A Probabilistic Framework for Molecular Network Structure Inference by Means of Mechanistic Modeling

Juho Timonen, Henrik Mannerström, Harri Lähdesmäki, and Jukka Intosalmi

Abstract—Ordinary differential equations (ODEs) provide a powerful formalism to model molecular networks mechanistically. However, inferring the model structure, given a set of time course measurements and a large number of alternative molecular mechanisms, is a challenging and open research question. Existing search heuristics are designed only for finding a single best model configuration and cannot account for the uncertainty in selecting the network components. In this study, we present a novel Markov chain Monte Carlo approach for performing Bayesian model structure inference over ODE models. We formulate a Metropolis algorithm that explores the model space efficiently and is suitable for obtaining probabilistic inferences about the network structure. The method and its special parallelization possibilities are demonstrated using simulated data. Furthermore, we apply the method to a time course RNA sequencing data set to infer the structure of the transiently evolving core regulatory network that steers the T helper 17 (Th17) cell differentiation. Our results are in agreement with the earlier finding that the Th17 lineage-specific differentiation program evolves in three sequential phases. Further, the analysis provides us with probabilistic predictions on the molecular interactions that are active in different phases of Th17 cell differentiation.

Index Terms—Biological network modeling, gene regulatory networks, Markov chain Monte Carlo, ODE models, model selection, Th17 cell differentiation

1 Introduction

In the field of computational biology, mechanistic models for biochemical networks are often constructed in the form of nonlinear ordinary differential equation (ODE) systems. Kinetic parameters of these mechanistic ODE models can be calibrated using statistical techniques if sufficient amount of time course data are available (see e.g., [1], [2], [3], [4]). Whereas the parameter estimation problem for ODE models is rather well studied, the related model structure inference problem remains unsolved in practice.

When mechanistic models are constructed, there is typically a large number of well-motivated, hypothetical models for the underlying biochemical mechanisms. A central challenge is to determine the most likely model structures given the experimental data. The number of alternative models can span from two to hundreds of thousands, and therefore the need for reliable, automatized, and computationally efficient methods that can reveal the underlying mechanisms is urgent.

Under the Bayesian formalism, different model structures can be compared by evaluating their posterior probabilities [5]. However, this is computationally demanding for nonlinear ODE models, because it involves computing the model marginal likelihoods by integrating over the model parameters [3]. Even if some approximation for the marginal likelihood is used [6], in practical applications it is not usually feasible to exhaustively compare all the alternative network combinations.

The combinatorial complexity of the network inference problem is often tackled by searching the optimal model using some greedy heuristic such as forward or backward search or their variants (see e.g., [7], [8], [9], [10]). The heuristic search methods are designed to find a single model which is the best according to some criterion. In the context of Bayesian model inference, the search methods can be used to find the maximum a posteriori (MAP) point estimate for the model structure. On the other hand, if we wish to obtain an approximation for the full model posterior, the simple search techniques are infeasible and we need to build an algorithm which explores the model space more thoroughly.

In this study, we present a novel approach to efficiently approximate the posterior distribution over alternative ODE model structures. The main idea is to use Markov Chain Monte Carlo (MCMC) methods for exploring the discrete model space. This is efficient because an MCMC chain is attracted towards the high-probability regions of the model space, and only the models that are encountered by the chain have to be evaluated. Furthermore, the uncertainty related to the structure inference can be accounted for by approximating the discrete model posterior distribution so that the models that have not been evaluated are given

• *The authors are with the Department of Computer Science, Aalto University, Aalto, FI-00076, Finland. E-mail: {juho.timonen, henrik.mannerstrom, harri.lahdesmaki, jukka.intosalmi}@aalto.fi.*

Manuscript received 27 June 2017; revised 14 Feb. 2018; accepted 2 Apr. 2018. Date of publication 10 Apr. 2018; date of current version 5 Dec. 2019. (Corresponding author: Juho Timonen.)
Digital Object Identifier no. 10.1109/TCBB.2018.2825327

zero probability. The proposed method therefore gives us an approximation of the entire model posterior distribution, and therefore provides much richer information than the MAP solution which can be obtained using a search heuristic such as the forward or backward search.

We provide an implementation of a discrete Metropolis algorithm with a simple yet efficient proposal distribution. The performance of our method is demonstrated using toy problems in which the data is artificial and the set of possible models is restricted so that we are able to compare the model posterior approximations with the full posterior. Using these test cases, we show that our MCMC-based approach is capable of producing accurate probabilistic inferences about the true molecular interactions even when only a fraction of all possible models are evaluated. Moreover, we argue that the proposed method has beneficial parallelization opportunities, because computation time required for reliably exploring the interesting parts of the model posterior can be reduced by combining information from multiple independent MCMC chains. This is also demonstrated in the simulated data experiments.

For a real world application, the proposed method is used to infer the transiently evolving core molecular network that steers the T helper 17 (Th17) cell differentiation. To capture the rewiring effects during the differentiation process, we utilize the recently developed latent effect mechanistic (LEM) modeling approach [9]. In this study, we also extend the LEM modeling approach to allow rigorous statistical testing about the type of the latent process. In our Th17 cell differentiation application, this extension enables us to test hypotheses about how many sequential phases are involved in the differentiation processes. Further, our novel approach also enables us to generate probabilistic predictions on the molecular interactions that are active in different phases of Th17 cell differentiation instead of merely finding a single molecular network structure with the highest posterior probability.

2 Materials and Methods

2.1 Constructing ODE Models for Biochemical Systems

Mechanistic ODE models can in principle be constructed for any biochemical system which consists of molecular species that interact through chemical reactions. The functional form of the ODE system can be determined by using, for instance, the law of mass action, Hill kinetics, or Michaelis–Menten kinetics (for an introduction on building mathematical models in systems biology, see e.g., [11]). In general form, the model construction process results in an ODE system

$$\frac{\mathrm{d}y}{\mathrm{d}t} = g(y,\theta), \tag{1}$$

where $y(t,\theta) : [0,T] \times \mathbb{R}^d \rightarrow \mathbb{R}^n$, $g(y,\theta) : \mathbb{R}^n \times \mathbb{R}^d \rightarrow \mathbb{R}^n$, and $\theta \in \mathbb{R}^d$ is a parameter vector. Here d is the dimension of the parameter space and n is the number of model components (chemical species). The interactions between model components can be of many types, for instance, activating or inhibiting. The type and number of chemical interactions between different chemical species determine the biochemical network structure that steers the time evolution of the system.

2.2 Latent Effect Mechanistic Models for Transiently Evolving Biochemical Systems

In many applications, the biochemical reaction system may evolve transiently in time (see e.g., [12]). As a simple example, a set of fixed biochemical interactions may be affected by fluctuations in the system temperature. Alternatively, in a more complex setting, the reactions that drive the gene regulatory interactions may be affected by transient epigenetic or signaling mechanisms that are difficult or impossible to incorporate into the standard ODE model due to the lack of detailed information about the hidden mechanisms. In such cases, the standard ODE modeling approach is insufficient and a more powerful modeling formalism is needed.

In a recent study [9], Intosalmi et al. introduced the so-called latent effect mechanistic models which facilitate the study of dynamically evolving gene regulatory networks. LEM modeling provides a powerful and rigorous means to model transiently evolving systems, but due to the large model space, the LEM model structure inference problem is even more challenging than for standard ODE models. The approach that we propose for model structure inference in this study is specifically designed for problems with a large model space and, thus, LEM modeling provides us with an excellent model class to test our approach.

A detailed formal definition of the LEM model, as well as illustrative examples, can be found in [9]. However, we present the central steps in constructing LEM models in the following. To formulate a LEM model with M different latent states, we need to define a set of N mechanistically motivated functions $f_j(y,\theta)$ that each correspond to a mechanism involving one or more of the components y_i. In addition, we need to design an M-dimensional latent process $x(t,\theta)$ that defines how the strength of each mechanism evolves over time. The exact functional form of this process needs to be designed based on the system properties. Given a latent process $x(t,\theta)$, the LEM model is defined by the equations

$$\frac{\mathrm{d}y_i}{\mathrm{d}t} = \sum_{j \in \mathcal{I}_i} f_j(y,\theta) w_j(t,x,Z,\theta), \tag{2}$$

where for each $i = 1,\ldots,n$, $\mathcal{I}_i$ is the set of indices $j \in \{1,\ldots,N\}$, for which the function $f_j(y,\theta)$ is affecting y_i. The matrix Z is an $N \times M$ binary matrix where the element $\{Z\}_{jk}$ determines if function f_j is present in the kth latent state or not (values 1 and 0, respectively). This state-dependent behavior of the functions f_j is mediated to the final ODE system through the weight function

$$w_j(t,x,Z,\theta) = \frac{\sum_{k=1}^{M} \{Z\}_{jk} x_k(t,\theta)}{\sum_{k=1}^{M} x_k(t,\theta)}. \tag{3}$$

Intuitively, the LEM model consists of a standard ODE model which is coupled with a latent process that determines how different subnetworks contribute to the dynamics over time. In the context of the LEM model formulation, the binary matrix Z determines the structures of the subnetworks. The standard ODE model also can be constructed using the same formalism and considering the latent process as a single component constant function. Thus, the structure of a standard ODE model is determined by a $N \times 1$ matrix (vector) Z. The correspondence between the network structures and Z

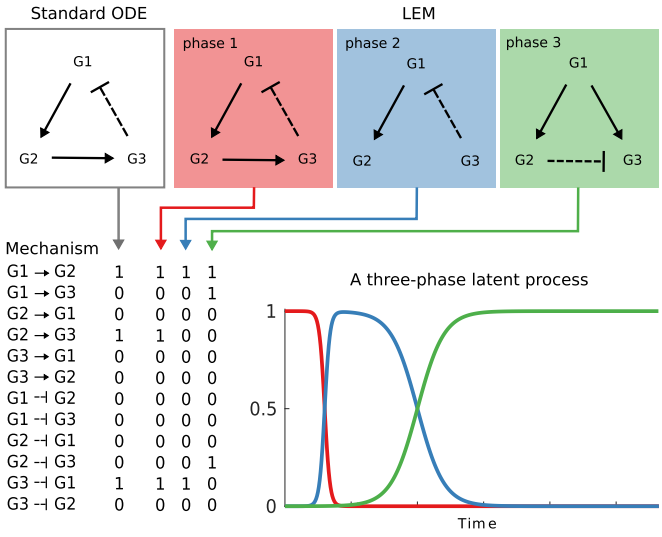

Fig. 1. Illustration of a standard ODE and a three-phase LEM model. Each row of the binary matrix corresponds to one mechanism and each column to one latent state. Here, the solid arrows represent activation links and the dashed lines are inhibition links. The element on row i and column j defines if mechanism i is present in phase j. An example design of the time-dependent latent process is illustrated with the red, blue, and green lines that represent the three latent states as a function of time. Each latent component represents a different, possibly overlapping phase of cellular process.

matrices in standard ODE models and LEM models is illustrated along with an example latent process design in Fig. 1. Because the standard ODE model is a special case of the LEM model, we will use Z to denote the model structure throughout this paper.

2.3 Statistical Inference of Rate Parameters and Model Structure

The model class that we introduced above provides us with powerful means to describe the dynamics of different kinds of biochemical systems. However, the rate parameters, possible latent process parameters, and also the model structure Z are typically unknown and they need to be calibrated from time course data using statistical techniques. To carry out this calibration task, we formulate the problem within the Bayesian framework (for an introduction to Bayesian inference for ODE models, see e.g., [3]).

Let us assume that we have a data set $D = \{D_{ik} \mid i = 1, \ldots, n; k = 1, \ldots, K_i\}$ where D_{ik} corresponds to a noisy measurement of the ODE model output $y_{ik}(Z, \theta_Z) = y_i(t_{ik}, Z, \theta_Z)$, i.e., the value of the ith component at time t_{ik}, given that the model is Z and has parameters θ_Z. For notational convenience, we will from now on drop the subscript from θ_Z as it is always clear that the parameters in question are those of model Z. Assuming that the data points are independent, the model response y can be combined with the time course data through the likelihood function

$$p(D \mid \theta, Z) = \prod_{i=1}^{n} \prod_{k=1}^{K_i} h(D_{ik} \mid \theta, Z), \tag{4}$$

where sampling distribution h should be selected based on the properties of measurements used in an application. By applying the Bayes' theorem, we can obtain the parameter posterior distribution

$$p(\theta \mid D, Z) = \frac{p(D \mid \theta, Z) p(\theta \mid Z)}{p(D \mid Z)}, \tag{5}$$

where $p(\theta \mid Z)$ is the prior distribution over the model parameters and

$$p(D \mid Z) = \int p(D \mid Z, \theta) p(\theta \mid Z) d\theta, \tag{6}$$

is the marginal likelihood. Further, we can express the posterior distribution over models Z in the form

$$p(Z \mid D) \propto p(D \mid Z) p(Z), \tag{7}$$

where $p(Z)$ is the prior distribution over all possible configurations of the matrix Z.

2.4 Testing Hypotheses about the Latent Process of LEM Model

Applications of LEM modeling can involve several alternative hypotheses about the possible latent process. For instance, in the gene regulatory network application that we consider in this study, the latent process can be constructed to have one, two, or three sequential phases. Bayesian methodology makes it possible to compute the posterior probabilities for the hypotheses about the latent process and thus obtain quantitative ranking for the hypotheses. In the following, this is derived formally by generalizing the notation of the previous section to allow conditioning on hypotheses about the latent process.

If we condition the whole model with the latent process hypothesis H, the parameter posterior takes the form $p(\theta \mid D, Z, H) \propto p(D \mid Z, H, \theta) p(\theta \mid Z, H)$ and the marginal likelihood $p(D \mid Z, H)$ can be computed by integrating the rate parameters out in a similar manner as in Eq. (6). The posterior distribution over the alternative models can also be expressed conditioned with H, i.e., $p(Z \mid D, H) \propto p(D \mid Z, H) p(Z \mid H)$. The posterior distribution over different latent process hypotheses is then

$$p(H \mid D) = \frac{p(D \mid H) p(H)}{\sum_H p(D \mid H) p(H)}, \tag{8}$$

where $p(H)$ is the prior distribution over alternative hypotheses and

$$p(D \mid H) = \sum_{Z \in \mathcal{M}_H} p(D \mid Z, H) p(Z \mid H), \tag{9}$$

where $\mathcal{M}_H$ denotes the set of alternative models consistent with the hypothesis H. Eq. (9) provides a well-defined posterior probability measure that can be used to evaluate the hypothesis H.

2.5 Probabilistic Inference of the ODE Model Mechanisms

If the full posterior distribution over alternative model configurations is available, we can determine how probable the individual mechanisms of the network are. For this purpose we can compute the so-called posterior weights that are presented in the following.

Let us assume that we have a model posterior distribution $p : \mathcal{M} \to [0, 1]$, and we wish to obtain predictions about the model structure. Now the matrix

$$W_p = \frac{\sum_{Z \in \mathcal{M}} p(Z) Z}{\sum_{Z \in \mathcal{M}} p(Z)}, \tag{10}$$

contains the posterior-weighted averages for each element in the model configuration matrix over all possible models. We can interpret the element (i, j) of this matrix as the probability that describes how likely it is that the corresponding molecular interaction i is present in the jth phase of the rewiring sequence. In the case of standard ODE models, W_p is just a vector that contains the posterior weights for each link in the full network.

In real world applications, the posterior weights have to be computed by using an approximation for p, since it often is not feasible to compute the posterior probability for all viable models. The following sections focus on how to obtain a good approximation efficiently. This approximative distribution can then also be used to perform the hypothesis testing presented in the previous section.

2.6 Marginal Likelihood Estimation

The estimation of the marginal likelihood is challenging in general and there exist several approaches to carry out the estimation task [6]. Throughout this study, we compute all marginal likelihoods using the approximation

$$\log p(D \mid Z, H) \approx \log p(D \mid Z, H, \theta_{\mathrm{ML}}) + \frac{d}{2} \log (T), \tag{11}$$

where θ_{ML} is the maximum likelihood estimate, d is the parameter dimension and T the total number of data points [13]. This crude approximation is essentially the same as the approximation behind the Bayesian information criterion [14]. The only computational task that is required for evaluating Eq. (11) is computing the maximum likelihood estimate, and consequently, it provides us with a computationally efficient estimate which can be computed for a large number of models in a reasonable time.

2.7 Exploring the Model Space Using Markov Chain Monte Carlo Methods

Even though estimation of the marginal likelihood via Eq. (11) is computationally rather efficient, the evaluation of this expression can be carried out for all possible configurations of Z only in the case of simple toy systems with a limited number of possible models. On the other hand, if the data are informative, only a small fraction of alternative configurations of matrix Z are likely to fit the data well. Embracing Markov chain Monte Carlo methods allows us to search the discrete model space cleverly without having to wade through the multitude of unsatisfactory models. The idea is to create an MCMC sampler which efficiently samples the model posterior distribution and is attracted towards the high probability models. Moving between already visited or proposed states is computationally light and can be implemented using the lookup method, since the parameters of any model have to be optimized only once. This means that time-consuming computations are needed only when a previously unseen model is being proposed. Furthermore, the MCMC sampler will explore the high-probability regions of the model space more thoroughly than any greedy search method, and therefore finds many more high-probability models. This is essential for obtaining information about the link uncertainties, since the model posterior can often be very vague in applications.

It is well known that the choice of proposal distribution is vital for MCMC methods that have desirable properties, such as rapid convergence or good mixing of the chains [15]. With a discrete model space, we wish to define a sensible neighborhood relation between the different models so that building an efficient proposal distribution is possible. Let us define a model space $\mathcal{M}$ which is a set of all $N \times M$ binary matrices. For model configuration $Z \in \mathcal{M}$, we define the k-neighborhood of Z as

$$\mathcal{N}_k(Z) = \{Y \in \mathcal{M} : \ 1 \leq ||Y - Z||_F^2 \leq k\}, \tag{12}$$

where $||A||_F$ is the Frobenius norm of matrix A. With binary matrices Y and Z, $||Y - Z||_F^2$ is the number of differing elements of the matrices. Now, if $Y \in \mathcal{N}_k(Z)$ with $k \ll N \times M$, then the models Z and Y have a high degree of similarity and many common components. Consequently, they are likely to have rather comparable posterior probabilities. Thus, a random walk Monte Carlo algorithm with a proposal distribution that has most weight on models in k-neighborhoods of the current state with small k presumably allows reasonably smooth moving with a practical acceptance rate.

2.8 A Random Walk Metropolis Algorithm for Model Posterior Sampling

We now formulate a Metropolis algorithm [16] for exploring a discrete model space consisting of binary configuration matrices, that can represent different ODE or LEM model structures. For a model Z, we have $p(Z \mid D, H) \propto p(D \mid Z, H) p(Z \mid H)$, where $p(Z \mid H)$ is the prior distribution of models, given the hypotheses H. In the following, we denote $\pi_H(Z) = p(D \mid Z, H) p(Z \mid H)$. We use a proposal distribution that is uniform over the 1-neighbors of the current model, i.e.,

$$Q(Z' \mid Z) = \begin{cases} \frac{1}{NM}, & \text{if } Z' \in \mathcal{N}_1(Z) \\ 0, & \text{otherwise,} \end{cases} \tag{13}$$

which clearly is symmetrical. When the models space consists of only standard ODE models, this proposal distribution is just uniform over removing or adding one link in the network. If the current state is Z, one Metropolis step consists of the following parts:

1) Draw a proposal model Z' from the discrete proposal distribution $Q(\cdot \mid Z)$.
2) Accept transition to Z' with probability

$$A(Z' \mid Z) = \min\left\{1, \frac{\pi_H(Z')}{\pi_H(Z)}\right\}, \tag{14}$$

or reject it with the probability $1 - A(Z' \mid Z)$.

Now the transition probability from state Z to Z' is the product $T(Z' \mid Z) = Q(Z' \mid Z) A(Z' \mid Z)$. For any distinct models Z and Z' such that $Z' \in \mathcal{N}_1(Z)$, we have

$$\pi_H(Z) T(Z' \mid Z) = \min\left\{\frac{\pi_H(Z)}{NM}, \frac{\pi_H(Z')}{NM}\right\} \tag{15}$$

$$= \pi_H(Z') T(Z \mid Z'), \tag{16}$$

which is the detailed balance condition. Because the detailed balance is satisfied and the proposal distribution Q is irreducible, it follows that this sampling algorithm produces an ergodic Markov chain, and thus its invariant distribution is $\pi_H(Z)$ [17]. Because $\pi_H(Z)$ is proportional to the model posterior, the algorithm will give unnormalized posterior probabilities for all the models that it has encountered at least once.

2.9 Model Posterior Distribution Approximation

Assume that we have first defined a model space $\mathcal{M}$ consisting of well-motivated ODE models that are viable under all hypotheses and prior information concerning the model components and parameters. After measuring time course data of the ODE model components, our goal is to obtain information about the actual underlying model structure. When the full model posterior $p(Z \mid D, H)$ is not computationally solvable, we begin the task by approximating this discrete probability distribution with a distribution that has a relatively small support. Of course, for this approximation to be good, the support must contain a remarkable proportion of the high posterior probability models.

Assume that we have run any search algorithm that has explored the set $\mathcal{Z} \subseteq \mathcal{M}$ and therefore provided us with the (possibly approximated) marginal likelihood values $p(D \mid Z, H)$ for each $Z \in \mathcal{Z}$. The model posterior approximation is then the distribution $q : \mathcal{M} \to [0, 1]$, where

$$q(Z) = \frac{p(D \mid Z, H)p(Z \mid H)}{\sum_{Y \in \mathcal{Z}} p(D \mid Y, H)p(Y \mid H)}, \tag{17}$$

for all $Z \in \mathcal{Z}$ and $q(Z) = 0$ for all $Z \in \mathcal{M} \backslash \mathcal{Z}$.

2.10 Assessing the Goodness of the Model Posterior Approximation Using Independent Chains

In our approach, we use the Metropolis algorithm to efficiently find a subset of models that is sufficient for obtaining a good model posterior approximation. In order to assess the reliability of the obtained results and to decide when to terminate the search, we use multiple independent MCMC chains. Naturally, the model posterior approximation is slightly different for different runs and, in the worst case if the independent MCMC chains have not reached the same high probability regions, the difference may be notable. Note that as the sampled distribution is discrete, we do not need to estimate the model posterior probabilities based on how frequently they occur in the MCMC chain, because the actual posterior probabilities (or their approximations) are available for each of the encountered model structures via Eq. (11). Consequently, we should not use the standard convergence diagnostics but rather need some measure to compare the approximative distributions that are obtained from independent chains using Eq. (17) and this way also assess the convergence of the approximations. For this purpose, we use the overlapping (OVL) coefficient [18], [19].

The OVL for two different posterior approximations $p(Z)$ and $q(Z)$ can be computed simply via the formula

$$\mathrm{OVL}(p, q) = \sum_{Z \in \mathcal{M}} \min\{p(Z), q(Z)\}. \tag{18}$$

Intuitively, the OVL coefficient measures the similarity of two distributions by computing the area that they share and the measure is bounded between 0 and 1, so that $\mathrm{OVL}(p, q) = 1$ if and only if $p = q$.

2.11 Computational Implementation

The Metropolis algorithm, and the marginal likelihood approximation in Eq. (11) are implemented in MATLAB (The MathWorks Inc., Natick, MA, USA). Even though the marginal likelihood approximation in Eq. (11) is computationally relatively simple compared to other methods, the related maximum likelihood parameter optimization problem is often difficult for nonlinear ODE models due to multimodality of the likelihood surface [3]. Following the recommendations presented in [1] and [2], we tackle the parameter optimization by using multiple local optimization starts with a latin hypercube setting. In order to achieve better performance, these starts are done in parallel. The CVODES solver from the SUNDIALS package [20] is used to solve the ODE systems numerically. This involves solving the related sensitivity equations simultaneously with the actual ODE system [21]. A more detailed description of the used methods and their reasoning can be found in the supplemental material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2018.2825327. The implementation is available at version.aalto.fi/gitlab/timonej3/ode_mcmc.

2.12 Exact Mathematical Formulation of the Dynamically Evolving Molecular Interactions in Our Experiments

In the present study, we will demonstrate the model structure inference with both simulated toy examples, and real data. In all our experiments, the construction of the ODE system that mathematically describes a given LEM model (Eq. (2)) is begun by creating a standard ODE system

$$\frac{dy_i}{dt} = k_i^{\mathrm{b}} + \sum_{j=1}^{n} k_{ij}^{\mathrm{a}} y_j + \sum_{\substack{k,j=1 \\ j>k}}^{n} k_j^{\mathrm{sa}} y_j y_k - \sum_{j=1}^{n} k_{ij}^{\mathrm{inh}} y_i y_j - k_i^{\mathrm{d}} y_i, \tag{19}$$

where k_i^{b}, k_{ij}^{a}, k_{ijk}^{sa}, k_{ij}^{inh} and k_i^{d} are the corresponding basal activation, independent activation, synergistic activation, inhibition and degradation rate constants, respectively. The summed components of Eq. (19) correspond to the functions f_j in Eq. (2). To include the time-dependent behavior of the mechanisms, we couple the ODE system with a latent process $x : [0, T] \to \Delta$, where $\Delta = \{x(t) \in [0, 1]^3 : \sum_i x_i(t) = 1\}$ using Eq. (3). In our design, $x(t)$ has the form

$$\begin{cases} x_1(t) = 1 - \frac{1}{1+\exp(-\lambda_1(t-\tau_1))} \\ x_2(t) = \frac{1}{1+\exp(-\lambda_1(t-\tau_1))} - \frac{1}{1+\exp(-\lambda_2(t-\tau_1-\tau_2))}, \\ x_3(t) = \frac{1}{1+\exp(-\lambda_2(t-\tau_1-\tau_2))} \end{cases} \tag{20}$$

where τ_1, τ_2 are location parameters and λ_1, λ_2 are shape parameters of the process. This formulation can be used for models that have up to three latent states. For models with two latent states, the parameters λ_2, τ_2 get suppressed from the final ODE system, and for models with only one latent

state the whole coupling weight function in Eq. (3) is constant at 1.

2.13 Experimental Data

To test the performance of our method in a real application, we make use of the extensive RNA sequencing data set provided by [22]. This data set provides rich information about the dynamic transcriptional changes during the T helper 17 differentiation process. These data are particularly good for testing model structure inference in the context of LEM modeling because Th17 cell differential process proceeds in sequential phases which can be modelled as latent states [9], [12]. The detailed experimental procedures that have been used to generate the RNA-seq data set are given in the original article [22]. In brief, the data are obtained by purifying naive $CD4^+$ cells from lymph nodes and spleen of wild-type mice, culturing the cells in Th17 conditions, and harvesting a proportion of cells at the time-points 0, 1, 3, 6, 9, 12, 16, 24, and 48 hours for RNA-seq processing. In this study, we use the fragments per kilobase of transcript per million mapped reads (FPKM) values of each core regulatory gene.

3 Results

3.1 Experiments with Simulated Data

In order to test the performance of our method, we formulate ODE model structure inference problems that involve realistically simulated data. In these experiments, the amount of different possible models is small enough to permit exhaustive computation of the BIC (Eq. (11)) for each model for reference. Running the Metropolis algorithm after obtaining the full model posterior distribution will not then increase the computational burden anymore, and we can easily investigate how the inferred model structure evolves as we run the algorithm. In particular, we test how the overlapping coefficient given in Eq. (18) between the real model posterior and the obtained approximation as well as the inferred posterior weights in Eq. (10) behave as a function of the approximation support size. Throughout the rest of the paper, we often refer to the size of the approximation support simply as the number of (evaluated) models. This number is the amount of models that have been proposed and possibly accepted by the Metropolis sampler, and reflects the computational burden that effectively increases only when a previously unseen model is proposed.

We present three different examples, one considering only standard ODE models and the other two involving LEM models with one to three phases. In the LEM experiments, we also demonstrate testing hypotheses about the latent process. Readers who are only interested in ODE model structure inference in general can study only Experiment 1, whereas readers who also wish to understand inferring the number of latent phases in the Th17 cell application are encouraged to see also Experiments 2 and 3.

Each experiment involves noisy data that is simulated from a mechanistically constructed ODE model. To be more exact, the data points D_{it} are generated so that $D_{it} = y_{it} + \varepsilon(y_{it}, \alpha, \beta)$, where y_{it} is the output of the data generating ODE system for component i at time t. The added measurement noise has a normal distribution with heteroscedastic variance that is proportional to the output, i.e.,

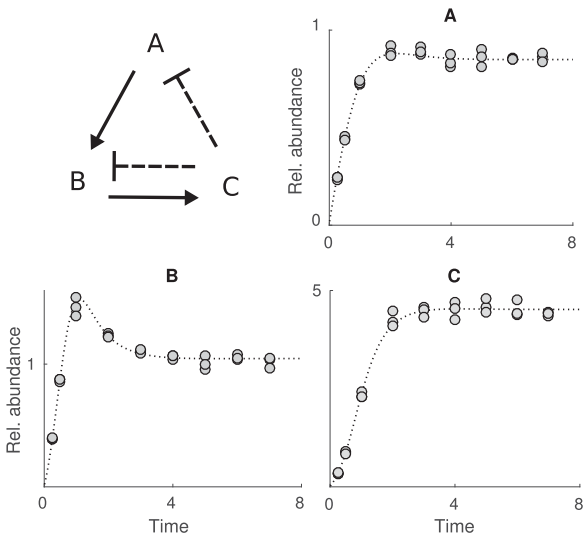

Fig. 2. Regulatory mechanisms of the data generating model and the simulated data in Experiment 1. The solid arrows and dashed turnstiles represent activating and inhibiting links, respectively. The dotted line represents the underlying model response and the noisy measurements are plotted using gray circles.

$$\varepsilon(y_{it}, \alpha, \beta) \sim \mathcal{N}(0, (\alpha + \beta y_{it})^2). \tag{21}$$

In each experiment, we use fixed noise parameters $\alpha = 10^{-4}$ and $\beta = 0.035$ and assume them known. Also, all experiments involve three genes and three replicates of each measurement. Thus, the likelihood in Eq. (4) takes the form

$$p(D \mid \theta, Z) = \prod_{i=1}^{3} \prod_{k=1}^{3\times K} \mathcal{N}(D_{ik} \mid y_{ik}, (\alpha + \beta y_{it})^2), \tag{22}$$

where K is the number of distinct measurement time points and $y_{it} = y_i(t, Z, \theta)$.

In these experiments, the ODE system for a given model is first constructed from its mechanisms according to Eq. (19). If the model consists of several latent states, we use the latent process design specified by Eq. (20). In each experiment, we apply a uniform prior distribution over models, meaning that $p(Z \mid D, H) \propto p(D \mid Z, H)$. Furthermore, we use fixed initial values $y_i(0) = 0.01$ for each $i = 1, 2, 3$. All experiments also involve some mechanisms that are fixed, i.e., appear in all feasible models. We note that the parameters of these fixed mechanisms are nevertheless estimated from the data.

3.1.1 Experiment 1

In the first example, we consider standard ODE models with three genes. We denote the genes by A, B and C and for each gene, we simulate three replicates of measurements at time points 0.25, 0.5, 1, 2, 3, 4, 5, 6 and 7. We consider only models where each gene is affected by basal activation and degradation. These mechanisms are thus included in the ODE system of each model. Further, we allow the genes to interact through all possible activating and inhibiting interactions. The model space can then be expressed as the union of all 12×1 binary matrices, where each row corresponds to one activation or inhibition link. This means that there are 2^{12} different models. The activation and inhibition links of the data generating model as well as the generated data are shown in Fig. 2. The corresponding kinetic rate

TABLE 1
Kinetic Rate Parameter Values of the Data Generating Models in Each Simulated Data Experiment

Mechanism	Exp. 1	Exp. 2	Exp. 3
basal activation of A	1	1.5	1
basal activation of B	1	-	-
basal activation of C	1	-	-
activation $A \rightarrow B$	5	3	3
activation $A \rightarrow C$	-	-	0.5
activation $B \rightarrow C$	2.5	2	-
activation $C \rightarrow B$	-	1.5	1.5
inhibition $C \dashv A$	0.15	-	-
inhibition $C \dashv B$	1	-	-
degradation of A	0.5	0.5	0.3
degradation of B	0.5	2	1
degradation of C	0.8	3	1.5

parameter values used in the data simulation are shown in Table 1. When performing model parameter calibration, our allowed range for the parameters is set to $[0.001, 30]$.

Fig. 3 displays the overlapping coefficient between the posterior approximation and the full posterior as a function of the number of evaluated models. Here the Metropolis algorithm is started from the empty model, i.e., from a 12×1 matrix of zeros, and it was run until the overlap reaches 0.99. This happened, when 339 different models where encountered. This means that the algorithm efficiently finds the models with high posterior probability and only a fraction of models is needed to obtain a good posterior approximation. Fig. 3 exhibits the posterior weights of each link computed using the approximations obtained after 80 and 100 evaluated models. The weights computed using the full posterior distribution are also plotted for reference. We see that after 100 models, when $OVL \approx 0.8$, the link weights are inferred very accurately. Another thing to note is that even though the OVL is only 0.35 after 80 models, we have inferred most of the posterior weights rather accurately.

This experiment demonstrates that for a standard ODE model structure inference problem, the algorithm indeed can provide a good model posterior approximation with a computational effort that is only a fraction from the effort required to compute the full model posterior distribution. It is clear that if this approximation is good, then also the structure inference is reliable. However, the results show that it is possible to obtain quite accurate predictions about the model structure even when the obtained approximation only partly overlaps with the full model posterior. Furthermore, the inference gives a posterior weight close to one for those mechanisms that actually were in the data generating model, when the data is informative enough.

3.1.2 Experiments 2 and 3

In the second and third experiment, we consider LEM models that have from one to three latent states. We generate three replicates of measurements at time points $1, 2, 3, 5, 8, 12, 18, 24$ and compare hypotheses H_M: "There are M latent states." for each $M = 1, 2, 3$. We consider three genes A, B, and C, and assume that basal activation of A and degradation of each gene are fixed mechanisms. The model space under the hypothesis H_M contains all possible combinations of the four links shown in Fig. 4 in M states. As explained in detail

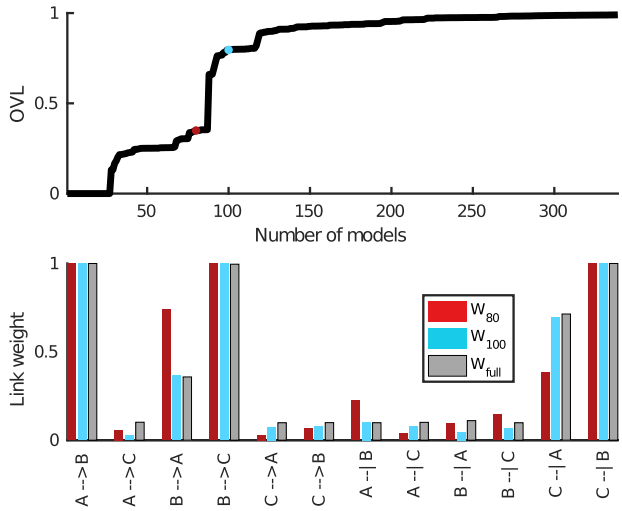

Fig. 3. Visualization of the model structure inference in Experiment 1. The top panel shows the value of the overlapping coefficient (OVL) between the full model posterior distribution and the approximative distribution as a function of the number of evaluated models. After 80 models, we have $\mathrm{OVL} \approx 0.35$ and after 100 models $\mathrm{OVL} \approx 0.80$, which is highlighted by the red and blue dots. The red and blue bars in the bottom panel represent the inferred link weights at these points. The gray bars represent the weights W_{full} which were computed using the entire model posterior. The total number of alternative models is 4,096.

in Section 2.2, the structure of a LEM model of this kind can be expressed using a $4 \times M$ binary matrix and, in total, the model space consists of 2^{4M} different models. In Experiment 2, the data are generated from a two-phase model and, in Experiment 3, from a three-phase model. Illustrations of these models and data are shown in Figs. 5 and 6, respectively. The used kinetic rate parameter values are shown in Table 1 and the used latent process parameters in Table 2.

For kinetic rate parameters, we use the range $[0.001, 50]$ in Experiment 2 and $[0.001, 30]$ in Experiment 3. In both experiments, the allowed range for λ_1 and λ_2 is $[0.75, 3]$. If the model has three phases, then the allowed range is $[1, 12]$ for τ_1 and $[3, 12]$ for τ_2. For two-phase models, we use the range $[1, 24]$ for τ_1. These parameter ranges allow flexible dynamics of the latent processes, but yet are restricted so that very rapid changes that would demand very dense data for reliable calibration are not possible. We note that this formulation with extreme parameters allows negative values for the second latent component of a three-phase model, but within our bounds these values are negligibly small. An example of a three-phase latent process obtained with similar formulation is illustrated in Fig. 1.

The structure inference is performed separately for each hypothesis H_M, $(M = 1, 2, 3)$, by starting one Metropolis chain in the corresponding model space from the $4 \times M$ matrix of zeros. Fig. 7 shows the overlap of the real model

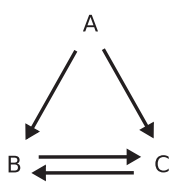

Fig. 4. Possible activation mechanisms in Experiments 2 and 3. In addition to these interactions, it is assumed that gene A has basal activation and all genes are allowed to degrade at some unknown rate.

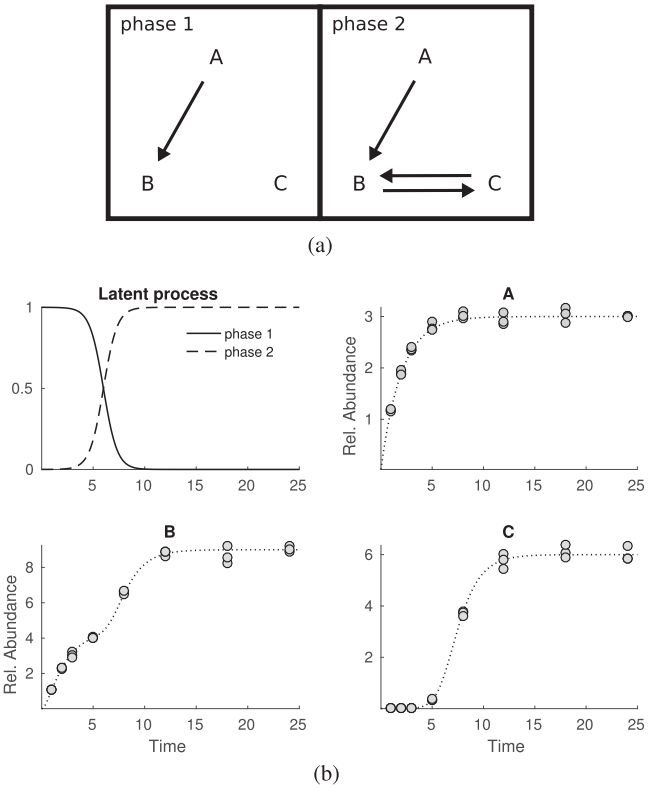

Fig. 5. Illustration of the activation mechanisms of the two-phase data generating model and the simulated data used in Experiment 2. (a) Activation links in each of the two latent states. (b) The latent process, model output, and noisy data plotted similarly as explained in Fig. 2.

TABLE 2
Latent Process Parameter Values of the Data Generating Models in the Simulated LEM Experiments

Interpretation	param.	Exp. 2	Exp. 3
rate of first state transition	λ_1	1.5	2.5
time of first state transition	τ_1	6	3
rate of second state transition	λ_2	-	1
time between state transitions	τ_2	-	5

posterior and the approximation obtained after different amounts of evaluated models for each $M = 1, 2, 3$. Again, we notice that the approximation is effectively same as the real model posterior after only a relatively small number of evaluated models. This is most evident in the case $M = 3$, when the model space is largest.

We assume a uniform prior distribution over the latent process hypotheses and compute the posterior probabilities $p(H_M \mid D)$ for each $M = 1, 2, 3$ given by Eqs. (8) and (9) to carry out the hypothesis comparison. The values for both experiments are computed from the full posterior distributions and are shown in Table 3. The values indicate that the hypothesis involving the real number of phases in the data generating model gets virtually all of the posterior probability mass in the hypothesis testing and we can infer the correct form of the latent process in both cases.

The successful exploration of the model space and the consequent rapid finding of the high probability models strengthen our belief in the power of the algorithm. Here, we computed the values $p(D \mid H_M)$ using the full model

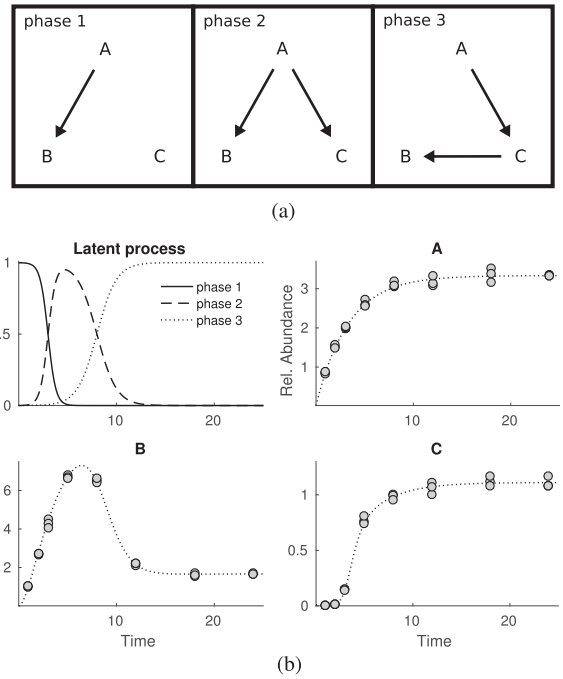

Fig. 6. Illustration of the activation mechanisms of the three-phase data generating model and the simulated data used in Experiment 3. (a) Activation links in each of the three latent states. (b) The latent process, model output, and noisy data plotted similarly as explained in Fig. 2.

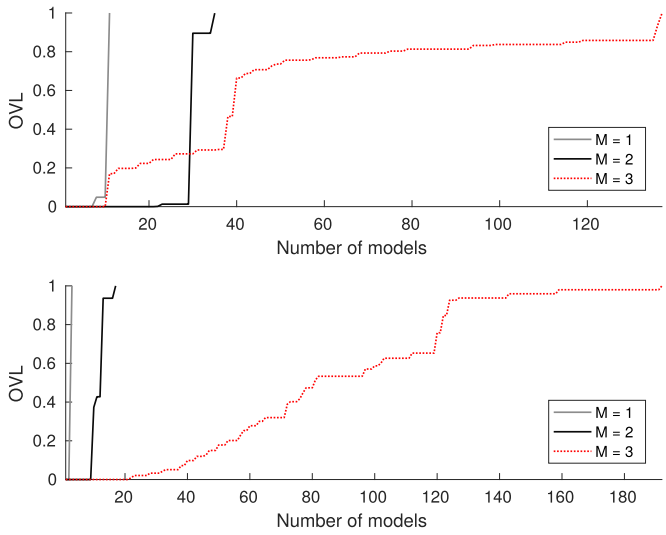

Fig. 7. Demonstration of the model posterior approximation in Experiments 2 (top panel) and 3 (bottom panel). The lines indicate the value of the overlapping coefficient between the full model posterior and the approximation obtained at different amounts of explored models for $M = 1, 2, 3$. The chains were stopped when the overlap reaches 0.99.

TABLE 3
Posterior Probabilities for Different Numbers of Latent States in the LEM Experiments

	$p(H_1 \mid D)$	$p(H_2 \mid D)$	$p(H_3 \mid D)$
Exp. 2	0	0.9955	0.0045
Exp. 3	0	$9.8 \cdot 10^{-223}$	1.0000

posterior, but since our approximations are good, it is clear that the same result is obtained without having to compute the full posterior exhaustively.

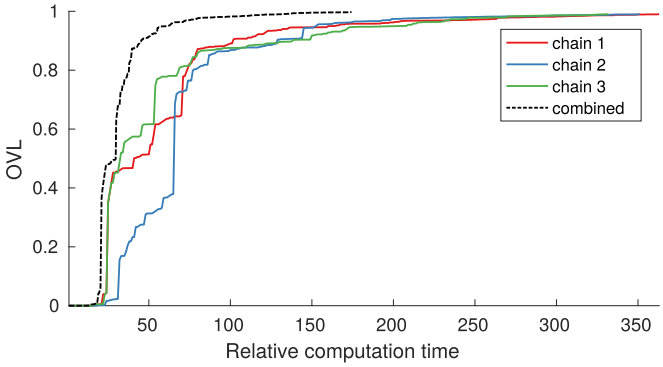

Fig. 8. Illustration of the benefit gained by using multiple parallel chains. The colored traces represent the overlapping coefficient between the posterior approximation given by that chain and the full model posterior as a function of the relative computation time. The dashed black trace represents the OVL between the full model posterior and the approximation created by combining the information of the independent chains. The approximation computed using information from all the three parallel chains reaches overlap of one in a shorter time.

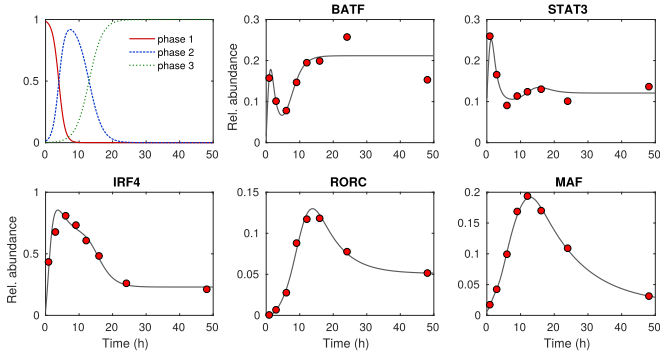

Fig. 9. The Th17 time-course data set used in the real world application along with the predictions given by the best found model. The top left panel illustrates the three-phase latent process that was fitted for the best found model. In other panels, gray lines are the output of the best model for each of the five genes and red dots correspond to measurements.

3.2 Combining Information from Multiple Chains

The previous experiments demonstrated that a single Metropolis-type sampler can indeed find the high-probability models and thus give a reliable approximation for the model posterior distribution. However, it is possible that a Metropolis chain gets stuck in a local optimum and does not escape it in a finite number of iterations. The inference results can then be different for chains that are started from different initial models. In order to obtain more reliable results, one can start several independent chains in parallel, possibly from different initial models, and create the posterior approximation in Eq. (17) using all evaluated models from each chain. Comparing different simulated chains is a common strategy for assessing the convergence of MCMC runs in continuous spaces [23]. For example, when sampling a multimodal distribution, results have a higher chance of being reliable, if all chains yield similar samples. On the other hand, if one chain has only sampled one mode and another chain has only sampled another mode for the same number of iterations, one cannot combine the samples. This is because the combined set of samples cannot generally be seen as a sample from the target distribution, since we have equally many samples from both modes, even though one of the modes might have a considerably larger total probability mass. However, in the case of a discrete model space, our information consists of the posterior probabilities for each model that has been sampled at least once by at least one chain. All this information can now be used to create the posterior approximation. This can be especially beneficial, if the independent chains can be run in parallel.

We demonstrate this strategy by returning to the model structure inference problem and the data set considered in Experiment 1. We start three independent chains from randomly chosen initial models, and test how much faster we would reach a good OVL if the chains were run in parallel and their information was combined after each iteration. Fig. 8 shows the OVL between the full model posterior and the approximations obtained by each chain on their own and the method of combining their information as a function of relative computation time. The relative computation time is defined to be the number of encountered models divided by the number of chains. For the individual chains, the horizontal axis is thus just the number of encountered models. Clearly, the latter method provides better results in a shorter time, as it always has more information than just a single chain.

3.3 Structure Inference for the Th17 Core Regulatory Network

The five-gene core network that we study was experimentally derived in [22]. The five transcription factors that are included in the network are the retinoic acid receptor-related orphan receptor gamma t (RORC), signal transducer and activator of transcription 3 (STAT3), basic leucine zipper transcription factor (BATF), transcription factor Maf (MAF) and interferon regulatory factor 4 (IRF4). We set the fixed part of the network such that each gene degrades at a constant rate and experiences a basal activation. An exception to this is RORC which has no basal activation. In addition, Fig. 10 shows the 15 possible free regulatory mechanisms that are motivated by [22]. The time course data are presented in Fig. 9.

We use the parameter bounds $[0.001, 100]$ for kinetic rate parameters and $[0.5, 3]$ for λ_1 and λ_2. Furthermore, for models with three latent states, we use the bounds $[1, 20]$ for τ_1 and $[4.5, 20]$ for τ_2 and for models with two latent states, our bounds for τ_1 are $[1, 40]$. The flexibility of the three-phase latent processes allowed with this setting is illustrated in Supplementary Fig. 1, available online.

The data set consists of measured FPKM values at timepoints 0, 1, 3, 6, 9, 12, 16, 24, and 48 hours. At time $t = 6$ h, there are three replicates but we treat those as a single measurement that is the mean of three replicates in order to get results that can more easily be compared with the ones in [9]. The initial values for the ODE models are fixed according to the measurements at time $t = 0$ h.

In likelihood computations, an underlying assumption that we make is that the measurement errors come from a normal distribution with heteroscedastic variance. The likelihood in Eq. (4) takes the form

$$p(D \mid \theta, Z) = \prod_{i=1}^{5} \prod_{k=1}^{8} \mathcal{N}(D_{ik} \mid y_{ik}, (\alpha + \beta y_{it})^2), \qquad (23)$$

$y_{it} = y_i(t, Z, \theta)$. We set the parameter values $\alpha = 10^{-4}$ and $\beta = 0.035$, which are taken from the previous study [9]. Also an approach where these parameters are estimated

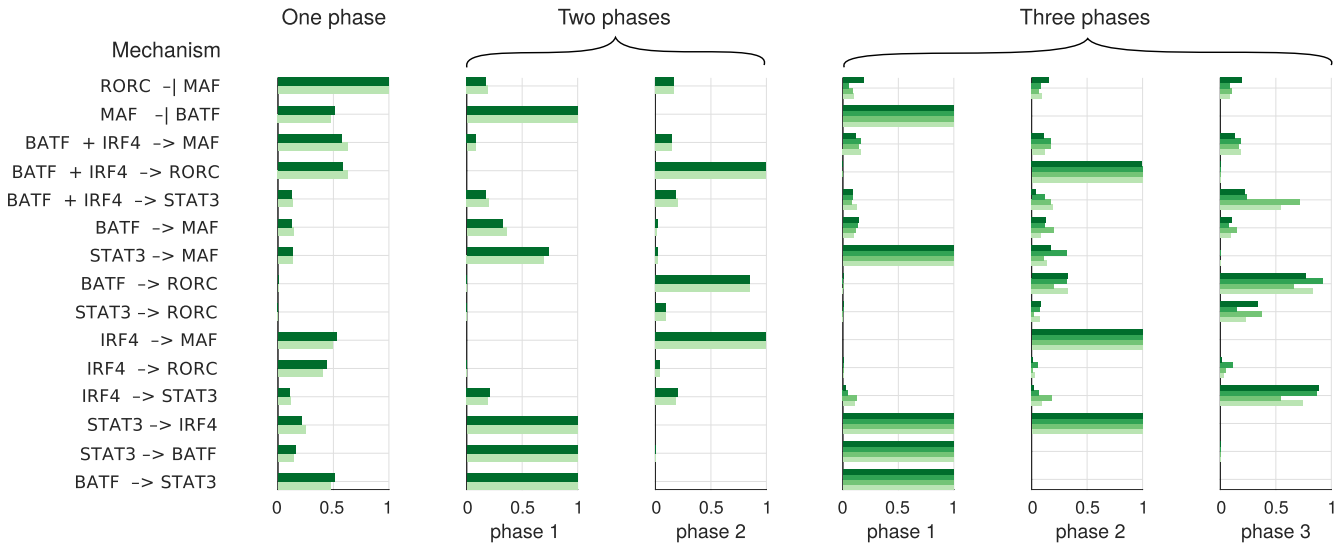

Fig. 10. Th17 core network inference results under different hypotheses about the number of distinct phases in the cellular differentiation. The bar length represents the posterior weight of the corresponding mechanism in the corresponding phase. The posterior weights obtained from different independent chains are color coded. Two independent chains were run for one- and two-phase models, and four chains for three-phase models.

simultaneously with the model parameters was applied, but this seemed to fit the models highly towards the first few data points and give very small values for α.

The model structure inference problem is tackled by starting independent Metropolis samplers from the empty model with different random seeds and comparing how the model posterior approximations provided by the independent chains converge close to each other. We perform the inference separately for three different hypotheses about M, the number of phases in the cellular differentiation. We start four independent chains in the case $M = 3$, and two independent chains in the cases $M = 1$ and $M = 2$. The number of alternative models in each case is $2^{15\times M}$ (i.e., 3.3×10^4 for $M = 1$, 1.1×10^9 for $M = 2$ and 3.5×10^{13} for $M = 3$), which means that exhaustive computation of the full model posterior is not computationally feasible.

The overlapping coefficients between the two independent chains for $M = 1$ and $M = 2$, and between chains 1 and 2 as well as chains 3 and 4 for $M = 3$, are shown in Fig. 11. We notice that in the first two cases the OVL approaches one and we get two model posterior approximations that are very similar. Thus we believe that the algorithm has explored the high probability regions accurately in both cases. In the first case, we see a temporary drop in the OVL trace when one chain moves to a new high posterior probability region and the other chain finds it only later. For $M = 3$, the model space is very large and our approximations obtained by independent chains only partially overlap after approximately 3600 evaluated models. Nevertheless, despite non-perfect overlap for $M = 3$, posterior probabilities for different mechanisms are surprisingly stable as shown below.

The posterior weight approximations for each mechanism in each phase from the different chains are presented in Fig. 10. For one- and two-phase models, the weights are naturally similar since our posterior approximations overlap remarkably. In addition, for the four chains that explore three-phase models, the obtained posterior weights are rather close to each other even though the OVL for the corresponding model posterior approximations does not reach one. The simulated data experiments, where good approximations to the weights were obtained even when the model posterior approximation did not fully overlap with the real model posterior, support our belief that the algorithm has provided us with meaningful information about the model structure given the mechanisms included in the network.

Hypotheses about the number of phases M are compared by computing the corresponding posterior probabilities $p(H_M \mid D)$, where H_M suggest that there are M different phases. The hypothesis H_3 gets practically all the posterior probability mass ($P(H_3 \mid D) \approx 1$), since models with three phases fit the data much better than ones with only one or two phases.

To summarize, our analysis supports the earlier findings that Th17 lineage specification occurs in three sequential phases [12]. Further, the posterior weights for different mechanisms (Fig. 10) obtained here with the novel and efficient search strategy coincide with the point estimate that was obtained in earlier LEM model analysis [9] through a greedy search. Additionally, the results in Fig. 10 provide interesting additional information about the uncertainty related to different mechanisms.

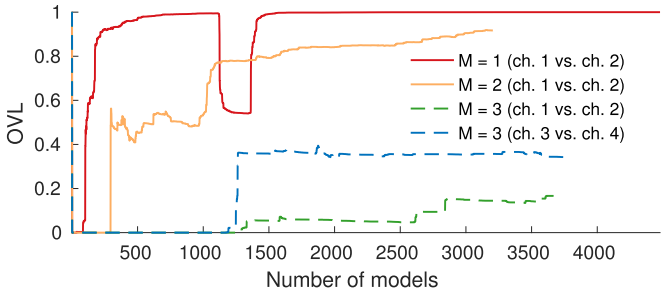

Fig. 11. Illustration of the converge of different chains in the Th17 core network structure inference under different hypotheses about the number of phases in cellular differentiation. Lines represent the overlapping coefficients (OVL) between approximations obtained from different independent chains as a function of evaluated models.

4 Discussion and Conclusion

In this study, we introduce a general framework which can be used to make inference about the structure of mechanistic ODE models. Our approach is especially suitable for inference problems in which the model space is rather large and exhaustive model evaluation is out of reach. The framework that we propose is built upon well-established MCMC techniques and provides an efficient yet simple means to obtain an approximation of the posterior distribution over alternative model structures. To our best knowledge, our approach is unique in the context of mechanistic ODE models and our results illustrate the applicability of the proposed algorithm.

The good performance of our sampling algorithm is due to the efficient parallel implementation and, on the other hand, due to successful choice of the proposal distribution of Metropolis sampler. A good proposal distribution enables efficient exploration of the model space as well as fast convergence to the high probability region of the distribution. However, the choice of the proposal distribution is at least to some extent an application specific problem and, in practice, it is always possible that the sampler gets stuck to a local mode. Consequently, it is of great importance to run several MCMC chains and carefully check how well the chains converge and cover the same high probability region (s). Good convergence of the chains is highly desirable, but it is important to note that chains that cover only partly the same high probability regions can also result in a good approximation of the target distribution because in discrete space we obtain rigorous probabilities of each visited state. Further, it is obvious that the resulting approximation provides us always with much richer information when compared with a point estimate that is obtained through a deterministic greedy search.

In applications in which the choice of the proposal distribution turns out to be especially challenging, the convergence properties can also be improved not only by changing the proposal distribution but also by constructing a different version of the sampler. For instance, population-based MCMC sampling [24] can be used to improve the mixing performance of the sampler or multiple-try Metropolis algorithm [25] can be used to make the sampler explore the space more efficiently. We have run preliminary tests using these two sampler variants and the results are promising. Importantly, both extensions can notably benefit from the parallelisms and lookup strategy in our implementation and we find these extensions as interesting future direction for sampler development.

The framework that we introduce in this study enables probabilistic model structure inference for mechanistic ODE models in larger scale than existing standard techniques. The inference problem is formulated within Bayesian framework and can be flexibly implemented in various forms. For example, the BIC approximation for the marginal likelihood can be replaced with more rigorous approximations such as the harmonic mean estimator [26] or power posterior estimator [27]. Thus, our framework provides an ODE modeller with a flexible and practical tool which makes it easier to carry out data-driven mechanistic modeling studies also when only weak prior information is available about the model structure.

Acknowledgments

This work has been supported by the Academy of Finland [Centre of Excellence in Molecular Systems Immunology and Physiology Research (2012-2017) as well as the project 275537]. The authors wish to acknowledge the Aalto Science-IT project and CSC – IT Center for Science, Finland, for computational resources.

References

[1] A. Raue, B. Steiert, M. Schelker, C. Kreutz, T. Maiwald, H. Hass, J. Vanlier, C. Tönsing, L. Adlung, R. Engesser, W. Mader, T. Heinemann, J. Hasenauer, M. Schilling, T. Höfer, E. Klipp, F. Theis, U. Klingmüller, B. Schöberl, and J. Timmer, "Data2Dynamics: A modeling environment tailored to parameter estimation in dynamical systems," *Bioinf.*, vol. 31, no. 21, pp. 3558–3560, 2015.

[2] A. Raue, M. Schilling, J. Bachmann, A. Matteson, M. Schelker, D. Kaschek, S. Hug, C. Kreutz, B. D. Harms, F. J. Theis, U. Klingmüller, and J. Timmer, "Lessons learned from quantitative dynamical modeling in systems biology," *PLoS One*, vol. 8, no. 9, pp. 1–17, 2013.

[3] M. Girolami, "Bayesian inference for differential equations," *Theoretical Comput. Sci.*, vol. 408, no. 1, pp. 4–16, Nov. 2008.

[4] F. Fröhlich, B. Kaltenbacher, F. J. Theis, and J. Hasenauer, "Scalable parameter estimation for genome-scale biochemical reaction networks," *PLoS Comp. Biol.*, vol. 13, no. 1, pp. 1–18, Jan. 2017.

[5] V. Vyshemirsky and M. A. Girolami, "Bayesian ranking of biochemical system models," *Bioinf.*, vol. 24, no. 6, pp. 833–839, 2008.

[6] N. Friel and J. Wyse, "Estimating the evidence–A review," *Statistica Neerlandica*, vol. 66, no. 3, pp. 288–308, 2012.

[7] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan, "Causal protein-signaling networks derived from multiparameter single-cell data," *Sci.*, vol. 308, no. 5721, pp. 523–529, 2005.

[8] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognit. Lett.*, vol. 15, no. 11, pp. 1119–1125, 1994.

[9] J. Intosalmi, K. Nousiainen, H. Ahlfors, and H. Lähdesmäki, "Data-driven mechanistic analysis method to reveal dynamically evolving regulatory networks," *Bioinf.*, vol. 32, no. 12, pp. 288–296, 2016.

[10] E.-M. Geissen, J. Hasenauer, S. Heinrich, S. Hauf, F. J. Theis, and N. E. Radde, "MEMO: Multi-experiment mixture model analysis of censored data," *Bioinf.*, vol. 32, no. 16, pp. 2464–2472, 2016.

[11] B. P. Ingalls, *Mathematical Modeling in Systems Biology: An Introduction*. Cambridge, MA, USA: MIT Press, 2013.

[12] N. Yosef, A. K. Shalek, J. T. Gaublomme, H. Jin, Y. Lee, A. Awasthi, C. Wu, K. Karwacz, S. Xiao, M. Jorgolli, D. Gennert, R. Satija, A. Shakya, D. Y. Lu, J. J. Trombetta, M. R. Pillai, P. J. Ratcliffe, M. L. Coleman, M. Bix, D. Tantin, H. Park, V. K. Kuchroo, and A. Regev, "Dynamic regulatory network controlling TH17 cell differentiation," *Nature*, vol. 496, no. 7446, pp. 461–468, 2013.

[13] B. D. Ripley, *Pattern Recognition and Neural Networks*, 1st ed. Cambridge, U.K.: Cambridge Univ. Press, 1996.

[14] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.

[15] J. S. Rosenthal, "Optimal proposal distributions and adaptive MCMC," in *Handbook of Markov Chain Monte Carlo*, S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, eds. London, U.K.: Chapman & Hall/CRC, 2011.

[16] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *J. Chemical. Phys.*, vol. 21, no. 6, pp. 1087–1092, 1953.

[17] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan, "An introduction to MCMC for machine learning," *Mach. Learn.*, vol. 50, no. 1, pp. 5–43, Jan. 2003.

[18] M. S. Weitzman, "Measure of the overlap of income distribution of white and negro families in the United States," U.S. Dept. Comm. Bur. Cen., Wash., DC, Tech. Rep. 22, 1970.

[19] H. F. Inman and E. L. Bradley Jr, "The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities," *Commun. Statist. Theory Methods*, vol. 18, no. 10, pp. 3851–3874, 1989.

[20] A. C. Hindmarsh, P. N. Brown, K. E. Grant, S. L. Lee, R. Serban, D. E. Shumaker, and C. S. Woodward, "SUNDIALS: Suite of nonlinear and differential/algebraic equation solvers," *ACM Trans. Math. Softw.*, vol. 31, no. 3, pp. 363–396, 2005.
[21] J. R. Leis and M. A. Kramer, "The simultaneous solution and sensitivity analysis of systems described by ordinary differential equations," *ACM Trans. Math. Softw.*, vol. 14, no. 1, pp. 45–60, 1988.
[22] M. Ciofani, A. Madar, C. Galan, M. Sellars, K. Mace, F. Pauli, A. Agarwal, W. Huang, C. N. Parkurst, M. Muratet, K. M. Newberry, S. Meadows, A. Greenfield, Y. Yang, P. Jain, F. K. Kirigin, C. Birchmeier, E. F. Wagner, K. M. Murphy, R. M. Myers, R. Bonneau, and D. R. Littman, "A validated regulatory network for Th17 cell specification," *Cell*, vol. 151, no. 2, pp. 289–303, 2012.
[23] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin, *Bayesian Data Analysis*, 3rd ed., Boca Raton, FL, USA: CRC Press, 2013.
[24] A. Jasra, D. A. Stephens, and C. C. Holmes, "On population-based simulation for static inference," *Statist. Comput.*, vol. 17, no. 3, pp. 263–279, 2007.
[25] J. S. Liu, F. Liang, and W. H. Wong, "The multiple-try method and local optimization in Metropolis sampling," *J. Amer. Statistical Assoc.*, vol. 95, no. 449, pp. 121–134, 2000.
[26] M. A. Newton and A. E. Raftery, "Approximate Bayesian inference with the weighted likelihood bootstrap," *J. Roy. Statistical Soc. Series B (Stat. Methodol.)*, vol. 56, no. 1, pp. 3–48, 1994.
[27] N. Friel and A. N. Pettitt, "Marginal likelihood estimation via power posteriors," *J. Roy. Statistical Soc. Series B (Stat. Methodol.)*, vol. 70, no. 3, pp. 589–607, 2008.

Juho Timonen received the MSc (Tech.) degree from Aalto University, Finland, in 2018, majoring in applied mathematics. His research interests include for example statistical inference methods, model assessment and selection, Bayesian hierarchical models, and non-parametric models such as Gaussian processes.

Henrik Mannerström is working toward the degree in mathematics at the University of Helsinki and works in research at Aalto University. His research interests include computational and theoretical aspects of Bayesian statistics and bioinformatics.

Harri Lähdesmäki received the MSc and DSc (Tech.) degrees from the Tampere University of Technology, in 2001 and 2005, respectively. Currently, he holds an associate professor position in the Department of Computer Science with the Aalto University in Finland. His research interests include computational biology, bioinformatics, probabilistic modeling, statistical data analysis, and machine learning.

Jukka Intosalmi received the MSc degree in mathematics and statistics from the University of Tampere, Tampere, Finland, in 2007 and the DSc (Tech.) degree in applied mathematics from the Tampere University of Technology, Tampere, Finland, in 2012. Currently, he works as a postdoctoral researcher with the Department of Computer Science in Aalto University, Espoo, Finland. His research interests include statistical data analysis, computational statistics, machine learning, dynamical systems modeling, and applications in the field of systems biology.