

Mathematical models are widely used to study natural systems. They allow us to test and generate hypotheses, and help us to understand the processes underlying the observed behaviour. However, such models are, by necessity, simplified representations of the true systems, so it is critical to understand the impact of assumptions made when using a particular model. Here we provide a method to assess how uncertainty about the structure of a natural system affects the conclusions we can draw from mathematical models of its dynamics. We use biological examples to illustrate the importance of considering uncertainty in both model structure and parameters. We show how solely considering the latter source of uncertainty can result in misleading conclusions and incorrect model inferences.

# Topological sensitivity analysis for systems biology

Ann C Babbie <sup>\*,1</sup>, Paul Kirk <sup>\*,1</sup>, and Michael P H Stumpf <sup>\*,2</sup>

<sup>\*</sup>Imperial College London, London SW7 2AZ, United Kingdom

Submitted to Proceedings of the National Academy of Sciences of the United States of America

**Mathematical models of natural systems are abstractions of much more complicated processes. Developing informative and realistic models of such systems typically involves suitable statistical inference methods, domain expertise and a modicum of luck. Except for cases where physical principles provide sufficient guidance it will also be generally possible to come up with a large number of potential models that are compatible with a given natural system and any finite amount of data generated from experiments on that system. Here we develop a computational framework to systematically evaluate potentially vast sets of candidate differential equation models in light of experimental and prior knowledge about biological systems. This topological sensitivity analysis enables us to evaluate quantitatively the dependence of model inferences and predictions on the assumed model structures. Failure to consider the impact of structural uncertainty introduces biases into the analysis and potentially gives rise to misleading conclusions.**

robustness analysis | parameter inference | model construction | model uncertainty

Abbreviations: GP, Gaussian process; ODE, ordinary differential equation; PSA, parameter sensitivity analysis; TSA, topological sensitivity analysis

Using simple models to study complex systems has become standard practice in different fields, including systems biology, ecology, and economics. Although we know and accept that such models do not fully capture the complexity of the underlying systems, they can nevertheless provide meaningful predictions and insights [1]. A successful model is one that captures the key features of the system, while omitting extraneous details that hinder interpretation and understanding. Constructing such a model is usually a non-trivial task involving stages of refinement and improvement.

When dealing with models that are (necessarily and by design) gross over-simplifications of the reality they represent, it is important that we are aware of their limitations and do not seek to over-interpret them. This is particularly true when modelling complex systems for which there are only limited or incomplete observations. In such cases, we expect there to be numerous models that would be supported by the observed data, many (perhaps most) of which we may not yet have identified. The literature is awash with papers in which a single model is proposed and fitted to a dataset, and conclusions drawn without any consideration of: (i) possible alternative models that might describe the observed behaviour and known facts equally well (or even better); or (ii) whether the conclusions drawn from different models (still consistent with current observations) would agree with one another.

We propose an approach to assess the impact of uncertainty in model structure on our conclusions. Our approach is distinct from — and complementary to — existing methods designed to address structural uncertainty, including model selection, model averaging and ensemble modelling [2–9]. Analogous to parametric sensitivity analysis (PSA), which assesses the sensitivity of a model's behaviour to changes in parameter values, we consider the sensitivity of a model's output to changes in its inherent structural assumptions. PSA techniques can usually be classified as: (i) local analyses, in which we identify a single 'optimal' vector of parameter values, and then quantify the degree to which small perturbations to these

values change our conclusions or predictions; or (ii) global analyses, where we consider an ensemble of parameter vectors (e.g. samples from the posterior distribution in the Bayesian formalism) and quantify the corresponding variability in the model's output. While several approaches fall within these categories [10–12], all implicitly condition on a particular model architecture. Here we present a method for performing sensitivity analyses for ordinary differential equation (ODE) models where the architecture of these models is not perfectly known, which is likely to be the case for all realistic complex systems. We do this by considering network representations of our models, and assessing the sensitivity of our inferences to the network topology. We refer to our approach as *topological sensitivity analysis* (TSA).

Here we illustrate TSA in the context of parameter inference, but we could also apply our method to study other conclusions drawn from ODE models (e.g. model forecasts or steady state analyses). When we use experimental data to infer parameters associated with a specific model it is critical to assess the uncertainty associated with our parameter estimates [13], particularly if we wish to relate model parameters to physical (e.g. reaction rate) constants in the real world. Too often this uncertainty is estimated only by considering the variation in a parameter estimate conditional on a particular model, while ignoring the component of uncertainty that stems from potential model mis-specification. The latter can, in principle, be considered within model selection or averaging frameworks, where several distinct models are proposed and weighted according to their ability to fit the observed data [2–5]. However, the models tend to be limited to a small, often diverse, group that act as exemplars for each competing hypothesis but ignore similar model structures that could represent the same hypotheses. Moreover, we know that model selection results can be sensitive to the particular experiments performed [14].

We assume that an initial model, together with parameters or plausible parameter ranges, has been proposed to describe the dynamics of a given system. This model may have been constructed based on expert knowledge of the system, selected from previous studies, or (particularly in the case of large systems) proposed automatically using network inference algorithms [15–19], for example. Using TSA, we aim to identify how reliant any conclusions and inferences are on the particular set of structural assumptions made in this initial candidate model. We do this by identifying alterations to model topology that maintain consistency with the observed dynamics and test how these changes impact the conclusions we draw

---

Reserved for Publication Footnotes

(Figure 1). Analogous to PSA we may perform local or global analyses — by testing a small set of ‘close’ models with minor structural changes, or performing large-scale searches of diverse model topologies respectively. To do this we require efficient techniques for exploring the space of network topologies and, for each topology, inferring the parameters of the corresponding ODE models.

Even for networks with relatively few nodes (corresponding to ODE models involving few interacting entities), the number of possible topologies can be enormous. Searching this ‘model space’ presents formidable computational challenges. We here employ a gradient-matching parameter inference approach that exploits the fact that the  $n$ -th node,  $x_n$ , in our network representation is conditionally independent of all other nodes given its regulating parents,  $\text{Pa}(x_n)$  [20–26]. The exploration of network topologies is then reduced to the much simpler problem of considering, independently for each  $n$ , the possible parent sets of  $x_n$  in an approach that is straightforwardly parallelised.

We use biological examples to illustrate local and global searches of model spaces to identify alternative model structures that are consistent with available data. In some cases we find that even minor structural uncertainty in model topology can render our conclusions — here parameter inferences — unreliable and make PSA results positively misleading. However, other inferences are robust across diverse compatible model structures, allowing us to be more confident in assigning scientific meaning to the inferred parameter values.

## Model structures

We consider systems consisting of  $N$  interacting variables that can be modelled using ODEs of the form,

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), t; \theta), \quad [1]$$

where  $\mathbf{x}(t) = [x_1(t), \dots, x_N(t)]$ ,  $x_n(t)$  is the value of the  $n$ -th variable at time  $t$ ,  $\dot{\mathbf{x}}(t)$  is the derivative of  $\mathbf{x}(t)$  with respect to  $t$ , and  $\theta$  is the vector of model parameters. The rate of change of the  $n$ -th variable over time is described by the  $n$ -th component of the vector  $\dot{\mathbf{x}}(t)$ ,

$$\dot{x}_n(t) = f_n(\mathbf{x}(t), t; \theta). \quad [2]$$

Typically,  $f_n$  will only act on a (relatively small) subset of the variables  $\mathbf{x}(t)$ ; we represent our systems as networks, in which nodes correspond to variables and a directed edge is drawn from  $x_m$  to  $x_n$  if the rate of change of  $x_n$  depends upon  $x_m$ . The set of parents of  $x_n$ ,  $\text{Pa}(x_n) \subseteq \{x_1, \dots, x_N\}$ , is therefore the collection of variables on which  $f_n$  acts. The parent set,  $\mathcal{C} = \{\text{Pa}(x_n)\}_{n=1}^N$ , provides a complete description of the system’s network representation.

## Defining the model search space

To specify the space of possible ODE models we need to define (i) the set of network topologies, and (ii) the corresponding ODE model(s) that can be represented by each network topology. Exploring large model spaces is, of course, challenging due to the number of possible topologies. However, since each network is completely defined by the parent set,  $\mathcal{C}$ , it is sufficient to consider (for each species  $n$  independently) the possible parents of  $x_n$ ,  $\text{Pa}(x_n)$ , in order to fully explore the space of networks. Considering the regulation of each species independently in this way can reduce the search space sufficiently to make an exhaustive search feasible (Figure S1A). Further simplification, such as assuming that each parent set,

$\text{Pa}(x_n)$ , is restricted to a relatively small size as in [23–25], may also be helpful.

To move from a network representation to an ODE model requires a set of rules that translate interactions (network edges) into parameterised mathematical functions. These rules are necessarily context- and model-specific and may allow a single network topology to represent multiple different ODE models. For example, we could permit several types of interaction (e.g. activation and inhibition) each represented by a different mathematical function, or several methods to model the combined effect of multiple interactions (e.g. synergistic and additive). Such rules allow us to map from each possible parent set,  $\text{Pa}(x_n)$ , to the possible functional forms of the corresponding differential equations,

$$\dot{x}_n(t) = f_n(\text{Pa}(x_n), t; \theta). \quad [3]$$

In order to test whether a particular ODE model can generate the desired dynamic behaviour, we require the associated model parameters,  $\theta$ . Crucially, to enable large-scale searches we need a method to infer the parameters of each ODE derivative component independently. Gradient-matching parameter inference approaches [26–28] avoid the need to simulate from the complete coupled ODE system (Equation 1); instead, parameters are optimised using data-derived estimates of  $\dot{\mathbf{x}}(t)$  and  $\mathbf{x}(t)$ , and minimising the discrepancy between two gradient estimates —  $\dot{x}_n(t)$  and  $f_n(\text{Pa}(x_n), t; \theta)$  — in a process which can be applied to each derivative component independently (Figure S1B). Here we use Gaussian process (GP) regression, a non-parametric Bayesian method for non-linear regression [28–31], to obtain data-derived estimates of  $\dot{\mathbf{x}}(t)$  and  $\mathbf{x}(t)$  (see SI Methods for details).

## Constructing and ranking ODE models

Given an initial parameterised candidate model we use the following method to identify and rank realistic alternative models:

1. Simulate time course data (at times  $t = t_1, \dots, t_T$ ) for each species in the system using the initial candidate model.
2. Define rules to construct possible ODE models for the regulation of each species  $x_n$ . We consider: the maximum number of parents allowed per species; possible parametric forms to represent interactions; and how to model combinatorial regulation if a node has several parents. The form of the initial candidate model can guide these rules.
3. Estimate state variables  $\hat{x}_n(t)$  and corresponding derivatives  $\hat{\dot{x}}_n(t)$  for all species in the system using GP regression and the simulated time course data (generated in step 1).
4. For each species  $n$ :
  - (a) Consider all possible parent sets in order to construct all possible models of the form  $\dot{x}_n(t) = f_n(\text{Pa}(x_n), t; \theta)$ , according to the rules defined in step 2.
  - (b) Infer parameters  $\theta$  for each test model using gradient-matching parameter inference and GP-derived estimates for  $\hat{\dot{x}}_n(t)$  and  $\hat{\mathbf{x}}(t)$ . We estimate  $\theta$  by minimising,

$$\text{distance} = \sum_{t=t_1}^{t_T} (\hat{\dot{x}}_n(t) - f_n(\hat{\mathbf{x}}(t), t; \theta))^2. \quad [4]$$

- (c) Rank models using the distance calculated in step 4(b) or an alternative metric, e.g. Akaike’s Information Criterion adapted for small samples ( $\text{AIC}_c$ ) [2].
5. Combine the component-wise models,  $\dot{x}_n(t)$ , to obtain complete ODE models  $\dot{\mathbf{x}}(t)$ .

Selected models are then used to explore how uncertainty in model structure impacts our inferences and/or predictions.

## Results

We first outline how to generate alternative models which generate dynamics consistent with the initial candidate model. We then use synthetic and experimental datasets to demonstrate how we can use selected models to test the robustness of our inferred conclusions (here, parameter estimates) to altered model assumptions. We include examples using both optimization and Bayesian approaches to parameter inference. ODE models were constructed as described in Materials and Methods. SI Methods lists the parameter values and initial conditions used for simulations from the initial candidate models.

**Automated model generation and ranking.** To illustrate, we assume a five-species gene regulatory network, Model A (Figure 6), has been proposed to model a hypothetical system. We suppose that this network corresponds to an ODE model for which parameters (including initial conditions) have already been estimated (see SI Methods for values). We take this as our initial candidate model, and use GP regression to estimate species concentrations,  $\hat{x}_n(t)$ , and corresponding derivatives,  $\hat{\dot{x}}_n(t)$ , from simulated trajectories (Figure 2). For  $n = 1, \dots, 5$ , we define 33 possible component-wise models of the form,  $\hat{x}_n(t) = f_n(\text{Pa}(x_n), t; \theta)$ , to describe the regulation of each species by allowing up to two regulating parents per gene, no self-regulation, and two types of interaction. For each putative parental set,  $\text{Pa}(x_n)$ , we test all permutations of interaction types — e.g. if  $\text{Pa}(x_1) = \{x_2, x_3\}$  we consider  $\text{Pa}(x_1) = \{x_2^+, x_3^+\}$ ,  $\{x_2^-, x_3^-\}$ ,  $\{x_2^-, x_3^+\}$  or  $\{x_2^+, x_3^-\}$ , where the superscript indicates activation (+) or inhibition (−). We rank these models using the distance (Equation 4) obtained during gradient-matching parameter inference (Figure S2A), and combine selected component-wise models to create coupled ODE models describing the network dynamics (Figure 2). Crucially, this combination step does not require further parameter estimation so by evaluating the  $33 \times 5 = 165$  possible component equations for  $\hat{x}_n(t)$ , we can easily — and rapidly — construct and rank all  $33^5 = 3.9 \times 10^7$  complete ODE systems,  $\hat{\mathbf{x}}(t)$ . If we have information about the system dynamics under multiple experimental conditions we can consider this when ranking our complete models (see SI Results and Figure S2).

### Global TSA of optimisation-based parameter estimation.

Once we have identified alternative ODE models with comparable dynamics to our initial model, we can use these to explore how dependent our inferences are on the structural assumptions inherent to a particular model. We illustrate this in the context of maximum likelihood parameter estimation, using a synthetic dataset simulated from a competitive population dynamics model, Model B (Figure 6), which we also take as the initial candidate.

As in the previous example, we assume that Model B corresponds to an ODE model of a hypothetical system for which parameters have previously been estimated (see SI Methods). We additionally assume that we have an experimental dataset,  $D_B$ , which, for this synthetic example, is generated by adding noise to values simulated from the candidate model (see SI Methods). Thus, in this illustrative example, the candidate model is also the ‘true’ (data-generating) model. We first perform an exhaustive search to identify models with consistent dynamics to our initial candidate. We allow a maximum of 3 parent species, no self-regulation, and a single type of interaction resulting in  $7.6 \times 10^5$  possible complete ODE models. To illustrate results, we select for the present analysis the top ten models, ranked by AIC<sub>C</sub> values (Figure 3A). The mod-

els ranked 2<sup>nd</sup> – 10<sup>th</sup> all contain the interactions present in the true model (ranked 1<sup>st</sup>), but have one or two additional ‘incorrect’ edges.

Next we can assess the variability in the conclusions (here, maximum likelihood parameter estimates) that we obtain from the candidate models when fitting them to a noisy dataset,  $D_B$ . We assess uncertainties in the estimated values using a parametric bootstrap [22] (Figures 3B and S3). Most distributions are consistent across the models, but network topology has a significant impact on some parameters. For example, with seven models (including the true model) the distributions for parameter  $r_3$  are centered close to 0.5 (which we know to be the true value for this synthetic example) with little variation; however, models containing an incorrect interaction from species 4 to 3 (Models 5, 6 and 10) result in broader distributions centered around a higher value.

Thus, even models that include all the true interactions can still lead to misleading conclusions about biophysical parameters if additional incorrect interactions are also present. If we solely rely on parameter uncertainty and sensitivity analyses, without considering the impact of potential structural mis-specifications in our model, we are therefore likely to overestimate the precision of our results.

**Local TSA of Bayesian parameter inference.** Instead of an exhaustive, global search, we may instead perform a local analysis by only considering models with minor structural modifications relative to the candidate model. This could be appropriate if we have a clear idea about key interactions, and consider large deviations from this topology to be biologically irrelevant; or for larger complex systems where exhaustive search becomes computationally infeasible. To illustrate this we use a synthetic dataset,  $D_C$ , simulated from a population dynamics model, Model C (Figure 6), which we also take as the initial candidate. We then consider the set of 20 ‘close’ models that differ by addition or deletion of a single network edge. For each model we obtain samples from the posterior parameter distribution using two Bayesian approaches: nested sampling [32, 33] and the Metropolis-Hastings algorithm [34, 35]. As well as generating posterior samples, nested sampling provides an estimate of the evidence (marginal likelihood) for each model, allowing us to rank them (Table S1). We select the five alternative models with evidence greater than or equal to the true model and compare the estimated posterior distributions.

In most cases the univariate marginal posterior distributions are broadly conserved across the selected models (Figure S5), but in a few cases the shape and/or location of these distributions vary with a slightly different model (e.g. in Figure S5, see parameters  $r_5$  and  $a_{51}$  in Model 18, or  $a_{23}$  in Model 8). As well as the values of individual parameters, we may also be interested in the dependencies *between* parameters. In particular, the related concepts of *sloppiness* and *identifiability* in biological models have recently received much attention — in the context of possible biological significance and for optimal experimental design [12, 36–39]. Figure 4 shows bivariate scatter plots illustrating the dependencies between particular parameter pairs for each of the selected models.

Again, while some dependencies are robust to varying model topology (e.g. the negative dependence between parameters  $a_{51}$  and  $a_{54}$  is observed in all cases) sometimes even small alterations to model topology significantly alter our conclusions about parameter dependencies — with a single edge missing from the true network, we may either infer a strong dependency (e.g.  $r_2$  and  $a_{23}$  in Model 8) that is absent in the true system, or miss a true dependency (e.g.  $r_3$  and  $a_{34}$  in Model 12). Of course, in this example we know the true model,

but these results serve to illustrate how drawing conclusions about the underlying system based on inferences drawn from a single model may be misleading. When dealing with simplified representations of reality we believe the approach outlined here is a useful, even essential, way to determine which results strongly rely on model assumptions and which ones are robust to structural modification and may therefore more likely be biologically relevant rather than an artefact resulting from a specific single model topology.

**TSA of a model of yeast gene expression dynamics.** We use time-resolved gene expression data for cell-cycle regulated genes in *Saccharomyces cerevisiae* [40]. Following Lu *et al.* [41], we used their D-NetWeaver algorithm [19] to construct an initial candidate ODE model describing gene expression dynamics (Figure S7). The resulting network comprises 41 nodes (corresponding to gene clusters) and 148 directed edges (regulatory interactions). The system size precludes an exhaustive search of alternative topologies, so we performed a local analysis in which we sampled models with 1-30 random edges rewired (relative to the initial candidate) and identified those with comparable dynamics using gradient-matching.

In total we sampled  $3.5 \times 10^5$  rewired models and selected seven models for comparison with our initial candidate (see SI Results for details). We obtained maximum likelihood estimates for the 189 parameters associated with each model by fitting to the mean clustered gene expression profiles, and compared estimates for parameters common to all selected models. Figure 5 illustrates the variety of estimated values we obtained for four parameters (results for all common parameters are shown in Figure S8).

Many estimated parameter values are consistent across these models (i.e. robust to topological changes). However, Figure 5 shows one example where the sign of an edge-associated parameter (p22) varies with model choice, suggesting we cannot reliably infer the nature of this regulatory interaction (activating or inhibitory). Relying solely on estimates of parameter uncertainty that condition on a chosen model structure (e.g. the confidence intervals estimated here), but ignoring the potential impact of structural uncertainty could lead us to draw unreliable conclusions about the true network. TSA provides us with a way to assess the latter component of uncertainty, and thus gain confidence in results which are consistent amongst the selected models, while identifying less reliable inferences.

## Discussion

For mechanistic mathematical modeling to be useful, we must assess the robustness of conclusions drawn from our models. This is particularly important in fields such as systems biology, where we usually rely on (knowingly) over-simplified representations of the true complex systems [7, 43, 44]. To gain meaningful insights into the real underlying processes, we must acknowledge that our conclusions are conditional on the chosen model architecture, and understand the impact of changing these model assumptions. Frequently, once a model has been selected, researchers only consider uncertainty at the level of model parameters but ignore the contribution of potential (and likely) structural mis-specifications. One reason for this is that it has been difficult to explore a diverse range of models in a computationally efficient manner.

Here we describe a rapid, parallelisable method to automatically generate models consistent with the observed dynamics of a biological system. Constraints on the parametric

forms of these models are selected for the system of interest, to ensure that proposed models are plausible given the types of interactions believed to be possible. Within computational limitations we can consider any number of possible parametric forms to describe the dynamics and interactions within a network, and investigate different rules for combinatorial regulation. This permits an extensive search of local or global model space for alternative, dynamically consistent models. These can then be used to test the sensitivity of inferences to the structural assumptions inherent to a specific model.

We demonstrate how our approach, TSA, may be used to explore the impact of changes to our model upon the conclusions that we draw, in the particular context of parameter inference. While Bayesian analysis and bootstrapping approaches are often employed in order to assess parameter variability (and identifiability), it is important to remember that these assessments condition upon a single, specific model. Even minor changes in the model structure can significantly alter the conclusions that we draw (in our case, the values and dependencies of the inferred parameters). Even if, for one particular model, the parameter values are very tightly constrained (and thus, for that model, we feel we have ‘high confidence’ in the parameter values), it does not follow that the same will be true for other feasible models. If our parameters have a real, biophysical interpretation, we therefore need to be very careful not to assert that we know the *true* values of these quantities in the underlying system, just because – for a given model – we can pin them down with relative certainty. Considering the sensitivity of model behaviour to both parameter *and* structural variation enables us to identify inferences that are robust to possible errors in our model. We can therefore be more confident about using these inferences to draw scientifically meaningful conclusions.

## Materials and Methods

**Models for synthetic gene regulatory networks.** Transcription regulation is modelled using equations of the form,  $\dot{x}_n(t) = s_n - \gamma_n x_n + \sum_{k \in \text{Pa}(x_n)} \beta_{nk} f_{nk}$ , where  $x_n(t)$  is species  $n$  mRNA concentration,  $s_n$  and  $\gamma_n$  are basal synthesis and degradation rates respectively, and the sum accounts for regulatory interactions influencing gene  $n$  (assuming additive combinatorial regulation) [45]. An interaction from species  $k$  to species  $n$  is modelled using a strength parameter  $\beta_{nk}$  and a Hill function  $f_{nk}$  (with associated parameters  $\theta_{nk}$  and  $m_{nk}$ ). The latter takes the form,  $f_{nk}^+ = x_k(t)^{m_{nk}} / (\theta_{nk}^{m_{nk}} + x_k(t)^{m_{nk}})$ , for activating interactions or  $f_{nk}^- = 1 / (1 + (x_k(t) / \theta_{nk})^{m_{nk}})$  for inhibitory regulation.

**Models for competitive population dynamics.** Population sizes,  $x_n$ , of  $N$  species competing for finite resources are modelled by [46],  $\dot{x}_n(t) = r_n x_n(t) (1 - \sum_{k=1}^N a_{nk} x_k(t))$ , where  $r_n$  is the inherent growth rate of species  $n$  and  $a_{nk}$  represents the influence of species  $k$  on species  $n$ . Here  $a_{nn} = 1$ ,  $r_n > 0$ , and  $a_{nk} \geq 0$  for all  $n$ .

**Models for yeast gene expression data.** We used D-NetWeaver [19] to construct an initial candidate model for the 613 complete  $\alpha$ -factor synchronised gene-expression profiles in ref. [40]. As in ref. [41], genes were clustered into 41 groups based on expression profile similarities, and a linear ODE model inferred to describe regulatory interactions between these clusters with the form,  $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}$ , where  $\mathbf{x}(t) = [x_1(t), \dots, x_N(t)]^T$ ,  $x_n(t)$  is the mean expression level of genes in cluster  $n$  at time  $t$ ,  $\mathbf{A}$  is a connectivity matrix with entry  $a_{ij} \neq 0$  indicating a directed interaction from cluster  $j$  to cluster  $i$  (with  $a_{ij} > 0$  and  $a_{ij} < 0$  indicating activation or inhibition respectively),  $\mathbf{b} = [b_1, \dots, b_N]^T$  is a vector of constant terms, and  $N$  is the total number of clusters.

**ACKNOWLEDGMENTS.** This work was supported by HFSP Grant RG0061/2011 and BBSRC Grant BB/K017284/1. MPHS is Royal Society Wolfson Research Merit Award Holder.

1. Box GEP, Draper NR (1987) Empirical model-building and response surfaces. (John Wiley & Sons).
2. Burnham KP, Anderson DR (2002) Model Selection and Multimodel Inference. A Practical Information-Theoretic Approach (Springer).
3. Kirk P, Thorne T, Stumpf MPH (2013) Model selection in systems and synthetic biology. *Curr Opin Biotechnol* 24(4):767–774.
4. Toni T, Stumpf MPH (2010) Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics* 26(1):104–110.
5. Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: A tutorial. *Statistical Science* 14(4):382–417.
6. Kuepfer L, Peter M, Sauer U, Stelling J (2007) Ensemble modeling for analysis of cell signaling dynamics. *Nat Biotechnol* 25(9):1001–1006.
7. Kaltenbach H-M, Dimopoulos S, Stelling J (2009) Systems analysis of cellular networks under uncertainty. *FEBS Lett* 583:3923–3930.
8. Sunnåker M, et al. (2013) Automatic generation of predictive dynamic models reveals nuclear phosphorylation as the key Msn2 control mechanism. *Sci Signal* 6(277):ra41.
9. Sunnåker M, et al. (2014) Topological augmentation to infer hidden processes in biological systems. *Bioinformatics* 30(2):221–7.
10. Rabitz H (1989) Systems analysis at the molecular scale. *Science* 246(4927):221–226.
11. Saltelli A, et al. (2008) Global Sensitivity Analysis. The Primer (John Wiley & Sons).
12. Erguler K, Stumpf MPH (2011) Practical limits for reverse engineering of dynamical systems: a statistical analysis of sensitivity and parameter inferability in systems biology models. *Mol Biosyst* 7(5):1593.
13. Vanlier J, Tiemann CA, Hilbers PAJ, van Riel NAW (2013) Parameter uncertainty in biochemical models described by ordinary differential equations. *Math Biosci* 246(2):305–314.
14. Silk D, Kirk PDW, Barnes CP, Toni T, Stumpf MPH (2014) Model Selection in Systems Biology Depends on Experimental Design. *PLoS Comput Biol* 10(6):e1003650.
15. Bonneau R, et al. (2006). The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol* 7(5), R36.
16. Greenfield A, Hafemeister C, Bonneau R (2013). Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics* 29(8), 1060–1067.
17. Thorne T, Stumpf, MPH (2012). Inference of temporally varying Bayesian networks. *Bioinformatics* 28(24), 3298–3305.
18. Oates CJ, et al. (2014). Causal network inference using biochemical kinetics. *Bioinformatics* 30(17), i46874.
19. Wu S, Liu Z-P, Qiu X, Wu H (2014). Modeling genome-wide dynamic regulatory network in mouse lungs with influenza infection using high-dimensional ordinary differential equations. *PLoS One* 9(5), e95276.
20. Friedman N, Geiger D, Goldszmidt M (1997) Bayesian Network Classifiers. *Machine learning* 29(2-3):131–163.
21. Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian networks to analyze expression data. *J Comput Biol* 7(3-4):601–620.
22. Murphy KP (2012) Machine Learning. A Probabilistic Perspective (MIT Press, Cambridge, MA).
23. Aijö T, Lähdesmäki H (2009) Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. *Bioinformatics* 25(22):2937–2944.
24. Penfold CA, Wild DL (2011) How to infer gene networks from expression profiles, revisited. *Interface Focus* 1(6):857–870.
25. Penfold CA, Buchanan-Wollaston V, Denby KJ, Wild DL (2012) Nonparametric Bayesian inference for perturbed and orthologous gene regulatory networks. *Bioinformatics* 28(12):i233–i241.
26. Dondelinger F, Husmeier D, Rogers S, Filippone M (2013) ODE parameter inference using adaptive gradient matching with Gaussian processes. *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics* 216–228.
27. Brunel NJ-B(2008) Parameter estimation of ODEs via nonparametric estimators. *Electron J Stat* 2(0):1242–1267.
28. Calderhead B, Girolami M (2008) Accelerating Bayesian inference over nonlinear differential equations with Gaussian processes. *Adv Neural Inf Process Syst* 21:217–224.
29. Kirk PDW, Stumpf MPH (2009) Gaussian process regression bootstrapping: exploring the effects of uncertainty in time course data. *Bioinformatics* 25(10):1300–1306.
30. Rasmussen CE, Williams CKI (2006) Gaussian Processes for Machine Learning (MIT Press).
31. Rasmussen CE, Nickisch H (2010) Gaussian processes for machine learning (GPML) toolbox. *J Mach Learn Res* 11:3011–3015.
32. Skilling J (2006) Nested sampling for general Bayesian computation. *Bayesian Anal* 1(4):833–860.
33. Johnson R, Kirk P, Stumpf MPH (2014) SYSBIONS: Nested sampling for systems biology. *Bioinformatics* doi: 10.1093/bioinformatics/btu675.
34. Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97–109.
35. Wilkinson DJ (2011) Stochastic Modelling for Systems Biology, Second Edition. (CRC Press).
36. Daniels BC, Chen Y-J, Sethna JP, Gutenkunst RN, Myers CR (2008) Sloppiness, robustness, and evolvability in systems biology. *Curr Opin Biotechnol* 19(4):389–395.
37. Gutenkunst RN, et al. (2007) Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput Biol* 3(10):1871–1878.
38. Apgar JF, Witmer DK, White FM, Tidor B (2010) Sloppy models, parameter uncertainty, and the role of experimental design. *Mol Biosyst* 6(10):1890.
39. Villaverde AF, Banga JR (2013) Reverse engineering and identification in systems biology: strategies, perspectives and challenges. *J R Soc Interface* 11:20130505.
40. Spellman PT, et al. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* 9:3273–3297.
41. Lu T, Liang H, Li H, Wu H. (2011) High dimensional ODEs coupled with mixed-effects modelling techniques for dynamic gene regulatory network identification. *J Am Stat Assoc* 106(496):1242–1258.
42. Statisticat, LLC. (2014) LaplacesDemon: Complete Environment for Bayesian Inference. <http://www.bayesian-inference.com/software> (R package version 14.06.23).
43. Schaber J, Klipp E (2011) Model-based inference of biochemical parameters and dynamic properties of microbial signal transduction networks. *Curr Opin Biotechnol* 22:109116.
44. Gunawardena J (2014) Models in biology: accurate descriptions of our pathetic thinking. *BMC Biol* 12:29.
45. Mazur J, Ritter D, Reinelt G, Kaderali L (2009) Reconstructing nonlinear dynamic models of gene regulation using stochastic sampling. *BMC Bioinformatics* 10(1):448.
46. Vano JA, Wildenberg JC, Anderson MB, Noel JK, Sprott JC (2006) Chaos in low-dimensional Lotka–Volterra models of competition. *Nonlinearity* 19(10):2391–2404.

**Fig. 1.** Overview of TSA applied to parameter inference. (A) Model space includes our initial candidate model and a series of altered topologies that are consistent with our chosen rules (e.g. all 2-edge, 3-node networks, where nodes indicate species and directed edges show interactions). One topology may correspond to one or several ODE models depending on the parametric forms we choose to represent interactions. (B) We test each ODE model to see whether it can generate dynamics consistent with our candidate model and the available experimental data. For TSA, we select a group of these compatible models and compare the conclusions we would draw using each of them. (C) Associated with each model  $m$  is a parameter space  $\Theta_m$  (grey); using Bayesian methods we can infer the joint posterior parameter distribution (dashed contours) for a particular model and dataset. (D) In some cases, equivalent parameters will be present in several selected models (e.g.  $\theta_1$  which is associated with the same interaction in models  $a - c$ ). We can compare the marginal posterior distribution inferred using each model for a common parameter to test whether our inferences are robust to topological changes, or rely on one specific set of model assumptions (i.e. sensitive). Different models may result in marginal distributions that differ in position and/or shape for equivalent parameters, but we cannot tell from this alone which model better represents reality — this requires model selection approaches [2–4].

**Fig. 2.** Example of automated model generation and ranking for the 5-species gene regulatory network (Model A). GP regression model means (left, solid lines) fitted to noiseless data simulated from the initial candidate model (circles). Simulated trajectories (right) from two example alternative ODE models automatically constructed using our method (see SI Results for details) — the best models accurately capture the desired dynamics, while lower ranked models deviate.

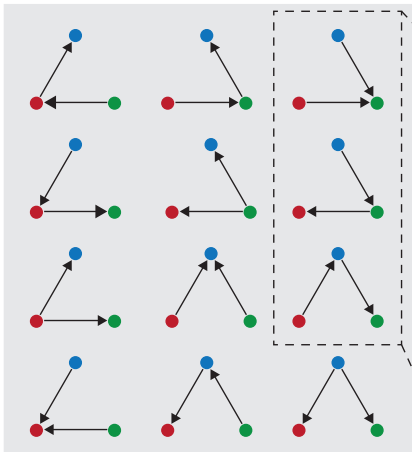
**Fig. 3.** Comparison of parameter inference results for the top 10 complete ODE models (ranked by AIC<sub>C</sub>) for a competitive population dynamics system. (A) Noisy data (circles),  $D_B$ , simulated from the true model (Model B) are used to obtain maximum likelihood parameter estimates for each of the ten models. Lines show the trajectories simulated using each of the resulting parametrised models (assuming the true initial conditions). (B) Comparison of parametric bootstrap distributions obtained using each model for three of the parameters present in all models; distributions for all the parameters present in the true model are shown in Figure S3. For each model, 1000 replicate datasets were obtained by adding Gaussian noise to the initial fitted trajectories (Figure 3A) and used for parameter estimation. Kernel density estimates of the resulting parameter distributions are shown by solid lines; vertical dashed grey lines indicate the true parameter values used for simulation. Incorrect edges, from species  $k$  to  $n$ , in models 2-10 are from the set:  $(k, n) \subseteq \{(1, 4), (1, 5), (4, 3), (5, 4), (3, 5)\}$ .

**Fig. 4.** Comparison of nested sampling inference results for the best ‘close’ models and the true model. Data simulated from the true model (Model C) assuming additive Gaussian noise are used to infer posterior parameter distributions for the true model and 20 ‘close’ models (those differing from the true model by addition or deletion of a single interaction); we assumed uniform prior distributions for all parameters ( $0.1 \leq r_n \leq 2, 0.1 \leq a_{nk} \leq 5$ ). Models with estimated evidence  $\geq$  to the true model are then compared. Example bivariate scatter plots are shown for three parameter pairs for these models; each circle shows a posterior sample, with colour corresponding to the likelihood value. Model 18 does not include an interaction from species 5 to species 4 hence there is no plot. See SI Results for additional results.

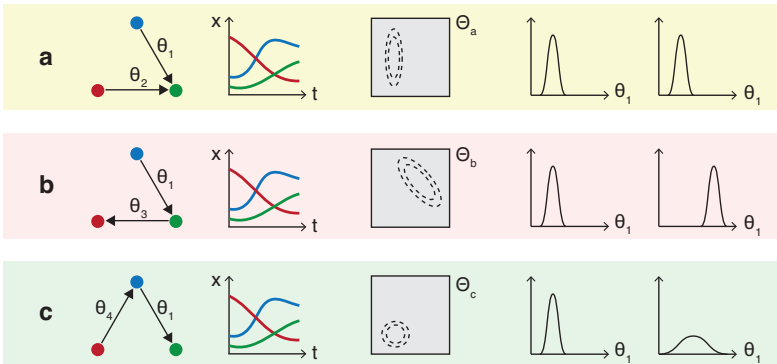
**Fig. 5.** Comparison of parameter estimation results for eight alternative ODE models fitted to the data of ref. [40]. Results for four parameters are shown; further estimates are given in Figure S8. Estimated values are shown by circles, with error bars indicating 95 % credible intervals estimated using a Laplace approximation [42]; estimates for equivalent parameters in different models are joined by a line to aid comparison. The selected models are ordered by the number of rewired edges relative to the initial candidate model (at  $x = 0$ ). Estimated values for many model parameters are robust to altered model topology (e.g. p9 and p78), but some are more sensitive (e.g. estimates for p29 are consistent in sign but not magnitude, while p22 estimates span positive and negative values in the selected models).

**Fig. 6.** Model topologies for synthetic datasets: model A represents a gene regulatory network [45], and models B and C represent population dynamics models.

(A) Model space



(B) Select models (a-c) with compatible dynamics



(C) Posterior parameter distributions

(D) Marginal posteriors for common parameters

robust sensitive



Species: ● 1 ● 2 ● 3 ● 4 ● 5

