}$ we can conclude that $\|\psi \mathbf{U}^{\leq k} \psi'\|_t, \|\psi \mathbf{U}^{k} \psi'\|_t \in \mathcal{B}$ since $\mathcal{B}$ is closed under countable intersections as well. The remaining two parts of the result follow from the definitions. □

On-line model checking

Here, we present our *on-line* model checking procedure to check if an ODE trajectory conforms to a property specified using a BLTL formula. On-line model checking combines the process of simulation with model checking i.e we simulate the system only until a decision about the satisfiability of the property can me made. This contrasts with off-line approaches, where the system is simulated for the whole time scale of interest, after which the model checking procedure is applied. At the end of the verification procedure, the model checker returns either a *Yes* or *No* for every ODE trajectory. On-line approaches have the advantage of conserving CPU, memory resources and have a lower amortized time complexity. Specifically, we use a tableau based model checking procedure, which we introduce as follows. The method relies on constructing and propagating a finite family of sets $\mathcal{F}$. Each set $F_i \in \mathcal{F}$ contains a finite number of formulas. Let φ, ψ and γ be BLTL formulas. A literal is defined as an atomic proposition A or its negation $\neg A$. For the purpose of illustration, let us assume that we convert the given BLTL formulas into a form in which only the atomic propositions can appear in negated

form. In other words, we assume our formulas will have the following syntax: (i) Every literal is a formula. (ii) If φ and φ' are formulas so are $\varphi \vee \varphi'$ and $\varphi \wedge \varphi'$, $\mathbf{O}\varphi$, $\mathbf{F}\varphi$, $\mathbf{G}\varphi$, $\varphi \mathbf{U} \varphi'$. It is easy to show that every formula in the original syntax can be expressed as a formula in the above syntax where only the atomic propositions are negated.

For a formula φ, we define the family of closure sets $cl(\varphi)$ by structural induction on φ:

- If φ is a truth constant or a literal then $cl(\varphi) = \{\{\varphi\}\}$.
- If $\varphi = \psi \vee \gamma$ then $cl(\varphi) = cl(\psi) \cup cl(\gamma)$.
- If $\varphi = \psi \wedge \gamma$ then $cl(\varphi) = cl(\psi) \times cl(\gamma)$.
- If $\varphi = \mathbf{O}\psi$ then $cl(\varphi) = \{\{\mathbf{O}\psi\}\}$.
- If $\varphi = \mathbf{F}\psi$ then $cl(\varphi) = cl(\psi) \cup cl(\mathbf{OF}\psi)$.
- If $\varphi = \mathbf{G}\psi$ then $cl(\varphi) = cl(\psi) \times cl(\mathbf{OG}\psi)$.
- If $\varphi = \psi \mathbf{U} \gamma$ then $cl(\varphi) = cl(\gamma) \cup (cl(\psi) \times cl(\mathbf{O}(\psi \mathbf{U} \gamma))$.

If we have a set of formulas $Y = \{\varphi_1, \varphi_2, \ldots, \varphi_n\}$, then the closure $cl(Y)$ can be written as $cl(Y) = cl(\varphi_1) \times cl(\varphi_2) \ldots \times cl(\varphi_n)$. We can also extend the notion of closure to families of sets of formulas such as $\mathcal{F} = \{Y_1, Y_2, \ldots, Y_k\}$, and say that the closure set of $\mathcal{F}$ is $cl(\mathcal{F}) = cl(Y_1) \cup cl(Y_2) \ldots cl(Y_k)$

We call the set of formulas Y a *leaf set* iff $cl(Y) = Y$. Further, a set Y is *inconsistent* iff (i) for an atomic proposition p, $p \in Y$ and $\neg p \in Y$ or (ii) for some formula φ, both $\mathbf{O}\varphi \in Y$ and $\mathbf{O}\neg\varphi \in Y$.

Proposition: The following assertions hold.

- Y is a leaf set iff each formula in Y is a literal or a $\mathbf{O}$ formula.
- $cl(\varphi)$ is a leaf family for each φ.
- $cl(Y)$ is a leaf family for every finite set of formulas Y.
- $cl(\mathcal{F})$ is a leaf family for every family of formula sets $\mathcal{F}$.

Suppose the current system state is s_t. If Y is a leaf set then Y is *dead* at time t iff Y is inconsistent or $s_t \not\models \ell$ for some literal $\ell \in Y$. Consequently, a family of leaf sets $\mathcal{F}$ is dead iff $\forall Y \in \mathcal{F}$: Y is dead. Furthermore, $\mathcal{F}$ is terminal iff $\exists Y \in \mathcal{F} : Y$ is not dead and $next(Y) = \emptyset$, where $next(Y) = \{\psi | \mathbf{O}\psi \in Y\}$.

Now assume we are given a formula φ and want to check in an on-line manner if the system trajectory satisfies φ. We propagate a family of sets and start with $\mathcal{F}^0 = cl(\varphi)$. Inductively, assume that we are given the family of sets $\mathcal{F}^t$ for $t < T$. If $\mathcal{F}^t$ is dead, then we set $\mathcal{F}^{t+1} = false$, and if $\mathcal{F}^t$ is terminal then we set $\mathcal{F}^{t+1} = true$. Otherwise, $\mathcal{F}^t$ is neither dead nor terminal. In this case we know that $\exists Y_1, Y_2, ..., Y_k \in \mathcal{F}^t, k \geq 1$ which are not dead. Since these sets are not dead, we know that $\forall i, 1 \leq i \leq k : next(Y_i) \neq \emptyset$. We can then build the family of sets for time $t+1$ as $\mathcal{F}^{t+1} = cl(next(Y_1)) \cup cl(next(Y_2)) \ldots \cup cl(next(Y_k))$.

The process terminates at time $t < T$ if $\forall Y \in \mathcal{F}^t$ is $false$ and returns $s(0) \not\models \varphi$ or if $\exists Y \in \mathcal{F}^t$ which is $true$, and returns $s(0) \models \varphi$. Furthermore, if $t = T$, if $\mathcal{F}^t$ is a terminal leaf family at $s(T)$, the process terminates and returns that $s(0) \models \varphi$. Otherwise it returns $s(0) \not\models \varphi$.

Statistical model checking properties

The core of our method is verifying properties of model dynamics using statistical model checking. We describe a few such properties along with their BLTL formulas and the result of verification in Table1.

Pathway	Property	Formula	Result
Thrombin-MLC	sustained activation	$(([\text{Phospho MLC} \leq 1]) \wedge (F^{\leq 20}(G^{\leq 20}([\text{Phospho MLC} \geq 3]))))$	*false*
Thrombin-MLC	transient activation	$(([\text{Phospho MLC} \leq 1]) \wedge F^{\leq 20}(([\text{Phospho MLC} \geq 3]) \wedge F^{\leq 20}(G^{\leq 20}((\text{Phospho MLC} \leq 1)))))$	*true*
Segmentation clock	oscillations	$(([\text{Lunatic fringe mRNA} \leq 0.4]) \wedge (F^{\leq 40}([\text{Lunatic fringe mRNA} \geq 2.2] \wedge F^{\leq 40}([\text{Lunatic fringe mRNA} \leq 0.4] \wedge F^{\leq 40}([\text{Lunatic fringe mRNA} \geq 2.2] \wedge F^{\leq 40}([\text{Lunatic fringe mRNA} \leq 0.4])))))))$	*true*

Table 1. Statistical model checking based verification

For the MLC phosphorylation pathway, it is known experimentally that the concentration of phosphorylated MLC starts at a low level, and then reaches a high steady state value. Our SMC method shows that the nominal model does not satisfy the property. However, when specifying a transient profile for phosphorylated MLC, the property is verified a *true*. This discrepancy has been studied in [30], and attributed to missing components and links in the proposed model. Table 1 also shows how an oscillation property for Lunatic fringe mRNA in the segmentation clock pathway can be encoded in our specification logic. The SMC verification shows that the specified oscillation is met.