OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Model Selection in Systems Biology Depends on Experimental Design

Daniel Silk, Paul D. W. Kirk, Chris P. Barnes, Tina Toni, Michael P. H. Stumpf*

Centre for Integrative Systems Biology at Imperial College London, London, United Kingdom

Abstract

Experimental design attempts to maximise the information available for modelling tasks. An optimal experiment allows the inferred models or parameters to be chosen with the highest expected degree of confidence. If the true system is faithfully reproduced by one of the models, the merit of this approach is clear - we simply wish to identify it and the true parameters with the most certainty. However, in the more realistic situation where all models are incorrect or incomplete, the interpretation of model selection outcomes and the role of experimental design needs to be examined more carefully. Using a novel experimental design and model selection framework for stochastic state-space models, we perform high-throughput *in-silico* analyses on families of gene regulatory cascade models, to show that the selected model can depend on the experiment performed. We observe that experimental design thus makes confidence a criterion for model choice, but that this does not necessarily correlate with a model's predictive power or correctness. Finally, in the special case of linear ordinary differential equation (ODE) models, we explore how wrong a model has to be before it influences the conclusions of a model selection analysis.

Citation: 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. doi:10.1371/journal.pcbi.1003650

Editor: Burkhard Rost, Tum, Germany

Received May 30, 2013; **Accepted** April 10, 2014; **Published** June 12, 2014

Funding: This work was funded by the Biotechnology and Biological Science Research Council (www.bbsrc.ac.uk) grant BB/K003909/1 (to DS and MPHS, the Human Frontiers Science Programme (ww.hfsp.org) grant RG0061/2011 (to PDWK and MPHS), and a Wellcome Trust-MIT fellowship (www.wellcome.ac.uk) to TT. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: m.stumpf@imperial.ac.uk

Introduction

Mathematical models provide a rich framework for biological investigation. Depending upon the questions posed, the relevant existing knowledge and alternative hypotheses may be combined and conveniently encoded, ready for analysis via a wealth of computational techniques. The consequences of each hypothesis can be understood through the model behaviour, and predictions made for experimental validation. Values may be inferred for unknown physical parameters and the actions of unobserved components can be predicted via model simulations. Furthermore, a well-designed modelling study allows conclusions to be probed for their sensitivity to uncertainties in any assumptions made, which themselves are necessarily made explicit.

While the added value of a working model is clear, how to create one is decidedly not. Choosing an appropriate formulation (e.g. mechanistic, phenomenological or empirical), identifying the important components to include (and those that may be safely ignored), and defining the laws of interaction between them remains highly challenging, and requires a combination of experimentation, domain knowledge and, at times, a measure of luck. Even the most sophisticated models will still be subject to an unknown level of inaccuracy – how this affects the modelling process, and in particular experimental design for Bayesian inference, will be the focus of this study.

Both the time and financial cost of generating data, and a growing understanding of the data dependancy of model and parameter identifiability [1,2], has driven research into experimental design. In essence, experimental design seeks experiments that maximise the expected information content of the data with respect to some modelling task. Recent developments include the work of Liepe et. al [2] that builds upon existing methods [3–8], by utilising a sequential approximate Bayesian computation framework to choose the experiment that maximises the expected mutual information between prior and posterior parameter distributions. In so doing, they are able to optimally narrow the resulting posterior parameter or predictive distributions, incorporate preliminary experimental data and provide sensitivity and robustness analyses. In a markedly different approach, Apgar et. al [8] use control theoretic principles to distinguish between competing models; here the favoured model is that which is best able to inform a controller to drive the experimental system through a target trajectory.

In order to explore the effects of model inaccuracies we work with a computationally efficient experimental design framework. We build on the methods of Flassig and Sundmacher [9] where expected likelihoods are predicted using efficient Sigma-point approximations and leveraged for optimal experimental design, and Busetto et al. [10] where choosing the optimal measurement readouts and time points is undertaken in an iterative fashion, using Sigma-point approximations to update the posterior distributions. Here we show how mixtures distributions may be exploited to cope with non-Gaussian parameter and predictive distributions and further, derive an extension to the case of stochastic state space models. The intuition behind the approach (described fully in Materials and Methods) is shown in Figure 1, where for identical inputs, two ODE models (illustrated in blue and red respectively) are simulated for a range of parameter

Author Summary

Different models of the same process represent distinct hypotheses about reality. These can be decided between within the framework of model selection, where the evidence for each is given by their ability to reproduce a set of experimental data. Even if one of the models is correct, the chances of identifying it can be hindered by the quality of the data, both in terms of its signal to measurement error ratio and the intrinsic discriminatory potential of the experiment undertaken. This potential can be predicted in various ways, and maximising it is one aim of experimental design. In this work we present a computationally efficient method of experimental design for model selection. We exploit the efficiency to consider the implications of the realistic case where all models are more or less incorrect, showing that experiments can be chosen that, considered individually, lead to unequivocal support for opposed hypotheses.

values, with times T_1 and T_2 representing two possible choices of times at which the true system can be measured and data gathered. Time T_2 represents an uninformative experimental choice since the behaviour of the two models is very similar, while data obtained at time T_1 is more likely to favour one model over another, since the distributions of simulated trajectories completely separate. More formally, the key steps in the method are as follow: Firstly we define the limited range of experimental options to be explored and encode them as parameterised extensions of the competing models. Secondly, the so called unscented transform (UT) [11] is used to approximate the prior predictive distribution as a mixture of Gaussians, for each model and a given experiment. Finally, optimisation is performed over the experiment parameters in order to best 'separate' the prior predictive distributions of the competing models. Parameters obtained by this optimisation represent an experiment whose generated data is predicted to maximise the differences in the subsequent marginal likelihood values of the models.

The contributions of this article are threefold; firstly, we extend a promising and computationally efficient experimental design framework for model selection to the stochastic setting, with non-Gaussian prior distributions; secondly, we utilise this efficiency to explore the robustness of model selection outcomes to experimental choices; and finally, we observe that experimental design can give rise to levels of confidence in selected models that may be misleading as a guide to their predictive power or correctness. The latter two points are undertaken via high-throughput *in-silico* analyses (at a scale completely beyond the Monte Carlo based approaches mentioned above) on families of gene regulatory cascade models and various existing models of the JAK STAT pathway.

Results

Identifying crosstalk connections between signalling pathways

We first illustrate the experimental design and model selection framework in the context of crosstalk identification. After observing how the choice of experiment can be crucial for a positive model selection outcomes, the example will be used to illustrate and explore the inconsistency of selection between misspecified models.

We consider pairs of regulatory cascades, each consisting of four transcription factors, modelled by ordinary differential equations of the form,

$$\frac{dx_j}{dt} = -k_{deg}x_j + \frac{k_j x_{j-1}^{n_j}}{K_j^{n_j} + x_{j-1}^{n_j}}$$

for $j=1,...,8$, where $k_{deg}=0.5$ is the rate at which protein x_j degrades, k_j represents the maximal rate of production of x_j, K_j is the amount of the transcription factor, x_{j-1}, needed for half the maximal response, and n_j is called the Hill-coefficient, and determines the steepness of the response. A range of crosstalk models are formed (Figure 2) by inserting additional regulatory links between $\{x_1,...,x_4\}$ and $\{x_5,...,x_8\}$ with the same kinetics as above. A single model is chosen as the 'true' biological system to which we perform experiments, and six others with equal prior probabilities are proposed as models of the true system – our task will be to identify the most suitable one.

An experiment is defined by the parameter $\phi=(s_1,s_5,\tau,T)$, where s_j denotes the strength of an external stimulus to the production of x_j, $j=1,5$ which is modelled as a term,

$$\frac{s_j}{t+0.1} \text{ for } j=1,\ t>0 \tag{1}$$

$$s_j \text{ for } j=5,\ t>\tau \tag{2}$$

$$0 \text{ otherwise} \tag{3}$$

added to the relevant ODE equations. The time delay between the two stimulus applications is given by τ, and T is the time at which a single measurement of the system (of species x_8 only) is taken. Prior distributions for the model parameters are set as Gaussian with means of 40 and covariances of 10 for both the k_i and K_i respectively, with the Hill coefficient fixed at 1.

The results of this round of experimental design are shown in the top left of Figure 3, where a good choice of ϕ is found to be (7.55,14.26,17.97,19.55), with a corresponding score of 31.5. From the figure, it can be seen that this experiment is predicted to distinguish some pairs of models better than others. In particular, the distribution of scores suggests that while the marginal likelihoods of most pairs of models are separated as desired, there is no power to discriminate between models M_2 and M_6, or models M_1 and M_5. Indeed, data obtained by performing the experiment upon our 'true' system, leads to posterior probabilities for each model with the same pattern.

As a sanity check, we first choose the true model from amongst the set of competing models (M_3), and as expected find that it is recovered by model selection with probability 1. However if the true model is not represented by $M_1,...,M_6$ (a far more realistic case) but instead the crosstalk model with a single connection from x_4 to x_8, then models M_2 and M_6 are found to have similar posterior probabilities of approximately 0.45. Likewise, M_1 and M_5 share a posterior probability of 0.045, while a clear difference exists between any other pair of models. To distinguish further between the pair of highest scoring models, a further round of experimental design was performed, with the resulting experiment and data providing strong evidence in favour of model M_2.

In an attempt to evaluate the added value of choosing ϕ rationally for this example, we calculate scores for a uniform sample of 1000 values of ϕ from the same range as explored above. The resulting score distribution shown in Figure 4a, peaks in the

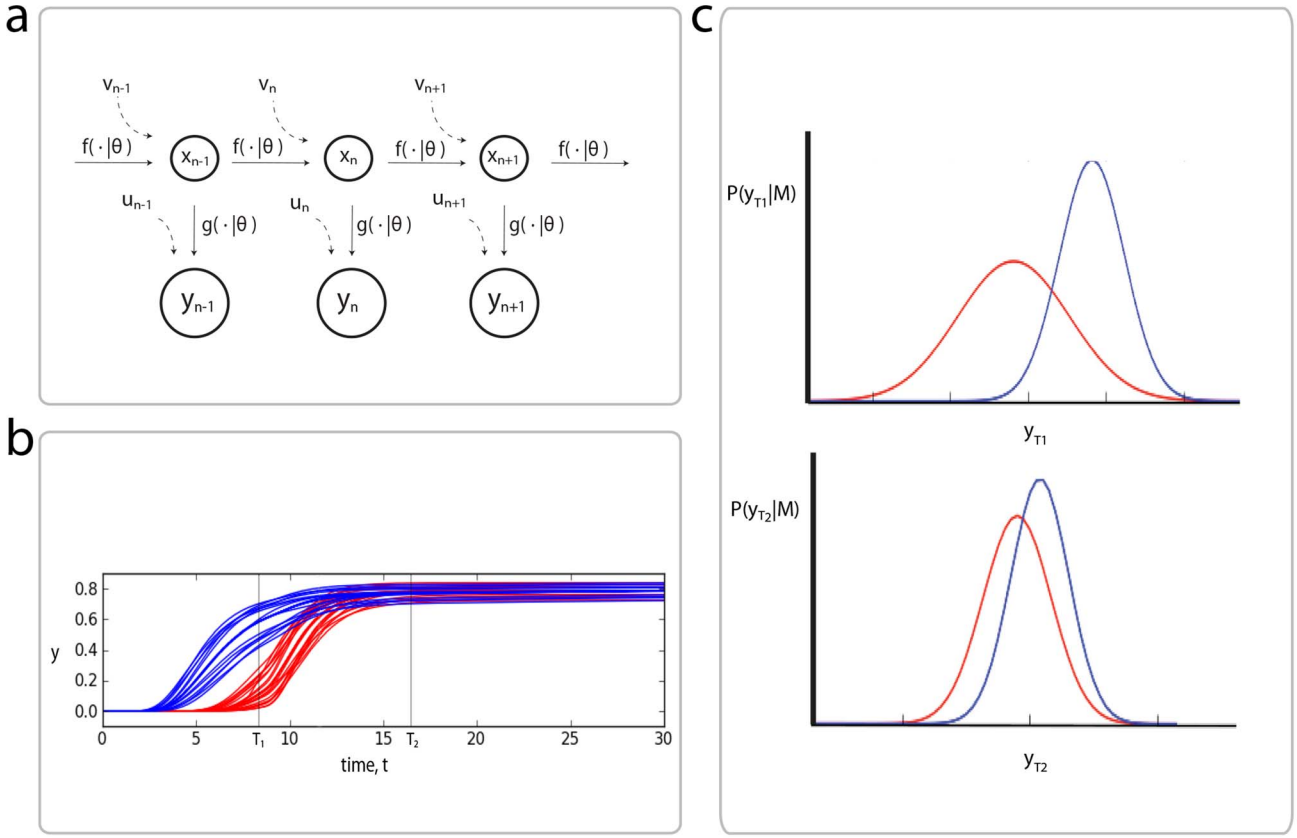

Figure 1. Outline of the proposed experimental design framework. a) We will be concerned with state-space formulations, which model a true state, x_n, as it evolves under the parametric function f subject to a process noise, v_n, and observations made of this process, y_n, via the 'observation' function, g, with measurement noise u_n. b) Plots of simulations from two different models (blue and red) for various parameter values, under the same experimental conditions. At time T_2, the behaviour of the two models is very similar, while at time T_1, the trajectories separate. c) Gaussian approximations of the model simulations at times T_1 and T_2 (in general these will be mixtures of Gaussians) obtained via the unscented transform. Time T_1 is likely to be more informative than time point T_2 for model selection purposes. Experiments can be scored by how separated these distributions are, which we quantify using the Hellinger distance.
doi:10.1371/journal.pcbi.1003650.g001

interval (15,16) which corresponds to an average Hellinger distance of <0.065 between the maximally separated marginal likelihoods of each pair of models. This is in contrast to the experiment found by our approach which lives in the tail of the distribution, with an average Hellinger distance of 0.74, and highlights how unlikely it is to find suitable experiments by chance alone. Experiments with even higher information content are found, which suggests that more care could be taken with the optimisation of ϕ, by for example, increasing the population size, or number of generations of the genetic algorithm used.

Perhaps unnervingly, the evidence in the first experiment is found to contradict (though not significantly in this case) the decision in favour of model M_2 over M_6, which is based on additional data from the second experiment. This suggests the possibility that the choice of experiment influences not only the amount of information available to select a particular model, but also the outcome of the model selection itself. Indeed the distribution of independently selected models from data generated by random experiments is surprisingly flat (Figure 4b). Even at very low levels of assumed noise, the most frequently selected model is chosen for less than half the experiments undertaken. This has been, to our knowledge, completely overlooked by the experimental design literature, but has important implications that we will explore further below.

The robustness of model selection to choice of experiment

To examine this last observation in more detail, we work with three of the crosstalk models described above, with connections between, (x_1,x_5), (x_1,x_6) and (x_4,x_8) respectively. The last of these is designated as the true model, and the others are considered as competing hypotheses about the location of the crosstalk connection. We perform 36100 experiments to collect data sets of size 1, 2, 4 and 8 equally spaced time points, each consisting of simulating the true model with different values of ϕ that correspond to changes in the delay between stimulus applications, and variation of the time at which the state of x_8 is first measured. An independent round of model selection is performed for each data set, and the posterior probabilities for each model are calculated.

The results for data sets of size 1 and 8 are illustrated in Figure 4c and 4d as heatmaps of posterior probabilities of the first model, and show that the vast majority of the space of experiments is split into distinct regions of high, low and equal probability for each model. In the case of a single time point, most of the explored experiment subspace is found to be uninformative, with the data providing equal support for each model. Three other distinct regions are identified, of which two show decisive support (on the

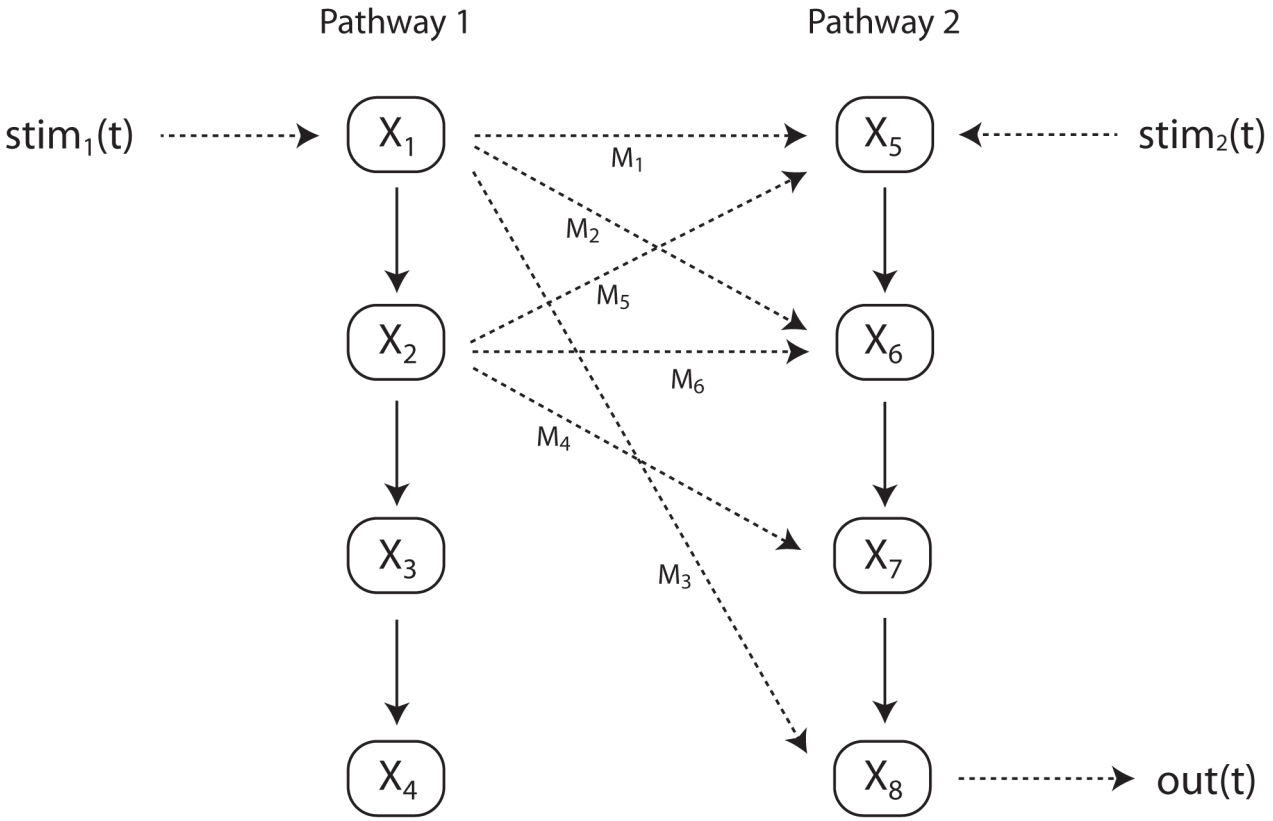

Figure 2. Crosstalk between regulatory cascades. Our task is to identify an unknown crosstalk connection between pathways 1 and 2. A limited range of experiments are considered, involving external stimulation of x_1 and x_5, and observation of x_8, and a set of models (M_1,...,M_6) corresponding to different crosstalk options are selected between. The times and strengths of the stimuli, and the time of measurement of x_8 are optimised to best distinguish between the competing crosstalk models.
doi:10.1371/journal.pcbi.1003650.g002

Jeffreys scale) for the first model, and one for which the second model is chosen decisively. In other words, by varying the experimental conditions an unequivocal choice (in isolation) for either model can be obtained. As more data points are considered, the uninformative region grows smaller, but regions of decisive support for each model remain. Interestingly, these regions are located in distinctly different places for single or multiple time points, although they remain similar for 2 or more time points. This reflects the added value of time series experiments – the marginal likelihoods now balance the ability of the models to reproduce each time point, with their ability to capture the autocorrelation of the time series.

In order to establish whether the observed inconsistencies are an artefact of the UT approximations, we perform a similar but necessarily course grained study using MultiNest [12,13], an implementation of nested sampling (a Monte Carlo based technique with convergence rate $O(n^{-\frac{1}{2}})$ [14]). Results obtained using MultiNest (shown in the upper right of figure 5) are almost identical to those of figure 4c, displaying the same regions of decisive support for each model. Given how difficult it is to estimate marginal likelihoods in general, the excellent performance of the UT (with only one Gaussian component) may seem rather surprising, until one notes that for the models and experiments considered, the prior predictive distributions are approximately Gaussian themselves (Figure 5). We discuss how the framework can deal with non-Gaussian effects, such as those found in the next examples, in the appendix.

JAK-STAT signalling

In this section we undertake an analysis of three mass action models of varying degrees of resolution of the JAK-STAT signalling pathway [15]. Each model describes the initial pathway activity after receptor activation (Figure 6), but before any feedback occurs. In brief, the signalling process consists of a receptor binding to JAK to form a complex that can dimerise in the presence of interferon-γ (IFN). This dimer is activated by phosphorylation by JAK, and in turn deactivated after being bound by tyrosine phosphatase (SHP_2). In its active state, the receptor complex phosphorylates cytoplasmic STAT1, which is then able to dimerise and act as a transcription factor [16].

We take the most detailed model, M_T, with 17 state variables and 25 parameters (published by Yamada et al. [16]), as our true system to which *in-silico* experiments can be performed, and select between two of the other models proposed by Quaiser et al. The first of these competing models, M_1, simplifies the true system, by neglecting a reaction – the re-association of phosphorylated STAT1 to the activated receptor – and thereby reducing the system to 16 states and 23 parameters. A series of five other 'biologically inspired' simplifications leads to our second model, M_2, which has 9 states and 10 parameters (these steps are summarised in Figure 6).

We set the parameter priors as a 10 component mixture of Gaussians fit to a uniform sample from the hypercube $[0,0.5]^d$, where $d\in\{10,23\}$ is the parameter dimension, such that all the parameter values inferred for each model by Quaiser et al. are supported. We define and undertake two classes of experiment

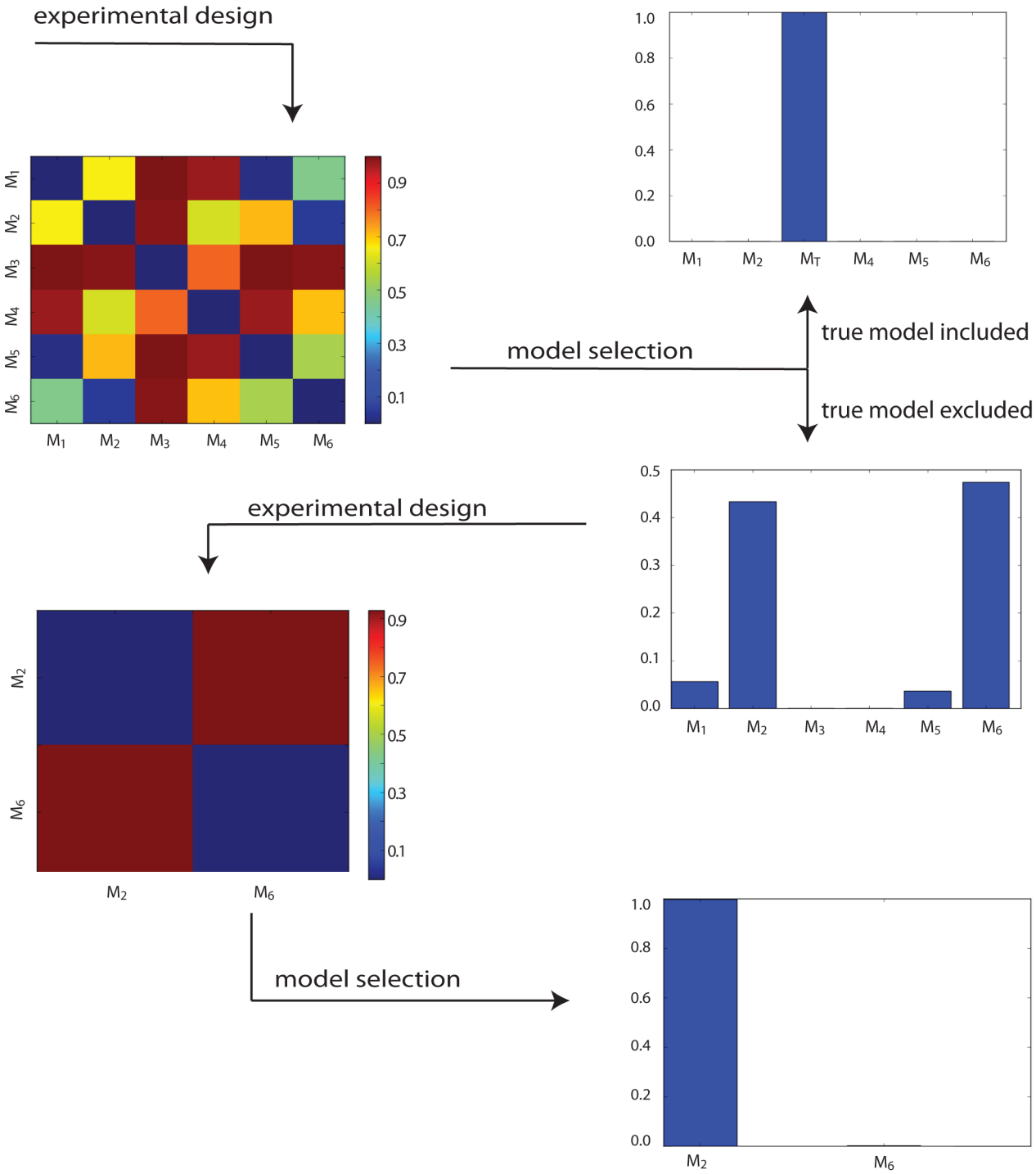

Figure 3. Flow diagram showing two rounds of experimental design and model selection. The heat maps on the left show the Hellinger distances between the prior predictive distributions of model pairs, for the chosen experiments. Bar plots on the right give the posterior probabilities of each model with respect to data produced by the chosen experiment. After the first experiment, models M_2 and M_6 have the most support, but evidence to choose between them is negligible. However a second experiment designed for only these two models (with priors set according to the posterior probability proportions after the first round of model selection) strongly favours model M_2.
doi:10.1371/journal.pcbi.1003650.g003

upon the true model (with parameters fixed to the published values); in the first, the IFN stimulus strength and the initial time point of a time series of 8 equally spaced measurements of the amount of JAK bound to the receptor are varied, and in the second, the species to be measured and the time at which this first measurement takes place are adjusted.

Model selection outcomes for each experiment (shown in Figure 7) show similar features to those for the crosstalk models,

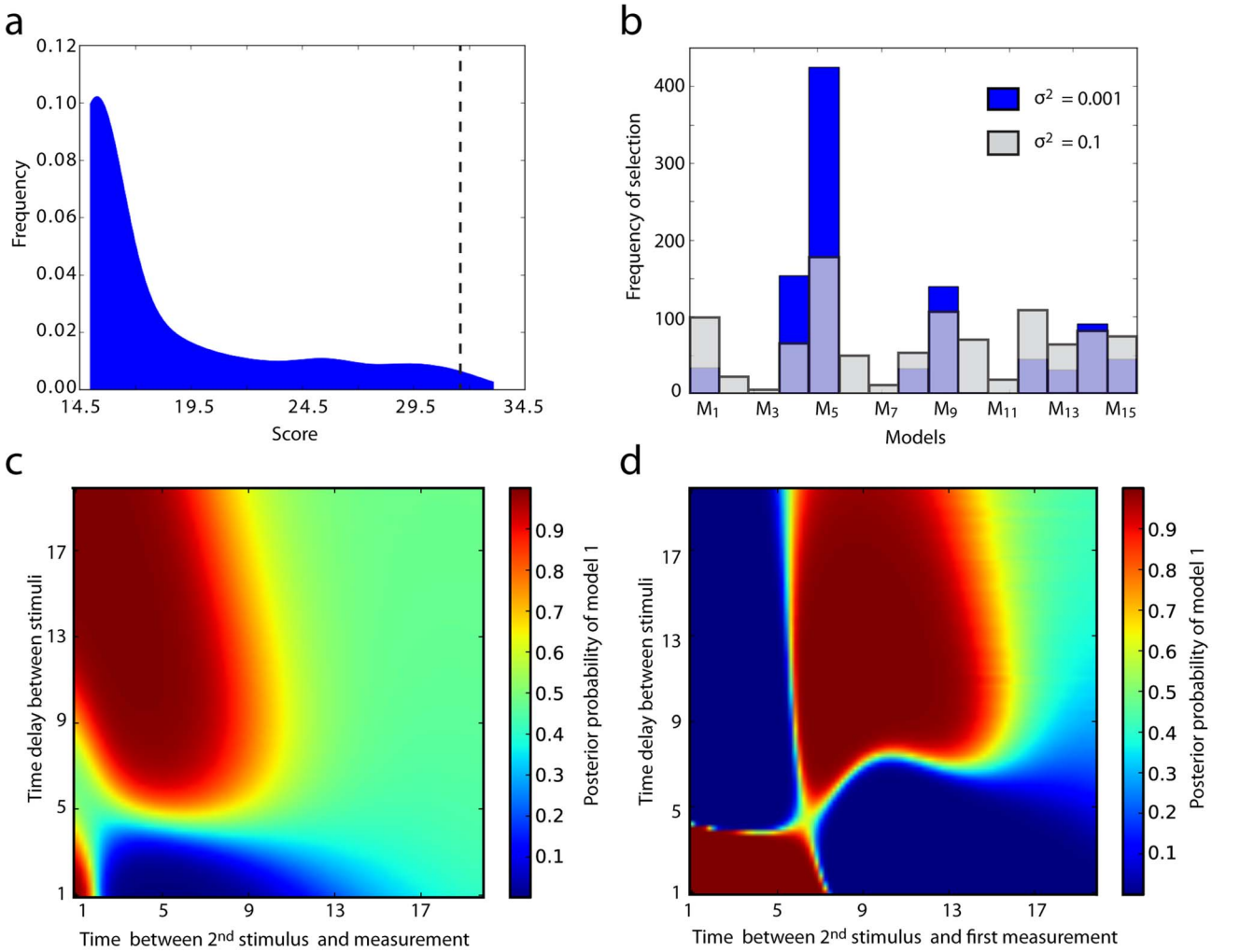

Figure 4. Robustness of model selection. a) Frequency distribution of scores for 1000 uniformly sampled values of ϑ. Scores concentrate around the interval (15,18), corresponding to very little information content. The dotted line indicates the score of ϑ^* chosen in the first round of experimental design. b) Using the 16 crosstalk models consisting of a single connection from pathway 1 to 2, a true model is fixed and 1000 uniformly sampled experiments are performed upon it. The frequencies at which the remaining 15 crosstalk models are selected, with each data set considered independently are shown. (blue) At a low level of measurement noise (with variance 0.01) model 5 is chosen most frequently, but is still outperformed for over half the experiments. (grey) When the measurement noise is increased to a variance of 0.1, the choice of model becomes even less robust. c, d) Each heatmap shows the posterior probabilities of model 1 (versus model 2), calculated independently for 9025 experiments, with data sets of different sizes (1 and 8 respectively). Each coordinate represents a different experiment, with variations to both the time delay between stimuli, and the measurement times.
doi:10.1371/journal.pcbi.1003650.g004

with distinct region of high posterior probability for each model. For the first class of experiments, selection between models M_1 and M_2 reveals strong support for the simpler model when data is gathered at earlier time points. The more complex model, M_1, is generally favoured for later time series, and also for a very limited range of IFN stimuli strengths at early time series. For the second class of experiments, the model selection outcome is found to depend strongly upon which species is measured. The simpler model is chosen decisively and almost independently of the measurement times considered when cytoplasmic phosphorylated STAT1, in monomeric or dimeric form, or two forms of the receptor complex (IFN_R_JAKPhos_2 and IFN_R_JAK) are measured. The same is true of the complex model for measurements of two other forms of the receptor complex (IFN_R_JAK2 and IFN_R_JAK-Phos_2_SHP_2). Otherwise the model selection outcome is time dependant or the choice of species is found to be uninformative.

Both these case studies make it clear that under the realistic assumption that all models are more or less incorrect, model selection outcomes can be sensitive to the choice of experiment. This observation has particular importance for studies that treat models as competing hypotheses that are decided between using experimental data; it is quite possible that if different experiments are undertaken, the conclusions drawn will also be different. In particular, the confidence calculated for such a conclusion (using the Jeffreys scale or another measure) can be misleading as a guide to how correct or predictive a model is (Figure 8a); in both the examples studied here, conditions exist such that any of the competing models can score a 'decisive' selection. The model selection outcome and associated confidence must therefore be strictly interpreted, as only increasing the odds of one model (with respect to others) for the data gathered under the specific experimental conditions.

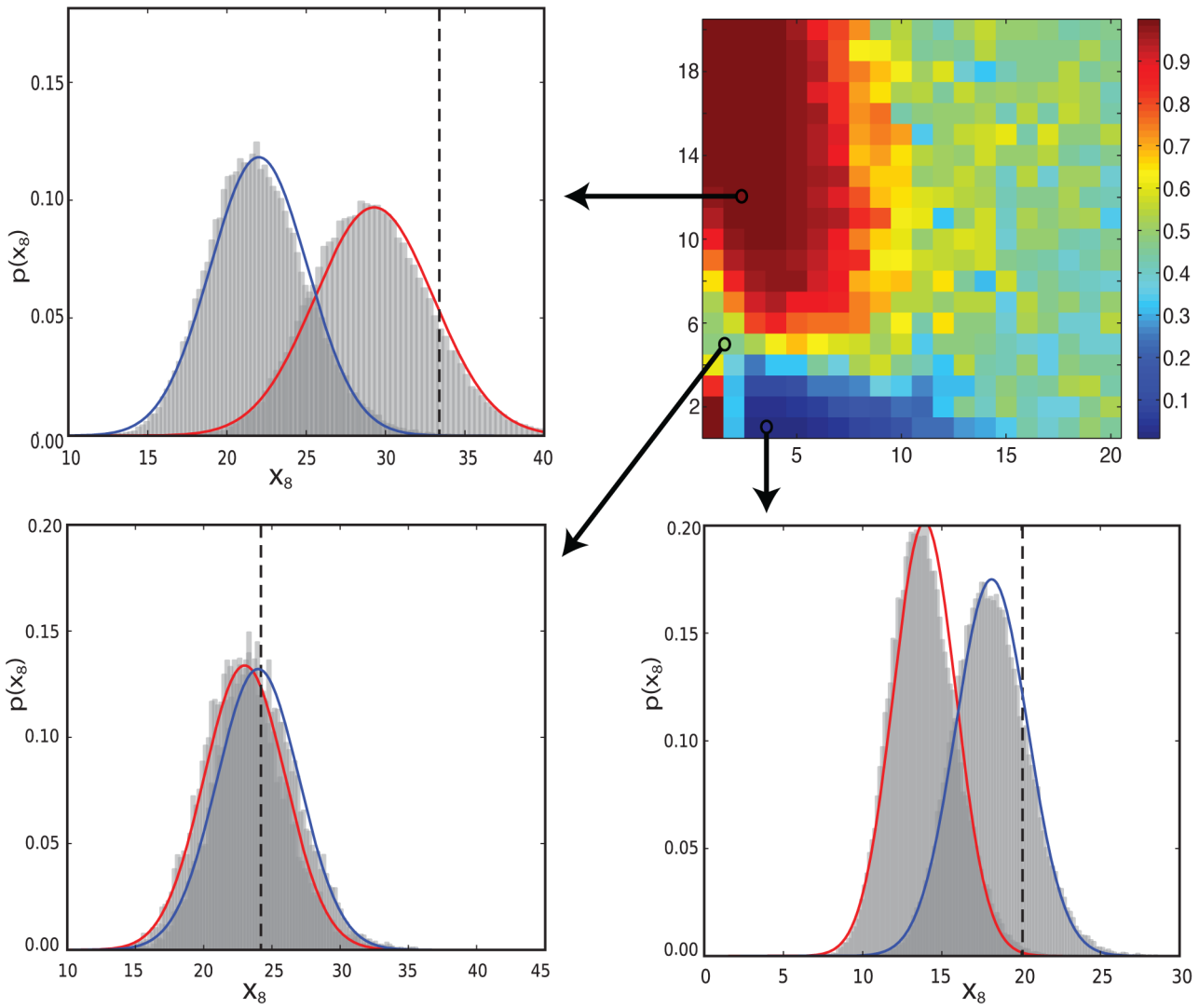

Figure 5. Monte Carlo validation. The top right plot shows posterior model probabilities obtained using MultiNest. The necessarily course grained results match those obtained by the UT in figure 4c. Each of the other plots compare UT approximations to the prior predictive distributions with Monte Carlo approximations using samples of size 10000, for different experimental conditions indicated by arrows. The red and blue lines correspond to UT approximations (using a single Gaussian component) for model 1 and 2 respectively. The dotted line indicates the data simulated from the true model.
doi:10.1371/journal.pcbi.1003650.g005

In light of this observation, the role of experimental design may need to be examined further. Since different models can be selected depending on the experiment undertaken, the use of experimental design will necessarily lead to choosing the model which, for some 'optimal' experiment, has the highest possible predicted level of confidence i.e. experimental design implicitly makes confidence a selection criterion. Is it misleading to claim high confidence in a model selection result when the models have been set up (by extensions to mimic the optimal experiment) for this purpose? Is a bias introduced into the inference via experiment design? In the context of experiment design for parameter estimation, MacKay suggests this is not a problem [17], stating that Bayesian inference depends only on the data collected, and not on other data that could have been gathered but was not. Our situation here is different since we consider changes not only to the data collection procedure, but also the data generation process and in turn the competing models themselves. It seems plausible that some models will gain or lose more flexibility than others with regards to fitting data for a particular choice of experiment. Even if the actual model selection is not biased, the confidence we associate with it will scale with the optimality of the experiment. After performing the optimal experiment, should there be any surprise that the selected model seems to have high support from the data? We feel these questions need further investigation.

Measuring sensitivity to model inaccuracies

In practical terms, the important question seems to be: how wrong does the model structure (or parameter values) have to be before the less predictive model (or that which captures less about the true system) is chosen? Clearly the answer is sensitive to the system and models under study, and moreover, the issue of how to compare the size of different structural inaccuracies is non trivial. Here, as a first attempt, we limit ourselves to considering the simple case of parameter inaccuracies in linear ODE models.

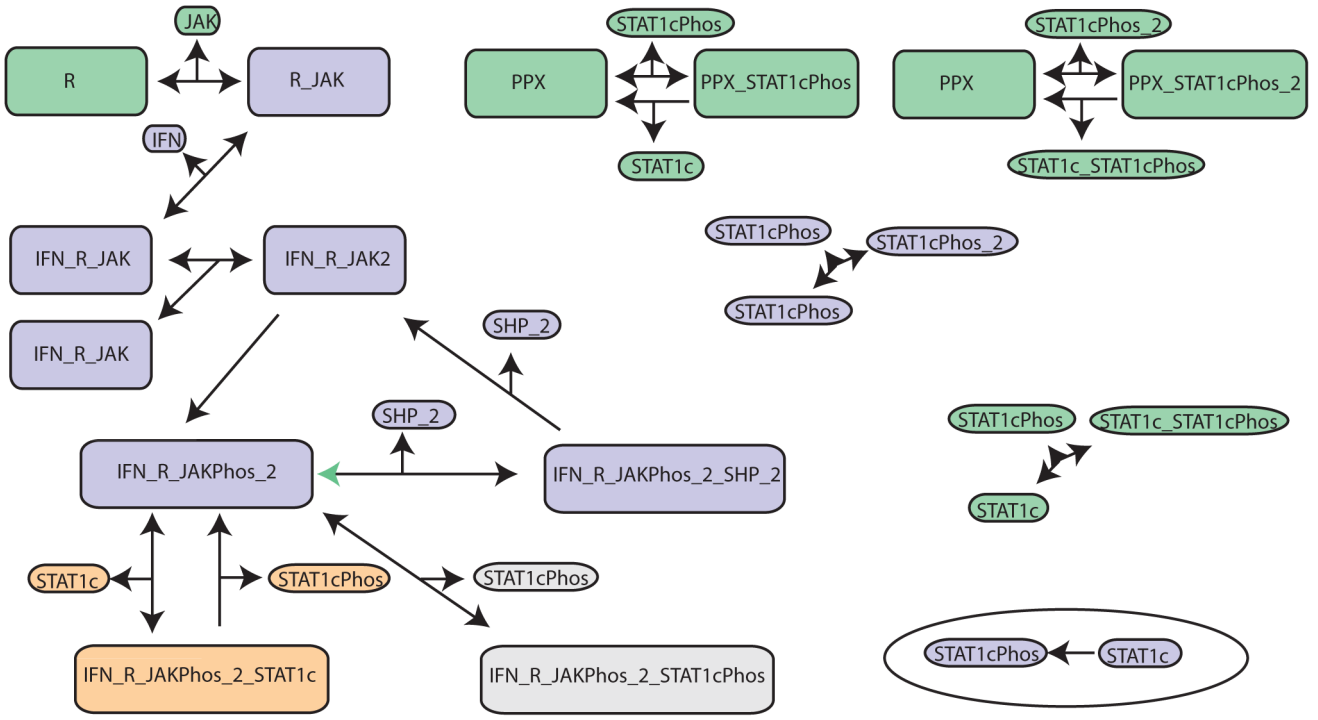

Figure 6. JAK STAT pathway models (adapted from Quaiser et al. [15]**).** Arrows indicate association or dissociation reactions between the protein species. Grey reactions only occur in the true model, (M_T). Model M_1 consists of the purple, orange and green components. Model M_2 is obtained by removing the green components, and replacing the orange reactions by the reaction in the bottom right oval.
doi:10.1371/journal.pcbi.1003650.g006

We define a 'base' model as the linear ode system defined by its Jacobian matrix with entries,

$$\begin{pmatrix} b_1 & b_2 \\ b_3 & b_4 \end{pmatrix}$$

and 'extensions' to this model as an extra row and column,

$$\begin{pmatrix} b_1 & b_2 & e_1 \\ b_3 & b_4 & e_2 \\ e_3 & e_4 & e_5 \end{pmatrix}$$

Biologically such an extension may represent the inclusion of an extra molecular species into the model, along with rules for how it interacts with components of the original system. Defining true base and extension models by $(b_1,b_2,b_3,b_4)=(-1.0,2.0,0.5,-4.0)$ and $(e_1,e_2,e_3,e_4,e_5)=(0,1,0,1,0)$, we consider two models,

$$\begin{pmatrix} b_1+p_1 & b_2 & \theta_{13}h_{13} \\ b_3+p_2 & b_4 & \theta_{23}h_{23} \\ \theta_{31}h_{31} & \theta_{32}h_{32} & \theta_{33}h_{33} \end{pmatrix}$$

and

$$\begin{pmatrix} b_1+p_1 & b_2 & \theta_{13}h'_{13} \\ b_3+p_2 & b_4 & \theta_{23}h'_{23} \\ \theta_{31}h'_{31} & \theta_{32}h'_{32} & \theta_{33}h'_{33} \end{pmatrix}$$

where $(h_{13},h_{23},h_{31},h_{32},h_{33})=(e_1,e_2,e_3,e_4,e_5)$ and $(h'_{13},h'_{23},h'_{31},h'_{32},h'_{33})=(1,0,1,0,1)$, are competing (true and false) hypotheses about the structure of the model extension, with a zero h'_{kj} or h_{kj} indicating a belief that species k does not directly affect the rate of increase of species j. Parameters θ_{jk}, are the unknown strengths of these interactions, over which we place a 50 component mixture of Gaussians prior, fit to a uniform distribution over the interval $[-5,5]$ for each parameter. We represent inaccuracies in modelling the base as additive perturbations p_1 and p_2. Data was generated by simulating the state of the first variable of the true model at times $t=0.1,0.2,0.3,0.4,0.5$, for initial condition $(1.0,1.0,1.0)$.

Model selection outcomes for 40,000 different pairs of values for the perturbations (p_1,p_2), are shown in Figure 9. Distinct regions for each possible outcome are found and colour coded in the figure, with red indicating that the true extension has been identified successfully, yellow representing a decision in favour of the false extension, orange that evidence for either model is not substantial on the Jeffreys scale, and finally blue indicating that the marginal likelihood for both models is found to be less than 10^{-10}, for which any conclusion would be subject to numerical error. Increasing this threshold has the effect of replacing red areas with blue.

In the majority of cases tested, the true extension is correctly identified despite inaccuracies in the base model. However, a set of perturbations are seen to confound the selection, and allow the false extension to obtain substantial support. Furthermore, the selection outcome is found to be more sensitive in some directions than others, with relatively small perturbations to base model entry (1,1) causing a change in outcome and creating decision boundaries near the lines $x=0$ and $x=4$. Prior to our analysis,

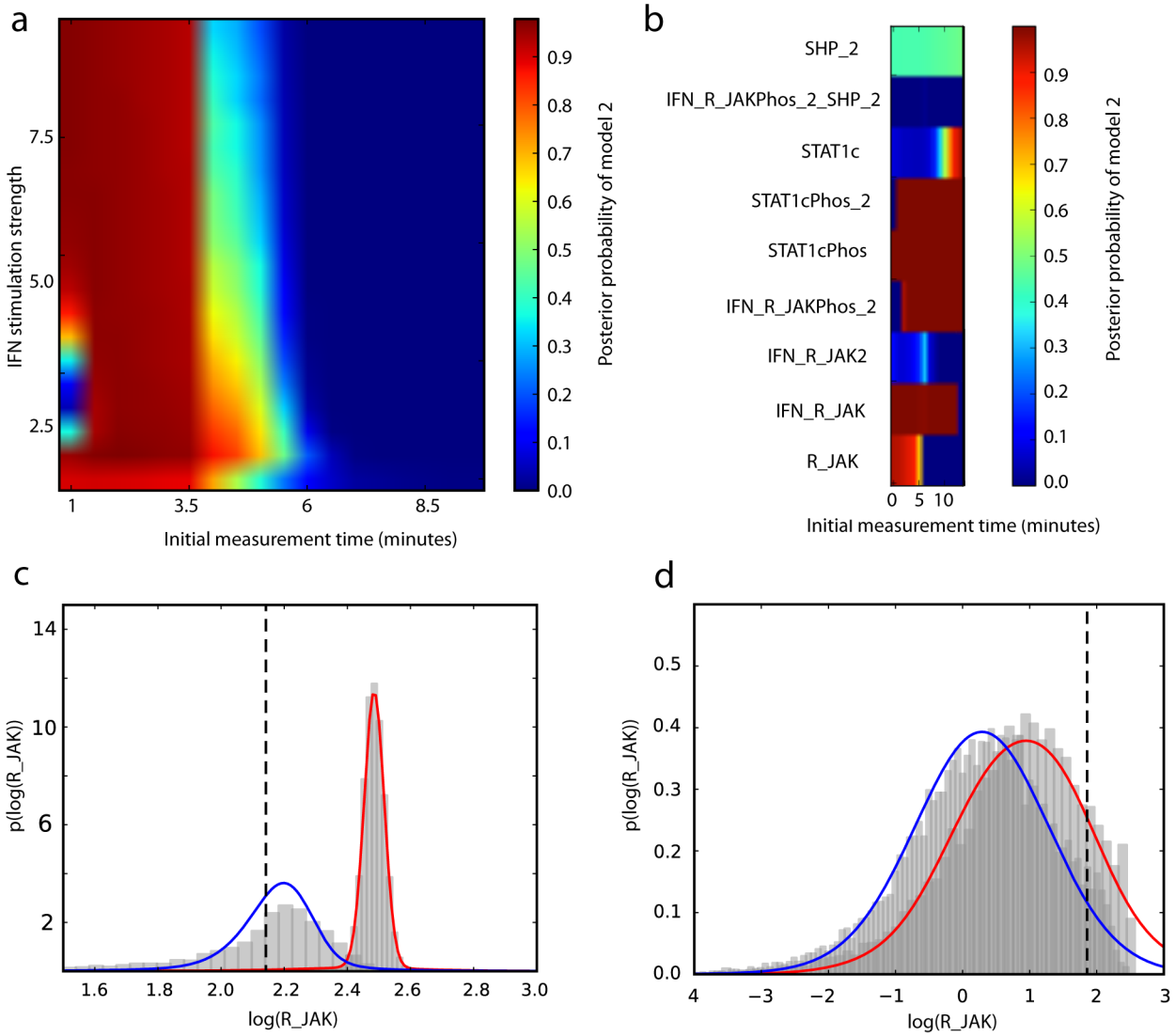

Figure 7. JAK STAT model selection sensitivity. a) IFN stimulus strength and the initial measurement time are varied. b) The species to be measured and the time at which this initial measurement takes place are adjusted. In both figures, distinct regions of high probability for each model can be seen. Comparison of UT and Monte Carlo approximations to the prior predictive distributions when c) M_1 is chosen and d) M_2 chosen for IFN stimulus strengths 1 and 0.6 at times 60 and 1 minute respectively. The 10 components used in the mixture distribution allow non-Gaussian effects to be captured. The error in the UT approximations is significantly smaller than the differences between models.
doi:10.1371/journal.pcbi.1003650.g007

it would be hard to predict these observations even when the true model is known and as simple as that explored here.

In real applications, where the true model is unknown and more complex, it may not be possible to tell whether a conclusion is an artefact of model inaccuracies, even when the truth of the conclusion itself can be tested by direct experimental measurement. However, the type of analysis undertaken here at least gives a measure of robustness for the conclusion to a range of model inaccuracies. Unfortunately, this remains difficult to implement in a more general setting – for example, in climatology, where the accepted method of coping with structural uncertainty is through the use of large ensembles of similar models produced by various research groups [18], a luxury that cannot be afforded on the scale of the most ambitious systems biology projects. While the practical challenges of dealing with large numbers of models is somewhat overcome by the model selection algorithm described above, a harder conceptual problem exists of how to define perturbations to more complicated classes of model, and to compare their strengths.

Finally, the example also highlights the difficulty of testing a hypothesis that represents only part of a model. The study shows that the implicit assumption that the base model is accurate, is not necessarily benign, and can affect any conclusions drawn – a result that is borne out by the logical principle that from a false statement, anything is provable.

Discussion

The scale of the analyses detailed above, comprising thousands of marginal likelihood computations, requires extreme computational efficiency. Indeed it is completely beyond Monte Carlo based methods such as that recently developed by Liepe et al. [2], which are limited to exploring small sets of models and experiments. Here, the efficiency was obtained by using the

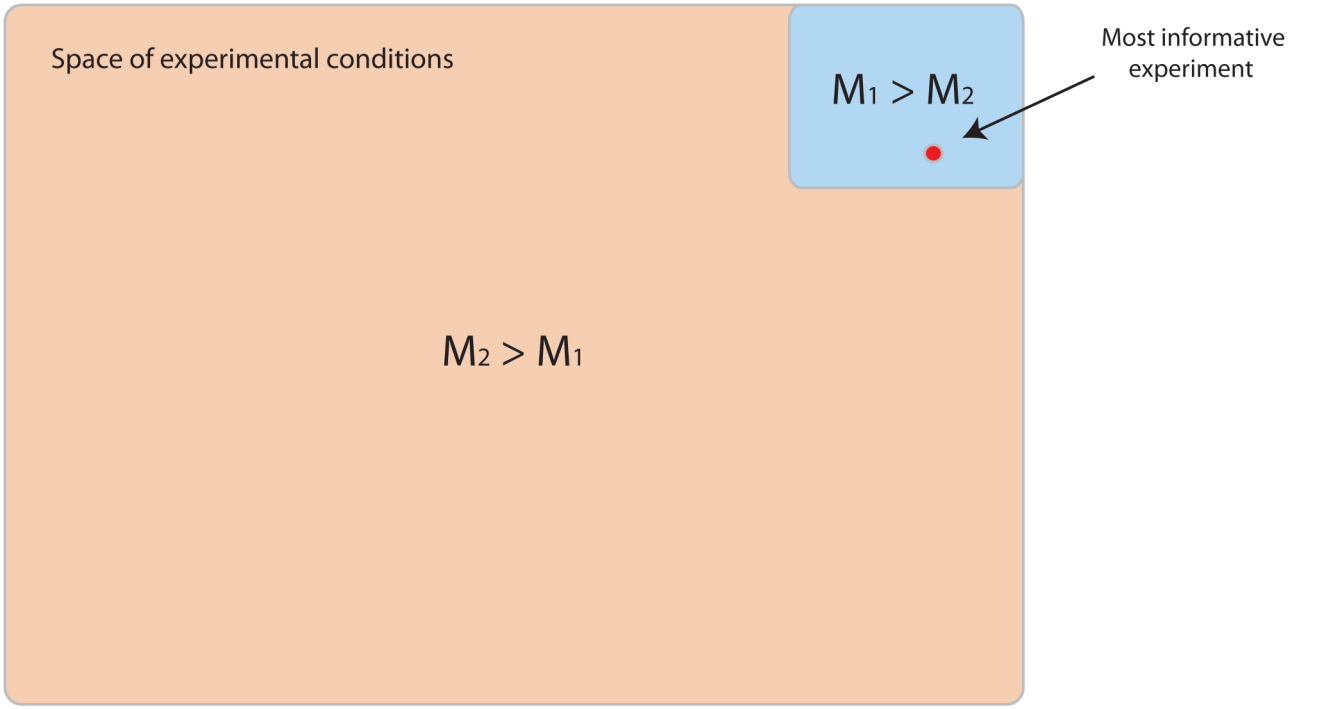

Figure 8. Model predictive power v.s. predicted confidence. Model M_1 explains data produced from experiments in the blue region better than model M_2. The opposite is true for the larger orange region. In this example, the most informative experiment generates data that favours model M_1. Performing model selection using such data will lead to the highest possible confidence we can generate for either model, and yet the chosen model will be the least predictive i.e. M_2 reflects reality better for the majority of considered experimental conditions. In this particular case, we have a greater chance of choosing the most predictive model by performing a random experiment.
doi:10.1371/journal.pcbi.1003650.g008

unscented transform for propagating Gaussian mixture distributions through non-linear functions. Further computational savings can be made by exploiting the highly parallelizable nature of Flassig and Sundmacher's method [9], which we have extended for use with mixture distributed priors and stochastic state space models.

This efficiency has allowed us to explore model selection problems involving relatively large numbers of models and experiments, and investigate the robustness of model selection results to both changes in experimental conditions and inaccuracies in the models. Results from the latter two studies illustrate some common, but often ignored, pitfalls associated with modelling and inference. Firstly, we show that the conclusions of a model selection analysis can change depending on the experiment undertaken. Related to this, we observe that confidence in such a conclusion is not a good estimator of the predictive power of a model, or the correctness of the model structure. Further we note that the use of experimental design in this context maximises the expected discriminatory information available, and implicitly makes confidence in the outcome a criterion for model selection. In the future we intend to investigate the desirability of this property and how it affects the interpretation of the confidence associated with model selection outcomes.

At the heart of these issues is a lack of understanding of the implications of model (or parameter) inaccuracies. Often improved fits to data or better model predictions are interpreted as evidence that more about the true system is being captured. This assumption underlines a guiding paradigm of systems biology [19], where a modelling project is ideally meant to be a cycle of model prediction, experimental testing and subsequent data inspired model/parameter improvement. However, it is possible that improved data fitting and predictive power (although desirable in their own right) can be achieved by including more inaccuracies in the model. In the context of parameter estimation, this concept of local optima is widely known, and their avoidance is a challenge when performing any non-trivial inference. One simple method to do so is to include random perturbations in the inference, in order to 'kick' the search out of a local optimum. Perhaps a similar strategy might be included in the modelling paradigm; by performing random experiments, or adding or removing interactions in a model structure, data might be gathered or hypotheses generated that allows a leap to be made to a more optimal solution.

While we have been concerned solely with the statistical setting, it is reasonable to expect similar results can be found for alternative model discrimination approaches e.g the use of Semidefinite programming to establish lower bounds on the discrepancy between candidate models and data [20]. Here the particular subset of models that are invalidated will be dependent upon the experiment undertaken. However, emphasis on invalidating wrong models instead of evaluating the relative support for each at least reduces the temptation for extrapolated and, perhaps, false conclusions.

George E. P. Box famously stated that 'Essentially, all models are wrong, but some are useful'. Here we would add that if nothing else, models provide a natural setting for mathematicians, engineers and physicists to explore biological problems, exercise their own intuitions, apply theoretical techniques, and ultimately generate novel hypotheses. Whether the hypotheses are correct or not, the necessary experimental checking will reveal more about the biology.

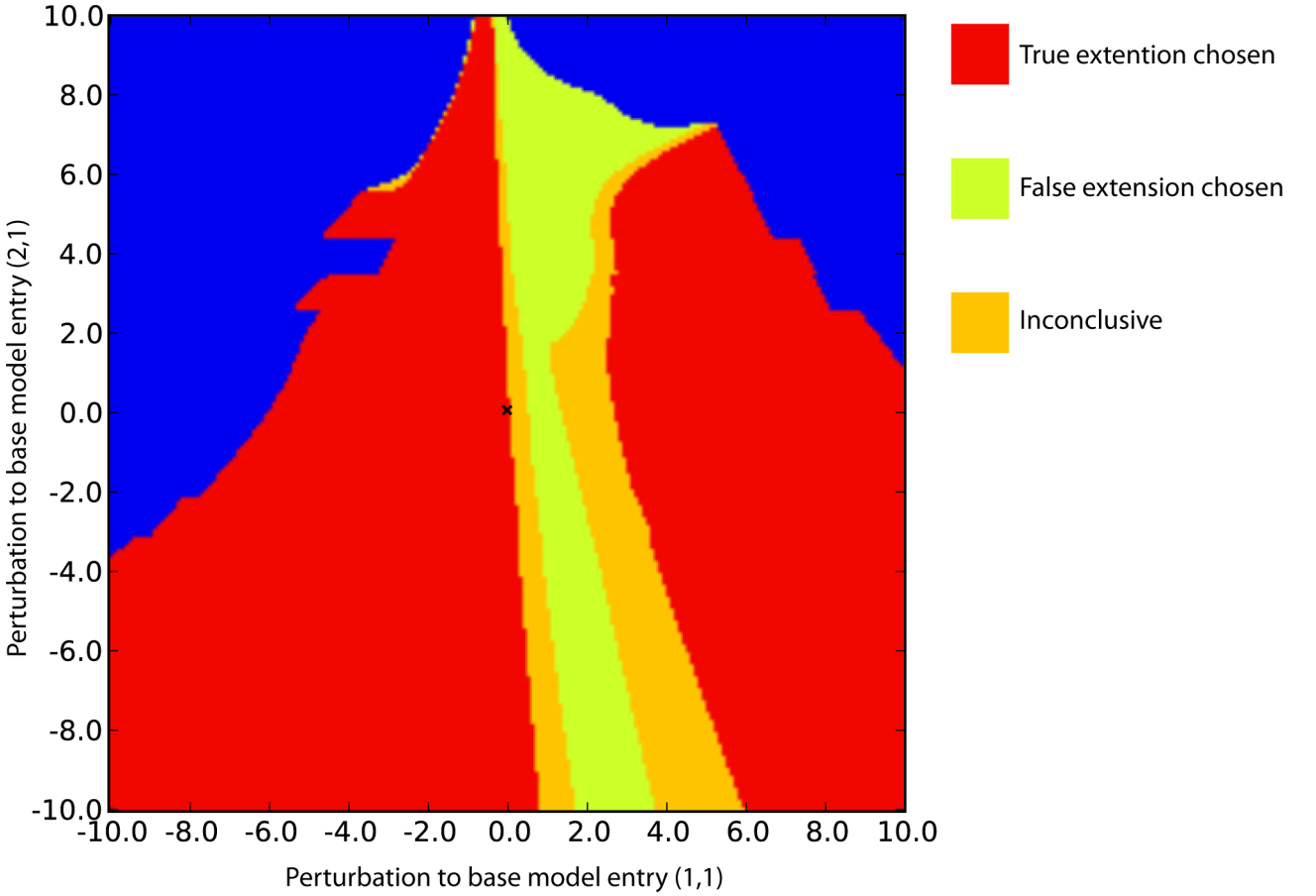

Figure 9. Model selection outcomes for 40,000**, different pairs of linear ode models.** Each model represents one of two competing hypotheses (the model extension), but with a different base model generated by perturbing Jacobian matrix entries (1,1) (x-axis) and (2,1) (y-axis). Regions where the different hypotheses receive support are given by the red (true extension), yellow (false extension), orange (no significant support for either extension), and blue (marginal likelihood values for both models are $<10^{-10}$) coloured regions. Increasing the threshold for the blue region to 10^{-2} results in reduction of the red region, but not of the yellow. Using the true base model (represented by the cross at (0,0)), the true extension is also identified.
doi:10.1371/journal.pcbi.1003650.g009

Materials and Methods

The unscented transform

The UT is a method that describes how the moments of a random variable, θ, are transformed by a non-linear function, g. The algorithm begins by calculating a set of weighted particles (called sigma-points) with the same sample moments up to a desired order as the distribution $p(\theta)$. For the results shown here, we use a scaled sigma-point set $\{\chi_k\}_{k=0,\ldots 2L}$ that captures both means and covariances [21],

$$\chi_0 = \mu_\theta$$
$$\chi_k = \mu_\theta + \left[\sqrt{(L+\lambda)\Sigma_\theta}\right]_k \quad k=1,\ldots,L$$
$$\chi_k = \mu_\theta - \left[\sqrt{(L+\lambda)\Sigma_\theta}\right]_k \quad k=L+1,\ldots,2L$$

where L is the dimension of θ, μ_θ and Σ_θ are the mean and covariance of $\theta \sim p(\cdot)$, $[A]_k$ represents the kth column of a matrix A, and

$$\lambda = \alpha^2(L+\kappa) - L.$$

The sigma-point weights $\{v_k^c, v_k^m\}_{k=0,\ldots 2L}$ are given by,

$$v_0^m = \frac{\lambda}{L+\lambda}$$
$$v_0^c = \frac{\lambda}{L+\lambda} + (1-\alpha^2+\beta)$$
$$v_k^m = v_k^c = \frac{1}{2(L+\lambda)} \quad k=1,\ldots,2L.$$

and finally, the parameters κ, α and β may be chosen to control the positive definiteness of covariance matrices, spread of the sigma-points, and error in the kurtosis respectively. For the results in this article we take $\kappa = 0$ as is standard in the literature [22], and $\beta = 2$ which is optimal for Gaussian input distributions, while α, controlling the spread of sigma-points is taken small as 10^{-2} to

avoid straddling non-local non-linear effects with a single Gaussian component [21].

The mean and covariance of the variable $g(\theta)$, can be estimated as the weighted mean and covariance of the propagated sigma-points,

$$\mu_{g(\theta)} \approx \sum_{k=0}^{2L} v_k^m g(\chi_k) \tag{4}$$

$$\Sigma_{g(\theta)} \approx \sum_{k=0}^{2L} v_k^c (g(\chi_k) - \mu_{g(\theta)})(g(\chi_k) - \mu_{g(\theta)})^T. \tag{5}$$

We denote the resulting approximate probability density function for $g(\theta)$, by $\mathcal{U}_{p(\theta)}(x)$.

By matching terms in the Taylor expansions of the estimated and true values of these moments, it can be shown that the UT is accurate to second order in the expansion. More generally, if the sigma-point set approximates the moments of θ up to the n^{th} order then the estimates of the mean and covariance of $g(\theta)$ will be accurate up to the n^{th} term [11]. Crucially, the number of points required ($2L+1$ for this scheme) is much smaller than the number required to reach convergence with Monte-Carlo methods.

Unscented model selection

We will consider discrete time state space models, $\mathcal{M}$, with state–transition (f) and observation (g) functions both parametrized by θ,

$$x_n = f(x_{n-1}|\theta, v_n) \tag{6}$$

$$y_n = g(x_n|\theta, u_n) \tag{7}$$

where $\hat{y} = (y_{t_0}, ..., y_{t_n}, ..., y_T)$, is the time series of M dimensional measurements that we are trying to model, x_n is the N dimensional true state of the system at time t_n, and u_n, and v_n are independent, but not necessarily additive, Gaussian white-noise process and measurement terms. Bayesian model selection compares competing models, $\{\mathcal{M}_i\}$, by combining the *a priori* belief in each model, encoded by the model prior distribution $p(\mathcal{M}_i)$, with the evidence for each model in the data $\hat{y}$, as quantified by the marginal likelihood,

$$p(\hat{y}|\mathcal{M}_i) = \int p(\hat{y}_\phi|\theta_i, \mathcal{M}_i) p(\theta_i|\mathcal{M}_i) d\theta,$$

where $p(\theta_i|\mathcal{M}_i)$ is the parameter prior for model $\mathcal{M}_i$. In the Bayesian setting, the relative suitabilities of a pair of models $(\mathcal{M}_1, \mathcal{M}_2)$ are often compared using the ratio of posterior probabilities, known as the Bayes factor,

$$B_{12} = \frac{P(\hat{y}|\mathcal{M}_1)}{P(\hat{y}|\mathcal{M}_2)},$$

with a Bayes factor of $\frac{1}{3} > B_{12} > 3$ seen as substantial [23]. However, for complex or stochastic models, the marginal likelihood can be intractable, and so approximate likelihood free methods, such as Approximate Bayesian Computation are becoming increasingly important and popular within the biosciences [24]. A big drawback of such Monte-Carlo based algorithms is the large number of simulations – and associated computational cost – required to estimate the posterior distributions or Bayes factors. Even with GPU implementation [25], applications are currently still limited to comparing pairs or handfuls of models.

In order to address the issues raised above, a higher-throughput model selection algorithm is needed. Our approach will be to fit mixture of Gaussian models to the prior parameter distribution for each model,

$$p(\theta|\mathcal{M}) \approx \sum_i \alpha_i p_i(\theta|\mathcal{M}),$$

so that we can exploit the UT within the state-space framework to drastically reduce the number of simulations necessary to estimate the distribution of the output of the model. Gaussian mixture measurement and process noise can also be considered, as in the work on Gaussian sum filters [26,27], although the number of mixture components required to model the output at each time point then increases exponentially, and in the case of long time series, component reduction schemes need to be implemented.

With this approximation, the marginal likelihood may be expressed as the sum,

$$p(\hat{y}|\mathcal{M}) \approx \int p(\hat{y}|\theta, \mathcal{M}) \sum_i \alpha_i p_i(\theta|\mathcal{M}) d\theta \tag{8}$$

$$= \sum_i \alpha_i \int p(\hat{y}|\theta, \mathcal{M}) p_i(\theta|\mathcal{M}) d\theta \tag{9}$$

$$\approx \sum_i \alpha_i \mathcal{U}_{p_i}(\hat{y}), \tag{10}$$

where the components, $\mathcal{U}_{p_i}(\hat{y})$, can be determined using the UT as described below. Note that the accuracy of the approximation can be controlled by the number of components used. However, in the presence of nonlinearities, choosing the number and position of components solely to fit the prior distribution may not be adequate. This is because we need to have enough flexibility to also fit a complex and possibly multi-modal output. Indeed, except at the asymptotic limit of dense coverage by the mixture components, it is possible to construct badly behaved mappings that will lead to loss of performance. For the applications visited in this article, the models proved well behaved enough such that a single component and 10 components respectively for the crosstalk and JAK-STAT systems sufficed for sufficient agreement with the nested sampling and Monte Carlo results. An improvement to the method described here would be to update the number of components automatically with respect to the model behaviour in a manner similar to how Gaussian mixtures can be adaptively chosen in particle based simulation of Liouville-type equations [28,29].

For the deterministic case including the examples considered in this article, we have $v_t = 0$, and the state–space model simplifies to,

$$\hat{y} = g(\theta) + u,$$

where might represent the simulation of certain variables of a system of ODEs, parameterised by θ, with additive measurement

error u. In this case the marginal likelihood can then be expressed as,

$$p(\hat{y}|\mathcal{M}) \approx \sum_i \alpha_i \mathcal{U}_{p_i}(\hat{y})$$

where each component $\mathcal{U}_{p_i}(\hat{y})$ is obtained simply through application of the UT with input distribution $p_i(\theta)$, and liklihood that is Gaussian with mean, $g(\theta)$, and variance, $\Sigma(u)$.

To estimate the marginal likelihood in the stochastic case ($v_t \neq 0$), we assume the observation function takes the form of a linear transformation of the true state and measurement noise at time n with additive noise,

$$g(x_n|\theta,u_n) = G(\theta)x_n + u_n \tag{11}$$

where $G(\theta)$ is an $N \times M$ matrix. In practice this might correspond to the common situation where observations are scaled measurements of the abundance of various homo- or heterogeneous groups of molecules.

We may then write the mean of the observation, y_n, in terms of the statistics of x_n,

$$\bar{y}_n = G(\theta)\bar{x}_n \tag{12}$$

for any n, and from the bilinearity of the covariance function, the covariance between any pair of observations, (y_n,y_m), as,

$$\begin{aligned}\Sigma(y_n,y_m) = G\Sigma(x_n,x_m)G^T + G\Sigma(x_n,u_m) \\ + \Sigma(u_m,x_n)G^T + \Sigma(u_n,u_m)\end{aligned} \tag{13}$$

$$= G\Sigma(x_n,x_m)G^T, \tag{14}$$

since x_n is independent of u_m for all n and m. We now need to find expressions for the process state covariance terms in equation 14. To do so we apply the UT iteratively for $n=0,\ldots,N-2$ to transform the state-variable, x_n through the state-transition function $f(x_n|\theta,v_n)$, with input distribution $p(x_n)$ given by,

$$p(x_n) \sim \begin{cases} p_i(x_n) & \text{if } n=0 \\ \mathcal{U}_{p(x_{n-1})}(x_n) & \text{if } n>0 \end{cases}$$

The result is a Gaussian approximation to the joint distribution $p_i(x_n,x_{n+1})$ for each n, and hence also to the conditional distributions $p_i(x_{n+1}|x_n)$. Given that x_n is a Markov process and that the product of Gaussian functions is Gaussian, we also have a Gaussian expression for the joint distribution, $p_i(x_0,\ldots,x_{N-1})$,

$$p_i(x_0,\ldots,x_{N-1}) = \prod_{n=0}^{N-1} p_i(x_{n+1}|x_n).$$

The covariance between any pair of observations y_n and y_m, may then be found by substituting relevant entries from the covariance matrix of the density of Equation into Equation 14. The subsequent Gaussian approximation to the joint distribution of y, given $p_i(x)$, constitutes one component in the mixture approximation of the marginal likelihood given in Equation 10.

Experimental design

We first introduce a vector of experiment parameters, ϕ, that describes how the dataset is created, specifying, for example, the times at which the system is stimulated, the strengths and targets of the stimuli, knockouts or knockdowns, along with the choice of observable to be measured at each time point. We can then model the system and experiments jointly, extending the f to include terms describing the possible experimental perturbations, and the g to capture the measurement options,

$$x_n = f(x_{n-1}|\theta,\phi,v_n) \tag{15}$$

$$y_n = g(x_n|\theta,\phi,u_n) \tag{16}$$

We assume that there is overlap between the system observables appearing in each model so that experiments that allow model comparison can be designed.

To illustrate how this might be done in practice, we consider a typical set of ordinary differential equations used to describe a gene regulatory mechanism,

$$\frac{dm}{dt} = -\beta_1 m + \frac{\alpha}{1+q} + \alpha_0, \tag{17}$$

$$\frac{dp}{dt} = -\beta_2 p + km, \tag{18}$$

where $\theta=(k,\beta_1,\beta_2,\alpha,\alpha_0)$ are the parameters controlling the rates of production and degradation of an mRNA, m, and a protein, p, subject to the concentration of a repressor protein, q. We define the state transition function f_i as their solution evaluated at the next measurement time-point $t_n(\phi)$ which is now dependant on the choice of ϕ, given the state at time $t_{n-1}(\phi)$, and subject to some additive noise v_n. These equations have be extended as,

$$\frac{dm}{dt} = \delta_k(\phi)[-\beta_1(\phi)m + \frac{\alpha(\phi)}{1+q} + \alpha_0] + s_m(\phi,t), \tag{19}$$

$$\frac{dp}{dt} = -\beta_2(\phi)p + k(\phi)m + s_p(\phi,t), \tag{20}$$

to model a range of possible experimental perturbations, e.g. setting $\delta_k(\phi)=0$ mimics a knockout of the gene producing mRNA m_k, and $s_x(\phi,t)$ an input stimulus to species x. The observation function g, as before can be some linear function of the states, however, the selection of variables and coefficients is now an experimental choice specified by ϕ

,

$$g(x_n|\theta,\phi,u_n) = G(\theta,\phi)x_n + u_n$$

Experimental design as an optimisation problem

Given a particular set of experimental options, ϕ, the marginal likelihood of model $\mathcal{M}$ *for any possible* data set $\hat{y}$ (the prior predictive distribution) can be estimated efficiently from equation 10,

$$p(\hat{y}|\mathcal{M},\phi)=\sum_i \alpha_i \mathcal{U}_{p_i}(\hat{y}),$$

with the components $\mathcal{U}_{p_i}$ calculated with respect to the extended system and experiment model. Comparisons between such prior predictive distributions for competing models provides a means to predict the discriminatory value of a proposed experiment. Intuitively, values of ϕ, for which the prior predictive distributions of two models are separated, correspond to experimental conditions under which the models make distinct predictions of the system behaviour. Data gathered under these conditions are thus more likely to yield a significant model selection outcome. More formally, we can quantify the value of an experiment ϕ, using the Hellinger distance between the prior predictive distributions,

$$H(P,Q)=\frac{1}{2}\int\left(\sqrt{P(x)}-\sqrt{Q(x)}\right)^2 dx$$

which takes the following closed form for multivariate Gaussian distributions, $P\sim N(\mu_P,\Sigma_P)$ and $Q\sim N(\mu_Q,\Sigma_Q)$,

$$H(P,Q)=1-\frac{|\Sigma_P|^{1/4}|\Sigma_Q|^{1/4}}{|\bar{\Sigma}|^{1/2}}e^{-\frac{1}{8}(\mu_P-\mu_Q)^T\bar{\Sigma}^{-1}(\mu_P-\mu_Q)},$$

where,

$$\bar{\Sigma}=\frac{\Sigma_P+\Sigma_Q}{2}.$$

or for Gaussian mixtures, it can be evaluated using the method suggested in [30].

The experimental design problem may then be posed as an optimisation problem (the results in this article used a genetic algorithm [31] of population size 100 and 20 generations) over ϕ - we search for the set of experimental parameters, $\phi^\star$, for which the Hellinger distance between the competing models $(\mathcal{M}_i,\mathcal{M}_j)$, $H(P(\hat{y}|\mathcal{M}_i,\phi^\star),P(\hat{y}|\mathcal{M}_j,\phi^\star))$, is maximal. $\phi^\star$ will then specify the experiment that gives the greatest chance of distinguishing between $\mathcal{M}_i$ and $\mathcal{M}_j$. In the case where more than two models are considered, the cost function is taken as

$$\sum_{i<j} e^{H(P(\hat{y}|\mathcal{M}_i,\phi^\star),P(\hat{y}|\mathcal{M}_j,\phi^\star))}.$$

where the sum of exponentials is introduced to encourage selection of experiments with a high chance of distinguishing between a subset of the model pairs, over experiments with less decisive information for any pair of models, but perhaps a larger average Hellinger distance over all model pairs.

Acknowledgments

We would like to thank the members of the *Theoretical Systems Biology Group* at Imperial College London for fruitful discussions. We dedicate this paper to Joe Silk.

Author Contributions

Conceived and designed the experiments: DS MPHS. Performed the experiments: DS. Analyzed the data: DS CPB. Contributed reagents/materials/analysis tools: DS CPB PDWK TT MPHS. Wrote the paper: DS CPB MPHS.

References

1. 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. Molecular bioSystems 7: 1593–1602.
2. Liepe J, Filippi S, Komorowski M, Stumpf MPH (2013) Maximizing the information content of experiments in systems biology. PLoS computational biology 9: e1002888.
3. Lindley DV (1956) On a measure of the information provided by an experiment. The Annals of Mathematical Statistics 27: 986–1005.
4. Vanlier J, Tiemann CA, Hilbers PAJ, van Riel NAW (2012) A Bayesian approach to targeted experiment design. Bioinformatics (Oxford, England) 28: 1136–1142.
5. Huan X, Marzouk YM (2012) Simulation-based optimal Bayesian experimental design for nonlinear systems. Journal of Computational Physics 232: 288–317.
6. Kutalik Z, Cho KH, Wolkenhauer O (2004) Optimal sampling time selection for parameter estimation in dynamic pathway modeling. Biosystems 75: 43–55.
7. Chu Y, Hahn J (2008) Integrating parameter selection with experimental design under uncertainty for nonlinear dynamic systems. AIChE Journal 54: 2310–2320.
8. Apgar JF, Witmer DK, White FM, Tidor B (2010) Sloppy models, parameter uncertainty, and the role of experimental design. Molecular bioSystems 6: 1890–1900.
9. Flassig RJ, Sundmacher K (2012) Optimal design of stimulus experiments for robust discrimination of biochemical reaction networks. Bioinformatics (Oxford, England) 28: 3089–3096.
10. Busetto AG, Hauser A, Krummenacher G, Sunnåker M, Dimopoulos S, et al. (2013) Near-optimal experimental design for model selection in systems biology. Bioinformatics 29: 2625–2632.
11. Julier S, Uhlmann J, Durrant-Whyte H (2000) A new method for the nonlinear transformation of means and covariances in filters and estimators. IEEE Transactions on Automatic Control 45: 477–482.
12. Feroz F, Hobson MP, Bridges M (2009) MULTINEST: an efficient and robust Bayesian inference tool for cosmology and particle physics. Mon Not Roy Astron Soc 398: 1601–1614.
13. Kirk P, Thorne T, Stumpf MP (2013) Model selection in systems and synthetic biology. Current opinion in biotechnology 24: 551–826.
14. Aitken S, Akman OE (2013) Nested sampling for parameter inference in systems biology: application to an exemplar circadian model. BMC systems biology 7: 72.
15. Quaiser T, Dittrich A, Schaper F, Mönnigmann M (2011) A simple work flow for biologically inspired model reduction–application to early JAK-STAT signaling. BMC systems biology 5: 30.
16. Yamada S, Shiono S, Joo A, Yoshimura A (2003) Control mechanism of JAK/STAT signal transduction pathway. Febs Letters 534: 190–196.
17. MacKay DJ (1992) Information-based objective functions for active data selection. Neural computation 4: 590–604.
18. Team CW (2010) Good practice guidance paper on assessing and combining multi model climate projections. In: IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections. p. 1.
19. Kitano H (2002) Systems biology: a brief overview. Science 295: 1662–1664.
20. Anderson J, Papachristodoulou A (2009) On validation and invalidation of biological models. BMC bioinformatics 10: 132.
21. Julier SJ The scaled unscented transformation. In: Proceedings of the 2002 American Control Conference. American Automatic Control Council. pp. 4555–4559.
22. Wan E, van der Merwe R (2000) The unscented Kalman filter for nonlinear estimation. Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000 AS-SPCC The IEEE 2000 : 153–158.
23. Jeffreys H (1961) Theory of Probability. 3rd ed. Oxford: The Clarendon Press.
24. Sunnåker M, Busetto AG, Numminen E, Corander J, Foll M, et al. (2013) Approximate Bayesian computation. PLoS computational biology 9: e1002803. doi:10.1371/journal.pcbi.1002803.
25. Zhou Y, Liepe J, Sheng X, Stumpf MP, Barnes C (2011) Gpu accelerated biochemical network simulation. Bioinformatics 27: 874–876.
26. Alspach D, Sorenson H (1972) Nonlinear Bayesian estimation using Gaussian sum approximations. Automatic Control, IEEE Transactions on 17: 439–448.
27. Faubel F, McDonough J (2009) The split and merge unscented Gaussian mixture filter. Signal Processing Letters.
28. Horenko I, Weiser M (2003) Adaptive integration of molecular dynamics. Journal of computational chemistry 24: 1921–1929.
29. Weiβe A, Horenko I, Huisinga W (2006) Adaptive approach for modelling variability in pharmacokinetics. Computational Life Sciences II : 194–204.
30. Kristan M, Leonardis A (2010) Multivariate online kernel density estimation. In: Computer Vision Winter Workshop. pp. 77–86.
31. Perone CS (2009) Pyevolve: a Python open-source framework for genetic algorithms. ACM SIGEVOlution 4: 12–20.