Designing experiments to understand the variability in biochemical reaction networks

Jakob Ruess, Andreas Milias-Argeitis and John Lygeros

Automatic Control Laboratory, ETH Zurich, 8092 Zurich, Switzerland

Exploiting the information provided by the molecular noise of a biological process has proved to be valuable in extracting knowledge about the underlying kinetic parameters and sources of variability from single-cell measurements. However, quantifying this additional information \textit{a priori}, to decide whether a single-cell experiment might be beneficial, is currently only possible in systems where either the chemical master equation is computationally tractable or a Gaussian approximation is appropriate. Here, we provide formulae for computing the information provided by measured means and variances from the first four moments and the parameter derivatives of the first two moments of the underlying process. For stochastic kinetic models for which these moments can be either computed exactly or approximated efficiently, the derived formulae can be used to approximate the information provided by single-cell distribution experiments. Based on this result, we propose an optimal experimental design framework which we employ to compare the utility of dual-reporter and perturbation experiments for quantifying the different noise sources in a simple model of gene expression. Subsequently, we compare the information content of a set of experiments which have been performed in an engineered light-switch gene expression system in yeast and show that well-chosen gene induction patterns may allow one to identify features of the system which remain hidden in unplanned experiments.

1. Introduction

Quantitative studies of biological systems with mathematical models strongly depend on an appropriate characterization of the underlying system, that is on good knowledge about the underlying mechanisms and kinetic parameters. While extracting such knowledge from averaged cell population data is common practice, it has only recently been realized that also the molecular noise observed in single-cell measurements may be a rich source of information about the parameters of stochastic kinetic models \cite{1–4}. Mathematically, one way to quantify the information provided by single-cell experiments is to determine the precision to which the model parameters can at best be estimated in a given experimental set-up, that is to determine the variances of the best possible unbiased estimators of the model parameters \cite{5}. Thanks to the Cramér–Rao inequality these variances can be computed from the Fisher information matrix. To compute the Fisher information for stochastic kinetic models, one has to solve the chemical master equation (CME) \cite{6} and take derivatives of its solution with respect to the model parameters. This is, however, only possible in the simplest cases and even approximation techniques either remain limited to very small systems \cite{7} or are based on strong assumptions \cite{8,9}, which are not always fulfilled in real applications. Consequently, experiments are usually designed based on the intuition of the experimenter, rather than on information theoretic criteria.

A second difficulty in the analysis and design of single-cell experiments is that stochastic kinetic models are usually based on the assumption that the same process governs the evolution in all cells of the population. This, however, is generally not the case, because the process of interest often interacts with other unmodelled factors which are themselves subject to fluctuations and differ between the cells.
For instance, differences in cell size, local growth conditions or expression capacity [10,11] may lead to additional variability in the cell population. In many instances, noise resulting from such extrinsic variability [12–14] has been reported to dominate the molecular noise of the process under study [15–17]. In such situations, methods which assume a homogeneous cell population and attribute all the observed variability to molecular noise of the modelled reactions may lead to biased results. Sometimes, it may be appropriate to assume that such unmodelled factors are static for the time scales of interest. For example, in the model of the stress response of budding yeast to osmotic pressure in Zechner et al. [3], the number of mitochondria affecting translation changes much more slowly than species in signalling and transcription cascades. In this case, the number of mitochondria can be taken as random but constant in time for the purposes of the model. In other cases, however, species which are not included in the model but affect reaction rates may evolve on time scales comparable to those of the reactions of interest [12,18,19], for instance global regulators affecting transcription. In theory, one should include such species in the model but this is often not practical because it would lead to models of intractable size. A convenient modelling abstraction may then be to include a stochastic process for some of the rates, to serve as a rudimentary abstraction of the complex mechanisms governing the fluctuations of the reaction rates [11].

In this paper, we propose a framework for optimally designing single-cell distribution experiments for identifying the parameters of stochastic kinetic models in which the reaction rates are possibly governed by stochastic differential equations. To this end, we first demonstrate on systems where one reaction rate is governed by a stochastic differential equation how equations describing the time evolution of the moments of the probability distribution can be derived. We then show how the Fisher information can be approximated from the first four moments and the parameter derivatives of the first two moments without the need of any assumption other than a sufficiently large measured cell population. Finally, we embed the approximated Fisher information into an optimization algorithm which returns the most informative experiment for a specified set of model parameters. This allows us to design optimal experiments for identifying specific features of the system. We demonstrate this by comparing dual-reporter and perturbation experiments in a simple model of gene expression, where the mRNA production rate is varying according to a stochastic differential equation. Finally, we study the variability in bation experiments in a simple model of gene expression, where the mRNA production rate is varying according to a stochastic differential equation. Finally, we study the variability in the sense that the time evolution of the moments of any order can be approximated by a stochastic differential equation how equations describing the time evolution of the moments of the probability distribution can be derived. We then show how the Fisher information can be approximated from the first four moments and the parameter derivatives of the first two moments without the need of any assumption other than a sufficiently large measured cell population. Finally, we embed the approximated Fisher information into an optimization algorithm which returns the most informative experiment for a specified set of model parameters. This allows us to design optimal experiments for identifying specific features of the system. We demonstrate this by comparing dual-reporter and perturbation experiments in a simple model of gene expression, where the mRNA production rate is varying according to a stochastic differential equation. Finally, we study the variability in the system which jointly describes the time evolution of the species and the reaction rate is a stochastic hybrid system [21].

The time evolution of the moments can be computed as

\[ \frac{d}{dt} \mathbb{E}[\phi(a, x(t))] = \mathbb{E}[\mathcal{L}\phi(a, x(t))], \]

where \( x(t) = [x_1(t), \ldots, x_m(t)] \) is a vector containing the molecule counts of the \( m \) species and \( \mathcal{L} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \) is chosen such that the left-hand side gives the derivative of the desired moment. \( L \) is the extended generator of the stochastic hybrid system, given by

\[ \mathcal{L}\phi(a, x) := \frac{\partial \phi(a, x)}{\partial a} f(a) + \frac{1}{2} \frac{\partial^2 \phi(a, x)}{\partial a^2} \sigma^2 a^2 + \sum_{j \in \mathcal{R}} \phi(a, x + v_j) - \phi(a, x) \sigma_j(a, x), \]

where \( f(a) = r(a - a), \sigma_j \) are the stoichiometric transition vectors and \( \sigma_j(a, x) \) the propensities of the \( K \) reactions. Note that the resulting system of moment equations may be non-closed in the sense that the time evolution of the moments of any order depends on moments of higher order. In such cases, the moments cannot be computed exactly and approximation techniques have to be used [22–24].

2.2. Approximating the Fisher information

The amount of information about model structure or parameters, which can be gained from measurements, may be highly dependent on the experimental set-up that is chosen [25–29]. Carefully planning an experiment reduces experimental effort and resources and may even allow one to answer questions which cannot be answered from unplanned experiments.

One way to assess the information about a vector of unknown model parameters \( \theta = [\theta_1, \ldots, \theta_n]^T \) that an experimental set-up can supply is through the computation of the Fisher information matrix \( I(\theta) \) [5,30]. The diagonal elements of the inverse of \( I(\theta) \) give lower bounds for the variances that any unbiased estimators of the model parameters can attain, and thus the Fisher information gives a measure of the accuracy to which the model parameters can be estimated in a given experimental setting (for a more detailed discussion, we refer the reader to [5]). The elements of the Fisher information matrix are given by

\[ (I(\theta))_{ij} = \mathbb{E} \left[ \left( \frac{\partial}{\partial \theta_j} \log f(Y; \theta) \right) \left( \frac{\partial}{\partial \theta_i} \log f(Y; \theta) \right) \right]. \]

where \( Y \) is the random variable which is experimentally measured and \( f(Y; \theta) \) is its distribution. In stochastic kinetic models, the parameter vector \( \theta \) typically contains reaction rates.

2. Material and methods

2.1. Moment equations for reaction systems with stochastically varying reaction rates

The time evolution of the probability distribution of stochastic kinetic models is governed by the CME. If variability in the reaction rates is present in a population, the distribution for each cell can be described by a CME, which is conditioned on the realizations of the reaction rates in the cell. In the study of Zechner et al. [3], it was shown how population moments can be computed from this conditional CME under the assumption that the reaction rates are constant in time. More generally, assume now that a reaction rate \( a_t \) is governed by a stochastic differential equation of the form

\[ da_t = r(a_t - a)dt + s\sqrt{a}dW_t, \]

where \( W_t \) is a standard Brownian motion. This process fluctuates around its mean \( \mu_a \) where the mean reversion speed \( r \) gives the autocorrelation time of the process, and thereby determines the time scale of the rate fluctuation. The noise coefficient \( s \) determines the size of the deviations from the mean, whereas the term \( \sqrt{a} \) prevents the process from taking negative values and is in accordance with the frequently used Langevin approximation for chemical reaction networks [9]. Note that this formulation includes constant reaction rates as for \( r = s = 0 \) the process is constant and always distributed according to its initial distribution.

The system which jointly describes the time evolution of the species and the reaction rate is a stochastic hybrid system [21].

The system which jointly describes the time evolution of the species and the reaction rate is a stochastic hybrid system [21].
and possibly also parameters describing the variability in the reaction rates, for instance, moments of the parameter distributions [3] or the coefficients $r_i$, $\mu_2$, and $s$ in equation (2.1). Because the values of the elements of the parameter vector often differ by orders of magnitude, it is sometimes more appropriate to compute the derivatives with respect to the logarithm of the parameters. It can be shown that these derivatives correspond to the sensitivity of the measured output with respect to the relative, instead of the absolute, changes in the parameters. It can be shown that these derivatives correspond to the sensitivity of the measured output with respect to the parameters. It can be shown that these derivatives correspond to the sensitivity of the measured output with respect to the parameters.

The Fisher information matrix for a logarithmic parametrization is targeted to specific parameters or to subsets of the parameter set. For instance, one might want to estimate the reaction rates of a noise model which have no biological meaning. An optimality criterion targeted to such questions is $D$-optimality [36,37]. It is based on partitioning the parameter vector $\theta = [\theta_1, \theta_2]^T$ in parameters of interest $\theta_1$ and nuisance parameters $\theta_2$. The experiment, which allows one to obtain the confidence region with minimum volume for $\theta_1$ can then be found by maximizing the determinant of

$$I_1(\theta) = I_{11}(\theta) - I_{12}(\theta)I_{22}(\theta)^{-1}I_{12}(\theta),$$

where $I(\theta) = [I_{11}(\theta), I_{12}(\theta); I_{12}(\theta), I_{22}(\theta)]$, which again gives the correct expression for the complete information.

### 2.3. Designing optimal experiments

The goal of experimental design is to find the experiment which is optimal according to some criterion reflecting information about the unknown parameters. The most frequently used criteria are $D$-optimality, $A$-optimality and $E$-optimality which correspond to maximizing the determinant, minimizing the trace of the inverse and maximizing the minimal eigenvalue of the Fisher information matrix, respectively [27,35]. In biological applications, it is often desirable to design experiments which are targeted to specific parameters or to subsets of the parameter set. For instance, one might want to estimate the reaction rates of a noise model which have no biological meaning. An optimality criterion targeted to such questions is $D$-optimality [36,37]. It is based on partitioning the parameter vector $\theta = [\theta_1, \theta_2]^T$ in parameters of interest $\theta_1$ and nuisance parameters $\theta_2$. The experiment, which allows one to obtain the confidence region with minimum volume for $\theta_1$ can then be found by maximizing the determinant of

$$I_1(\theta) = I_{11}(\theta) - I_{12}(\theta)I_{22}(\theta)^{-1}I_{12}(\theta),$$

where $I_1(\theta)$ and $I_2(\theta)$ are the information matrices for $\theta_1$ and $\theta_2$, respectively, and $I_{12}(\theta)$ and $I_{21}(\theta)$ give the cross terms between $\theta_1$ and $\theta_2$.

The computation of $I(\theta)$ and $I_1(\theta)$ requires knowledge of the true parameters $\theta$ which are not available. This difficulty can be overcome by evaluating the information at some initial estimate $\hat{\theta}$ [30]. If, however, this initial estimate differs significantly from the true parameter vector, the resulting experiment may be far from optimal. This is especially important for biological applications where initial estimates, if available at all, usually involve large uncertainties. Here, we chose an approach which includes the uncertainty of the initial estimate in the form of a prior distribution $\pi(\theta)$ and computes the expected information with respect to $\pi(\theta)$ [38,39] (for an overview of other methods, see electronic supplementary material, §4.4). The corresponding optimal experiment $e^*$ can then be obtained by solving the following optimization problem:

$$e^* = \arg \max_{e \in E} \{E_e[\det I_1(\theta, e)]\},$$

where the expectation is taken over $\theta \sim \pi(\theta)$. $I_1(\theta, e)$ is the information matrix for experiment $e$ evaluated at $\theta$ and $E$ is the set of possible experiments. We can now state a procedure for designing optimal experiments for the estimation of parameters of stochastic kinetic models from single-cell measurements of a cell population.

Some comments on practical applicability of this procedure are given in the electronic supplementary material, §5.6. This optimal experimental design procedure can be performed in iterations with experiments. Starting from some prior distribution $\pi(\theta)$, the computations lead to optimal experiments that yield data which can be used in a parameter inference scheme to compute posterior distributions. These can then in turn serve as new prior distributions for the computation of a new optimal experiment. This can be continued until the uncertainty about the parameters has been sufficiently reduced.
3. Results

3.1. In silico study of a gene expression system

We demonstrate the proposed experimental design framework on a simple example of gene expression. The model we consider consists of the two species mRNA \( M \) and protein \( P \) and the four reactions

\[
\begin{align*}
\emptyset & \rightleftharpoons^a g(t) \quad M, \\
M & \rightleftharpoons^b \emptyset, \\
M & \rightleftharpoons^c M + P, \\
P & \rightleftharpoons^d \emptyset,
\end{align*}
\]

where \( b = 0.03, c = 0.5, d = 0.04 \) and \( a_i \) is varying according to a stationary stochastic process of the form equation (2.1). The dynamics of the moments of order up to two of \( M \) and \( P \) are then completely determined by the mean reversion speed \( r \) and mean \( \mu_a \) and variance \( \nu_a \) of the stationary distribution of \( a_i \) (see electronic supplementary material, §S.2.1).

Here, we assume that the values of these parameters are \( \mu_a = 0.5, \nu_a = 0.1 \) and \( r = 0.005 \). We further assume that the gene can be switched between an on state (where \( g(t) = 1 \)) and an off state (where \( g(t) = 0 \)) using some external input, for example, either by adding different nutrients in nutrient shift experiments [40], by adding salt to induce the osmotic stress response [41] or with light pulses [20]. Furthermore, throughout this section, we assume that it is known that no molecules are present at time \( t = 0 \) (loosely speaking, the gene has been off for some time at the start of the experiment) and that the degradation rates \( b \) and \( d \) are known, whereas \( \mu_p \), \( \nu_p \), \( r \) and \( c \) have to be determined from the measurements. Finally, for simplicity, all the computations of this section are performed locally at the 'true' parameter values and prior uncertainty about the parameters is not included.

We compare four experimental methods in terms of the information they can provide about the unknown parameters. For all methods, we assume that the experiments are limited to a time length of \( t = 300 \) time units and that at most 10 measurements of the protein distribution are taken during that time. The first two methods we consider are standard time course experiments, where the gene is switched on only at time zero and measurements are taken in regular time intervals (every 30 time units). In the first method (referred to as unplanned experiments), we assume that a single reporter protein is measured, whereas in the second method (unplanned dual-reporter experiments), an identical copy of the gene is added to the cells, such that a second reporter protein which is conditionally independent of \( P \), given the history of \( a_0 \), can be measured [15,42,43]. These two methods are compared to more sophisticated experiments where informative gene-switching patterns and measurement times are identified using our experimental design framework.

For the optimally designed experiments, we again consider both single and dual-reporter experiments (referred to as optimal perturbation and optimal dual-reporter experiments, respectively). In both cases, the search for the most informative experiments proceeds in two steps. First, we fix equally spaced measurement times and use a Markov chain Monte Carlo-like algorithm to perform the optimization on the space of possible gene-switching patterns. Second, we fix the resulting gene-switching pattern and sequentially place the measurement times, when they yield maximal information. More details on this algorithm are given in the electronic supplementary material, §S.2.2. Figure 1 gives the resulting optimal perturbation experiments when either \( r \) or \( V_a \) is taken as parameter of interest (\( t_i \) in equation (2.5)) and the remaining parameters are taken as nuisance parameters (\( t_i \) in equation (2.5)). Results for the remaining parameters and results for the optimal dual-reporter experiments are given in the electronic supplementary material, figures S2 and S3. It can be seen that different perturbations and measurement times are optimal for identifying different parameters.

The results of the comparison of the four methods are summarized in table 1. Note that the reported values correspond to a logarithmic parametrization to allow one to compare the information obtained for different parameters (rows). From the first column, we see that the information which can be gained from unplanned experiments is very small. This indicates that the parameters may be practically unidentifiable (see also the discussion in the electronic supplementary material, §S.2.3). Unplanned dual-reporter experiments, on the other hand, lead to much more information (second column in table 1) and appear to be suitable for identifying all the parameters of the system. Only \( r \) potentially remains difficult to identify. In general, \( r \) and \( V_a \) are harder to identify than \( \mu_p \) and \( c \) because the protein mean does not depend on \( r \) and \( V_a \), and hence they can only be identified from the protein variance. The information of the optimal perturbation experiments which were found by our experimental design procedure are given in the third column of table 1, where for all parameters the first value corresponds to the information which is obtained if the experiment is specifically targeted at the parameter (taking all other parameters as nuisance parameters) and the second value corresponds to the information which is obtained if the experiment is targeted at estimating all parameters. It can be seen that, compared to dual-reporter experiments, more information is obtained for the parameter \( r \), whereas dual-reporter experiments lead to more information for \( \mu_p \), \( V_a \) and \( c \). Hence, depending on the objective of the study, different experimental strategies are preferable.

The fourth experimental method, which combines optimal perturbations and dual reporters, naturally leads to the most information for all parameters (fourth column in
Note, however, that the increase in information compared to unplanned dual-reporter experiments is very small if $m_a$ or $c$ are the parameters of interest, which indicates that additionally perturbing the system may not be worth the effort. Furthermore, contrary to the optimal perturbation experiments, targeting the dual-reporter experiment at specific parameters (first value in the fourth column) yields only a minor increase in information compared with an experiment which is targeted at all parameters (second value in the fourth column). The experiment targeted at identifying $r$ even leads to less information about $r$ than the experiment targeted at all parameters which shows that the optimization algorithm used for finding informative experiments converged to a local minimum in the former case.

The maximum-likelihood estimator for the parameter vector is asymptotically normally distributed with covariance matrix equal to the inverse of the Fisher information matrix. We can therefore further visualize our results by computing confidence regions for the maximum-likelihood estimator in the different experiments. Figure 2 shows two-dimensional 95% confidence ellipses for all pairs of parameters for the optimal perturbation experiments targeted at each of the parameters where we assume that at each time point a cell population of size $n = 10000$ is measured. The red ellipses correspond to the experiment which is targeted at all parameters and are therefore overall the smallest. However, the variance in the direction of each of the parameters can be reduced by specifically targeted experiments. The most significant difference can be seen in the experiments targeted either at $m_a$ or $c$ (black and magenta) where the size of the ellipses is reduced in both the directions of $m_a$ and $c$ at the cost of making $r$ and $V_a$ much harder to identify. The
parameter $r$ is the hardest to identify and can only be reasonably constrained by experiments targeted either directly at $r$ or at all parameters. Figure 3 shows two-dimensional 95% confidence ellipses for all pairs of parameters for the unplanned experiment, the unplanned dual-reporter experiment, the optimal perturbation experiment targeted at all parameters and the optimal dual-reporter experiment targeted at all parameters. Again, we assume that at each time point a cell population of size $n = 10,000$ is measured. As already indicated by the information values in table 1, the parameters are almost unidentifiable from unplanned experiments. All the other experiments, on the other hand, can constrain the parameters to relatively small regions. Finally, we also computed the information under a Gaussian assumption. Our results (electronic supplementary material, table S1 and figure S4) show that for many objectives a Gaussian assumption leads to information estimates which are overly optimistic. This is most probably owing to the independence assumption of sample mean and variance which is implicitly imposed by a Gaussian approximation and is not valid for the system in question.

3.2. Characterizing variability in a light-induced gene expression system

Next, we study a light-switch gene expression system which has been engineered in yeast [20]. The authors used a light responsive module to initiate transcription by shining red light on the yeast culture and to terminate transcription by shining far-red light. They then proposed a control scheme to regulate the mean amount of protein in the population. The development of more sophisticated control schemes (for example, to allow one to also control the protein variance) requires a proper characterization of the sources of variability in the system. To this end, our framework can be used to compare the utility of different experiments, and ultimately to

Figure 2. Comparison of optimal perturbation experiments targeted at different parameters. Ninety-five per cent confidence ellipses for all pairs of parameters are shown in the different panels. Experiments are targeted at $\mu_a$ (black), $V_a$ (orange), $c$ (magenta), $r$ (green) and all parameters (red). Note that these ellipses correspond to confidence regions for the original parameters (not their logarithms).
design the experiments which are optimal for the characterization of the different noise sources. We thus developed a stochastic version of the model [20] which includes a stochastic differential equation of the form equation (2.1) for the mRNA production rate. This introduced four additional parameters which were not identified in Milias-Argeitis et al. [20]. To characterize uncertainty about these parameters, we chose independent uniform prior distributions and computed the expectation of the information (for a logarithmic parametrization), which is provided by the experiments reported in fig. 1 of Milias-Argeitis et al. [20] about each of the additional parameters according to equations (2.5). Thereby, the remaining parameters which were already identified in Milias-Argeitis et al. [20] were fixed to their known values (see electronic supplementary material, §S.3.1 and §S.3.2). The results are shown in tables S2 and S3 in the electronic supplementary material.

Which experiment is best again depends on the objective. For instance, the experiment where a red light pulse is applied at the beginning and a far-red light pulse after 30 min is best for identifying the protein production rate but worst for identifying the mean reversion rate \( r \) of the mRNA production rate. This is most probably owing to the fact that if the gene is switched off, the mRNA production rate is set to zero and the parameters describing this rate do not influence the dynamics anymore.

Furthermore, our results show that even though the experiments in fig. 1e of Milias-Argeitis et al. [20] were performed over a shorter time and contain fewer measurements than the experiments in fig. 1c of Milias-Argeitis et al. [20], they provide much more information about the parameter \( r \). This suggests that experiments where the gene is expressed for short time intervals with silent periods in between could allow one to
determine $r$, and thus to test whether the mRNA production rate can be assumed to be constant or whether a stochastic process description as used in this paper is required. Our results in the electronic supplementary material, §S.3.3 indicate that indeed, contrary to other experiments, the variance measured in the experiments in fig. 1 indicates that the light-pulse pattern and measurement times we determined with our method are shown in figure 4. They lead to an experiment which carries close to four times more information than any of the experiments in Milias-Argeitis et al. [20] suggesting that our experimental design framework can be a valuable tool for characterizing variability in this system.

4. Discussion

While knowledge about biological mechanisms is constantly growing, our understanding of the stochasticity of biological systems and its influence on system dynamics remains rather limited. Many different sources of variability may play a role and neglecting any one of them may lead to over- or underestimating the effect of others. In the gene expression model, for instance, one could use the methods of Friedman et al. [44] to estimate the protein burst size and frequency from measurements of the stationary protein distribution, under the assumption that the mRNA production is the same for all cells in the population. If, however, the data actually come from a population with variable mRNA production rate these estimates would be biased—potentially by more than an order of magnitude (see electronic supplementary material, §S.2.3). Allowing reaction rates to vary between individuals in a cell population offers a way to incorporate variability stemming from unknown factors and enables model-based studies of heterogeneous cell populations. We showed how the information about unknown parameters of such models which is provided by means and variances of measured populations can be approximated if the first four moments and the parameter derivatives of the first two moments of the underlying process can be computed. The derived formulae are applicable as long as the measured population is of sufficient size for the application of the central limit theorem. This opens up the possibility to pose many interesting questions: do the measurements contain enough information to separate different noise sources? How much information can be gained by measuring the variance in addition to the mean? And most importantly: what is the most informative experiment? By means of examples, we demonstrated that unplanned experiments may not contain enough information to identify the model parameters and that designing experiments based on intuition alone may not be sufficient. For instance, placing all the measurements either very early or very late in the experiment turns out to be better for identifying the mean reversion speed in our examples (figures 1 and 4) but appears very intuitive at a first glance.

Our results (table 1 and figure 2; electronic supplementary material, figure S2) show that the optimal experiments are highly dependent on the chosen objective. In some cases, introducing a dual reporter yields high information, in other cases perturbing the system with input stimuli is preferable. A study of the system using the experimental design framework presented in this paper allows a comparison of the different experimental approaches and enables one to choose the approach which is most likely to be successful for the given objective. The resulting experiments can in turn be used to refine the model and to update the parameter estimates, giving rise to an iterative procedure of successive rounds of computations and experiments.

In the light-switch gene expression system, the computation of the information contents of the different experiments shows that perturbing the system with different light-pulse sequences can highlight different features of the system. This suggests that well-chosen gene induction patterns may allow one to uncover features of the system which remain hidden in unplanned experiments. For instance, the electronic supplementary material, figure S5 suggests that temporal fluctuations in the mRNA production rate may play a role for this system. Perturbing a system with light pulses to understand the variability may seem to be limited to this specifically engineered system. However, a similar strategy could also be employed by exploiting naturally occurring biological mechanisms. For example, in the study of Zechner et al. [3], the authors studied gene expression in yeast in response to osmotic pressure. Different salt concentrations led to different residence times of the signalling molecule Hog1 in the nucleus, and thereby created different input signals to the downstream gene expression system. In that system, multiple subsequent salt stresses, which can for instance be implemented using a microfluidic device as in Uhlenendorf et al. [41], could serve as the equivalent of the multiple light pulses used in this paper and might give further insights into the specific nature of the system.

Acknowledgements. J.R. and J.L. designed research. J.R. developed the method and performed the computations. A.M. assisted with supplementary calculations. J.R. and J.L. wrote the paper.

Funding statement. The work was supported in part by the European Commission under the project MoVeS.