Iterative experiment design guides the characterization of a light-inducible gene expression circuit

J. Ruess*, F. Parise*, A. Milias-Argeitis, M. Khammash, J. Lygeros

S.1 Nomenclature

Table S.1: **Symbols used in the supplementary material.**

	Symbol	Variable
Species	M	mRNA
	P	dark protein
	F	fluorescent protein
	I	protein fluorescence intensity
	A	autofluorescence artifacts and technical noise
	Y	total fluorescence
Parameters	M_{k_M}	mean of the mRNA production rate across the population
	V_{k_M}	variance of the mRNA production rate across the population
	k_P	protein production rate
	k_F	protein maturation rate
	c_M	mRNA degradation rate
	c_P	protein degradation rate
	d_r	dark reversion: decay parameter
	h	dark reversion: lag parameter
	r	scaling parameter to convert F into I
Moments	$\tilde{\mu}^e$	vector of all the uncentered moments up to order 4 in experiment e
	μ_i^A	centered moment of order i of the autofluorescence and technical noise A
	$\mu_i^{e\,F}$	centered moment of order i of F in experiment e
	$\mu_i^{e\,I}$	centered moment of order i of I in experiment e
	μ_i^e	centered moment of order i of Y in experiment e
	$\hat{\mu}_i^e$	centered moment of order i of Y, estimated from measured data of experiment e
Others	L	sequence of red and far-red light pulses
	$\tilde{\gamma}$	vector of all parameters except the scaling factor: $[M_{k_M}\ V_{k_M}\ k_P\ k_F\ c_M\ c_P\ d_r\ h]$
	γ	vector of all parameters: $[\tilde{\gamma}\ r]$
	$\hat{\gamma}^i$	MAP estimate of the parameter vector after i optimal experiments
	$I(\gamma, e)$	Fisher information matrix for experiment e, according to the parameter vector γ
	N	number of samples taken in every measurement
	Q	number of particles sampled from the parameter posterior distribution
	$u(t; \gamma, L)$	input that models the effect of the light pattern L on the mRNA production rate, given the dark reversion parameters d_r, h

*J.R. and F.P. contributed equally to this work.

Table S.2: **Nomenclature associated with the experiments used to infer the parameters.**

Name	Type	Figure illustrating the data
O1	first optimal	Figure 2B (main text)
O2	second optimal	Figure 2C (main text)
R1	random	Figure S.5 R1
R2	random	Figure S.5 R2
R3	random	Figure S.5 R3
I1	experience-based	Figure S.6 I1
I2	experience-based	Figure S.6 I2
I3	experience-based	Figure S.6 I3

Table S.3: **Nomenclature associated with the validation and control experiments.**

Name	Type	Figure illustrating the data
V1	validation	Figure 3 (main text)
V2	validation	Figure 4 (main text)
C1	control (mean)	Figure 5A (main text)
C2	control (mean)	Figure 5B (main text)
C3	control (variance)	Figure 5C (main text)
C4	control (CV)	Figure 5D (main text)

S.2 Stochastic modeling of the light-inducible gene expression circuit

S.2.1 Description of the biochemical reaction network

To model the light-inducible gene expression circuit we use the biochemical reaction network illustrated in Figure 1 in the main text, that consists of the following reactions:

$$\begin{array}{llcl}
\text{Reaction 1:} & \emptyset & \xrightarrow{k_M \cdot u(t)} & M \\
\text{Reaction 2:} & M & \xrightarrow{c_M} & \emptyset \\
\text{Reaction 3:} & M & \xrightarrow{k_P} & M + P \\
\text{Reaction 4:} & P & \xrightarrow{k_F} & F \\
\text{Reaction 5:} & P & \xrightarrow{c_P} & \emptyset \\
\text{Reaction 6:} & F & \xrightarrow{c_P} & \emptyset,
\end{array}$$

where the empty set $\emptyset$ denotes that a certain species is produced or degraded without involving the other species, in other words enters of leaves the system. This reaction network is similar to the one used in [1] but contains a crucial difference in the way the light signal $u(t)$ is incorporated. Here, we assume that the mRNA production rate is multiplied by the signal $u(t) = u(t; \gamma, L)$, where γ is the vector of parameters and L is the applied light-pattern,

$$u(t; \gamma, L) = U \frac{e^{-d_r(t-t_c)}}{e^{-d_r(t-t_c)} + h},$$

see Section S.2.1.1 for a more detailed derivation. When a red pulse is applied, t_c is reset to the current time and U is set to one, so that mRNA transcription takes place with maximum rate. The unknown parameters d_r and h capture the natural decay of the signal after a red pulse due to dark reversion [1]. When a far-red pulse is applied U is set to zero, so that transcription is arrested until a new red pulse is applied.

S.2.1.1 A theoretical justification for the input signal form

In this section, based on the available knowledge of the light-responsive PhyB-PIF3 module [2] used in the circuit, we derive a model that maps the applied light-induction pattern into a time-varying mRNA transcription rate profile. Since PhyB and PIF3 are highly abundant in the cell ($\gg 1$), we use for this specific task a deterministic model. More in detail, the PhyB-PIF3 system can be modeled using the following reaction scheme (further details can be found in [3]):

$$\varnothing \xrightarrow{k_{B_r}} B_r \xrightarrow{c_D} \varnothing \tag{1}$$

$$\varnothing \xrightarrow{k_{P_3}} P_3 \xrightarrow{c_D} \varnothing \tag{2}$$

$$B_r \underset{d_r,FR}{\overset{R}{\rightleftharpoons}} B_{fr} \tag{3}$$

$$B_{fr} + P_3 \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} B_{fr}P_3 \tag{4}$$

$$B_{fr}P_3 \xrightarrow{d_r,FR} B_r + P_3 \tag{5}$$

$$B_{fr} \xrightarrow{c_D} \varnothing \tag{6}$$

$$B_{fr}P_3 \xrightarrow{c_D} \varnothing \tag{7}$$

In these reactions, PhyB (denoted by B) is assumed to be found in two states: the red-absorbing (inactive) form, B_r and the far-red absorbing (active) form, B_{fr}. PIF3 is denoted by P_3. Both proteins are produced constitutively in our system (with rates k_{B_r} and k_{P_3}, respectively), and are known to be very stable, which means that they are removed only by dilution due to cell growth and division (with rate c_D). Reaction (3) encodes the fact that red light (denoted by R) activates PhyB, while far-red light (denoted by FR) deactivates it[1]. The active form of PhyB can also spontaneously revert to the inactive form in the dark, through a phenomenon called *dark reversion*, that happens with rate d_r. The dark reversion half-life of PhyB is known to be in the order of tens of minutes, or even hours, depending on the cellular context. Active PhyB is able to bind to PIF3 to form a complex $B_{fr}P_3$. This complex can also decay through dark reversion, as modeled by reaction (5). Moreover, it is known that PhyB (in)activation by light is a very fast reaction that is completed in a fraction of a second, and that complex formation proceeds much faster than dark reversion or cell growth. Thus, after the application of red light, we can consider (4) to be at equilibrium at any given time. That is,

$$\frac{[B_{fr}] \cdot [P_3]}{x} = \frac{k_{-1}}{k_1} = K, \tag{8}$$

where $x = [B_{fr}P_3]$ denotes the concentration of the complex $B_{fr}P_3$.

Let the total amount of PhyB be denoted by B^{TOT} and the total amount of PIF3 by P_3^{TOT},

$$\begin{aligned} B^{TOT} &= B_r + B_{fr} + B_{fr}P_3, \\ P_3^{TOT} &= P_3 + B_{fr}P_3. \end{aligned}$$

In the following we assume that the total amounts of PhyB and PIF3 are constant over all experiments, since their respective genes are transcribed from constitutive promoters, as described in [1]. Hence, B^{TOT} and P_3^{TOT} are constant over the course of the experiment.

Our main assumption is that the mRNA transcription rate is proportional to the concentration of the complex $B_{fr}P_3$ present in the cell, which is able to recruit the Gal4 activation domain to the Gal4 promoter. To this end let

$$s(t) = \frac{[B_{fr}(t)] + x(t)}{[B^{TOT}]} = 1 - \frac{[B_r(t)]}{[B^{TOT}]}, \tag{9}$$

[1] To be more precise, red light also partially inactivates active PhyB, while far-red light partially activates the inactive form. In other words, red and far-red light both establish a *photoequilibrium* of phytochrome molecules. For simplicity, the residual (in)activation by the opposite wavelength can be safely ignored, since it is relatively small.

denote the relative concentration of PhyB molecules that are in the active form, both free or bound to PIF3. When a red pulse is applied all the inactive PhyB becomes active ($B_r = 0$). Therefore, according to equation (9), the effect of a red pulse is to reset the signal s to 1. On the other hand, if a far-red pulse is applied all the active PhyB (even the one bound with PIF3) becomes inactive, hence $B_r = B^{TOT}$ and $s = 0$. Finally, in the dark the amount of active PhyB slowly decays according to reactions (3) and (5), due to the dark-reversion effect. In particular, in between pulses, the amount of inactive PhyB follows the differential equation

$$\frac{d}{dt}[B_r(t)] = d_r([B_{fr}(t)] + x(t)).$$

This equation allows us to find the evolution of s in between pulses. In fact

$$\frac{d}{dt}s(t) = \frac{d}{dt}\left(1 - \frac{[B_r(t)]}{[B^{TOT}]}\right) = -\frac{\frac{d}{dt}[B_r(t)]}{[B^{TOT}]} = -d_r\frac{[B_{fr}(t)] + x(t)}{[B^{TOT}]} = -d_r s(t).$$

Overall, the evolution of s is thus given by

$$s(t; \gamma, L) = Ue^{-d_r(t-t_c)}, \tag{10}$$

where t_c is the time when the last pulse was applied and U is set to 1 when a red-pulse is applied and to 0 when a far-red pulse is applied.

The amount of the complex $B_{fr}P_3$ is related to the signal s by the following algebraic equation.

$$\begin{aligned} x(t) &= \left[B^{TOT}\right] s(t) - [B_{fr}(t)] \\ &= \left[B^{TOT}\right] s(t) - K\frac{x(t)}{[P_3(t)]} \\ &= \left[B^{TOT}\right] s(t) - K\frac{x(t)}{[P_3^{TOT}] - x(t)}, \end{aligned} \tag{11}$$

where we used the definition of B^{TOT} in the first line, equation (8) in the second line and the definition of P_3^{TOT} in the last line. From (11) we get the quadratic equation

$$x^2 - x\left(K + \left[P_3^{TOT}\right] + \left[B^{TOT}\right] s\right) + \left[B^{TOT}\right]\left[P_3^{TOT}\right] s = 0,$$

where for simplicity we dropped the dependence on time. This has solutions

$$\begin{aligned} x_{1,2} &= \frac{\left(K + [P_3^{TOT}] + [B^{TOT}]\, s\right) \pm \sqrt{y}}{2}, \text{ where} \\ y &= \left(K + [P_3^{TOT}] + [B^{TOT}]\, s\right)^2 - 4\,[P_3^{TOT}]\,[B^{TOT}]\, s. \end{aligned}$$

The solution x_1, with positive sign, is not feasible since it would imply

$$2x = K + \left[P_3^{TOT}\right] + \left[B^{TOT}\right] s + \sqrt{y} = K + [P_3] + x + [B_{fr}] + x + \sqrt{y} \;\Rightarrow\; 0 = K + [P_3] + [B_{fr}] + \sqrt{y} > 0,$$

which is indeed impossible.

Hence the relation between x and s is given by

$$x = \frac{\left(K + [P_3^{TOT}] + [B^{TOT}]\, s\right) - \sqrt{\left(K + [P_3^{TOT}] + [B^{TOT}]\, s\right)^2 - 4P_3^{TOT}\,[B^{TOT}]\, s}}{2}. \tag{12}$$

Fig. S.1 illustrates this behavior for three different concentrations of total PhyB: $50, 500$ and 1000 nM. In each plot we used 10 different values of K, equally spaced between 10 and 100 (since the dissociation constant of the PhyB-PIF6 complex has been previously estimated to be between 20 and 100 nM [4]), and we assumed a total PIF3 concentration of 100 nM ($\sim$ 2000 molecules per cell). From these plots it

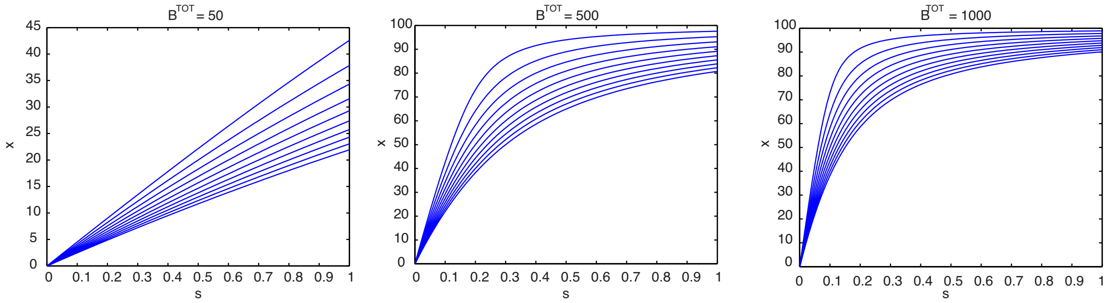

Figure S.1: Relation between x and s according to equation (12), for $\left[P_3^{TOT}\right] = 100$, $K \in \{10, 20, \dots, 100\}$ and three different values of $\left[B^{TOT}\right]$.

appears that the exact functional relation between x and s given in equation (12), can be approximated using the Michaelis-Menten equation

$$x \sim x_{max} \frac{s}{s+h}, \tag{13}$$

where h is an unknown parameter related to K, $\left[P_3^{TOT}\right]$ and $\left[B^{TOT}\right]$, and x_{max} is a scaling parameter. In particular, we observe that if the total PhyB (B^{TOT}) is less abundant than the total PIF3 P_3^{TOT} (left plot), then the relation between x and s is linear (obtainable for h large) while in the opposite case (middle and right plots) it has a hyperbolic shape (obtainable for h small). The quality of the approximation given in equation (13) is ultimately validated by the fact that this model can precisely fit and predict measured data (as presented in the main paper).

Finally, we expect the mRNA transcription rate (r_M) to be approximately proportional to the concentration of x, hence by using equations (10) and (13), we finally get

$$r_M \sim \rho x \sim \rho x_{max} \frac{s}{s+h} := k_M \cdot U \frac{e^{-d_r(t-t_c)}}{e^{-d_r(t-t_c)} + h} := k_M \cdot u(t; \gamma, L), \tag{14}$$

where d_r, k_M and h are unknown parameters that have to be estimated from real data.

S.2.2 Conditional chemical master equation and population moment equations

In the previous section we described a biochemical reaction network that consists of 3 distinct chemical species and 6 reactions. If the system is well stirred and in thermal equilibrium, it can be modeled by a continuous-time Markov chain (CTMC) $X(t) = [M(t)\ P(t)\ F(t)]^\top$ that describes how the amounts of molecules of the different species change in time [5]. $X(t)$ can take states $x = [x_1\ x_2\ x_3]^\top$ in $\mathbb{N}^3$. The rate of occurrence for each possible reaction k is described by the reaction propensities $a_k(x;\theta) : \mathbb{N}^3 \times (\mathbb{R}_0^+)^7 \mapsto \mathbb{R}_0^+$, $k = 1, \dots, 6$, which depend on the amounts x of molecules of the different species which are present in the system and on the parameter vector $\theta = [k_M\ k_P\ k_F\ c_M\ c_P\ d_r\ h]^\top$. The propensity $a_1(x;\theta) = a_1(\theta_1, u(t))$ of the mRNA production reaction also depends on the light-induction pattern L through the signal $u(t) = u(t;\gamma, L)$. In other words $a_k(x;\theta)dt$ is the probability that the k-th reaction takes place in an infinitesimal time interval $[t, t+dt]$ given that $X(t) = x$.

As stated in the main text, we assume that the parameter k_M varies between the cells according to a gamma distribution P_{k_M} with unknown mean M_{k_M} and variance V_{k_M}. Note that the gamma distribution is a generalization of the exponential distribution and can be used to approximate a wide variety of shapes. This assumption means that we allow the parameter k_M to vary among cells but not in time. Other formulations in which k_M is allowed to additionally fluctuate in time would also be possible [6, 7], but we do not investigate such approaches here. Under our assumption, the time evolution of the amounts of molecules in an individual cell can be described by a CTMC conditioned on the value of k_M in that cell. Consequently, we obtain a conditional chemical master equation (CME)

$$\frac{\mathrm{d}}{\mathrm{d}t}p(x,t|k_M) = \sum_{k=1}^{6} -p(x,t|k_M)a_k(x;\theta) + \sum_{k=1}^{6} p(x-\nu_k,t|k_M)a_k(x-\nu_k;\theta), \tag{15}$$

where $p(x,t|k_M)$ is the probability that x molecules are present at time t conditional on the value of the parameter k_M in the cell. $\nu_k \in \mathbb{Z}^3,\ k = 1,\ldots,6$ are the stoichiometric transition vectors of the 6 chemical reactions. Equation (15) implies that the stoichiometric transition vectors ν_k, the reaction propensities $a_k(x;\theta)$, the value of k_M and the initial probability distribution $p(x,0|k_M)$ determine the probabilities of having any possible combination of molecule counts x present in the cell at any time t, i.e. they determine $p(x,t|k_M)$ for all $x \in \mathbb{N}^3$ and $t > 0$. Solving (15) for $p(x,t|k_M)$ is, however, usually very difficult and would anyway only give the probabilities for a cell with a specific value of k_M and not the population distribution.

By integrating (15) over all possible values of k_M with respect to the probability measure P_{k_M} and simultaneously multiplying by different polynomials in x and summing over all possible values of x, we can derive a system of population moment equations from (15) (see reference [8]) that depends on those rate parameters that are fixed for all cells in the population and on the moments of the distribution P_{k_M}. Since we assumed that P_{k_M} is an unknown gamma distribution parametrized by its mean and variance, we obtain a system of population moment equations that depends on the parameter vector $\tilde{\gamma} = [M_{k_M}\ V_{k_M}\ k_P\ k_F\ c_M\ c_P\ d_r\ h]^\top$

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\mu}(t;\tilde{\gamma}) = A(\tilde{\gamma},u(t;\tilde{\gamma},L))\,\tilde{\mu}(t,\tilde{\gamma}) + B(\tilde{\gamma},u(t;\tilde{\gamma},L)), \tag{16}$$

where L is the light-induction pattern and $\tilde{\mu}(t;\tilde{\gamma})$ is a vector which comprises moments up to a desired order (in our case four) of the joint distribution of $X(t)$ and the parameter k_M (the tilde here serves to refer to moments of the entire joint distribution whereas the symbol μ will be used later on to refer to moments of a marginal distribution which can be associated with the measured fluorescence distribution). The matrix $A(\tilde{\gamma},u(t;\tilde{\gamma},L))$ is determined by the model, the parameters $\tilde{\gamma}$ and the light induction pattern. For our model, (16) is a system of 65 non-linear ordinary differential equations for the moments up to order four that is closed in the sense that it does not depend on moments of higher order. Hence, its solution can be computed numerically using standard solvers for ordinary differential equations. For non-closed moment systems, (16) would have to be replaced by an approximate closed system using some approximation technique [9, 10].

S.2.3 Measurement error model and moments of the fluorescence intensity distribution

From the solution of (16) we can extract the mean $\mu_1^F(t;\tilde{\gamma}) \in \mathbb{R}_0^+$ and the centered moments up to order four $\mu_2^F(t;\tilde{\gamma}) \in \mathbb{R}_0^+, \mu_3^F(t;\tilde{\gamma}) \in \mathbb{R}, \mu_4^F(t;\tilde{\gamma}) \in \mathbb{R}_0^+$ of the marginal distribution $p_t^F(\cdot;\tilde{\gamma})$ of the amount of fluorescent protein $F(t)$. To compare this model output with the measurements we first need to convert these moments into fluorescence intensity units. We assume that each molecule emits a deterministic but unknown amount of fluorescence. This means that there exists an unknown scaling parameter $r \in \mathbb{R}^+$ by which we can multiply the molecule count output of the model to obtain the fluorescence intensity $I(t) \in \mathbb{R}_0^+$, i.e. $I(t) = rF(t)$. Since r is unknown it has to be included as an unknown model parameter leading to the final 9-dimensional parameter vector $\gamma = [\gamma_1 \cdots \gamma_9]^\top = [M_{k_M}\ V_{k_M}\ k_P\ k_F\ c_M\ c_P\ d_r\ h\ r]^\top$. The mean $\mu_1^I(t;\gamma)$ and the centered moments up to order four $\mu_2^I(t;\gamma), \mu_3^I(t;\gamma), \mu_4^I(t;\gamma)$ of the distribution $p_t^I(\cdot;\gamma)$ of $I(t)$ can then be obtained as follows:

$$\begin{aligned}
\mu_1^I(t;\gamma) &= r\mu_1^F(t;\tilde{\gamma})\\
\mu_2^I(t;\gamma) &= r^2\mu_2^F(t;\tilde{\gamma})\\
\mu_3^I(t;\gamma) &= r^3\mu_3^F(t;\tilde{\gamma})\\
\mu_4^I(t;\gamma) &= r^4\mu_4^F(t;\tilde{\gamma}).
\end{aligned} \tag{17}$$

At every measurement time t, the collected data are the measured fluorescence intensities of the N cells of the sample, which we denote as $\{y_n(t)\}_{n=1}^N$. We assume that these measurements are affected by additive noise terms due mainly to two different sources: autofluorescence artifacts and technical errors. Accordingly, each measurement $y_n(t)$ can be decomposed as follows:

$$y_n(t) = rf_n(t) + e_n^B + e_n^T$$

where $rf_n(t)$ is the quantity of interest, that is a realization of the random fluorescence intensity $I(t) = rF(t)$ due to the fluorescent protein $F(t)$, e_n^B is a particular autofluorescence artifact and e_n^T is a particular realization of the technical noise. In our model we assume that both the autofluorescence and technical errors are independent of the gene expression process and that they are realizations of two random variables whose distributions do not depend on time. More analytically, we assume that every cell is characterized by a different value of autofluorescence e_n^B and that the measurement of every cell is affected by a different realization e_n^T of the technical noise. Since in every sample we measure different cells, we can assume that there exists a unique random variable A such that for every measurement time t the elements of $\{e_n^B + e_n^T\}_{n=1}^N$ are independent realizations of A. The random process $Y(t)$ of which the real measurements are taken is then given by

$$Y(t) = I(t) + A = rF(t) + A.$$

We assume that at time $t = 0$ no fluorescent protein is present in the cells. Consequently, any measured fluorescence signal at $t = 0$ stems from A and we can estimate the distribution $p^A(\cdot)$ of A, its mean μ_1^A and the centered moments up to order four $\mu_2^A, \mu_3^A, \mu_4^A$ from the realizations of the fluorescence intensity $Y(0)$ measured in the N cells of the sample at time 0. Since $p^A(\cdot)$ is the same for all experiments this estimation can also be performed using measurements at $t = 0$ obtained from multiple different experiments. Accordingly, we assume that data from sufficiently many cells is used in this estimation such that the statistical error of the estimates is negligible, i.e. that the moments $\mu_1^A, \mu_2^A, \mu_3^A, \mu_4^A$ can be computed from the data without any significant error. This allows us to perform the final step that is required to align the model predictions with the real measurements: we need to convolute the distribution $p_t^I(\cdot;\gamma)$ of $I(t)$, predicted by the model, with $p^A(\cdot)$ and obtain distributions $p_t^Y(\cdot,\gamma)$ of $Y(t) = I(t) + A$ which are compatible with the measured distributions. To be precise, we only require the mean $\mu_1(t;\gamma)$ and the centered moments up to order four $\mu_2(t;\gamma), \mu_3(t;\gamma), \mu_4(t;\gamma)$ of $p_t^Y(\cdot,\gamma)$ for our sequential experiment design procedure (the subscript Y is dropped in the moments notation for simplicity). These can easily be obtained from the moments up to order 4 of $I(t)$ and A, that is $\{\mu_i^I(t,\gamma), \mu_i^A\}_{i=1}^4$. Consequently, we do not actually have to perform the full convolution and we also do not require the full distributions $p_t^I(\cdot,\gamma)$ and $p^A(\cdot)$. This is important because the full distribution $p_t^I(\cdot,\gamma)$ cannot be computed from the population moment equations. For the first two moments we obtain

$$\mu_1(t;\gamma) = \mu_1^I(t;\gamma) + \mu_1^A,$$
$$\mu_2(t;\gamma) = \mu_2^I(t;\gamma) + \mu_2^A,$$

therefore the effect of the measurement noise A is to introduce a constant offset on the mean and variance of $I(t)$.

Finally, notice that the data, $\mu_1(t_s)$ and $\mu_2(t_s)$, to which we compare the model predictions are *estimated* from the measurements $\{Y_n(t_s)\}_{n=1}^N$, that is

$$\hat{\mu}_1(t_s) = \frac{1}{N}\sum_{n=1}^{N} Y_n(t_s), \tag{18}$$
$$\hat{\mu}_2(t_s) = \frac{1}{N}\sum_{n=1}^{N} (Y_n(t_s) - \hat{\mu}_1(t_s))^2.$$

Regarding the measurements $\{Y_n(t_s)\}_{n=1}^N$ as random variables, the estimated mean and variance are random quantities as well, whose distribution depends on the distribution of $Y(t_s)$ and on the number of samples taken. According to the central limit theorem, for large values of N, the distribution of the random vector $\hat{\mu}(t_s) := [\hat{\mu}_1(t_s)\ \hat{\mu}_2(t_s)]^\top$ is approximately Gaussian with mean and variance that depend on the moments up to order 4 of $Y(t)$ [8]. The data distribution predicted by the model is therefore

$$\begin{aligned} \mathbb{E}\left[\hat{\mu}(t_s;\gamma)\right] &= \left[\mu_1(t_s;\gamma)\ \mu_2(t_s;\gamma)\right]^\top, \\ \mathrm{Var}\left[\hat{\mu}(t_s;\gamma)\right] &= \frac{1}{N}\left[\begin{array}{cc} \mu_2(t_s;\gamma) & \mu_3(t_s;\gamma) \\ \mu_3(t_s;\gamma) & \mu_4(t_s;\gamma) - \frac{N-3}{N-1}\mu_2(t_s;\gamma)^2 \end{array}\right]. \end{aligned} \tag{19}$$

S.2.4 Noise properties of the system

To summarize, our model encodes three main sources of variability. Firstly, we assume the presence of time-independent and process uncorrelated measurement noise, which produces an offset in the predictions of the fluorescence intensity mean and variance. Secondly, since we estimate means and variances from a finite sample, the data is approximately normally distributed around the predicted means and variances according to formula (19). Thirdly, variability coming for example from the fact that the cells are not identical, is allowed in our model through the parameter k_M, which models cell heterogeneity. Further noise sources could be added by allowing additional parameters to be heterogeneous among the population, or as suggested in [6], by including simultaneous events (e.g. bursts in the protein production) or more complicated non-constant models for the variability in the rates.

In order to judge whether our model can explain the variability observed in the data, we considered the experiment in Figure 4 of the main text. Specifically, we used the data shown in Figure 4 (black dots) to identify the vector of parameters $\hat{\gamma}^{F4}$ that produces the best possible fit for this particular experiment, that is its MAP estimator. In Figure S.2 we show the predictions of mean and variance obtained using $\hat{\gamma}^{F4}$, together with their uncertainty coming from our noise model.

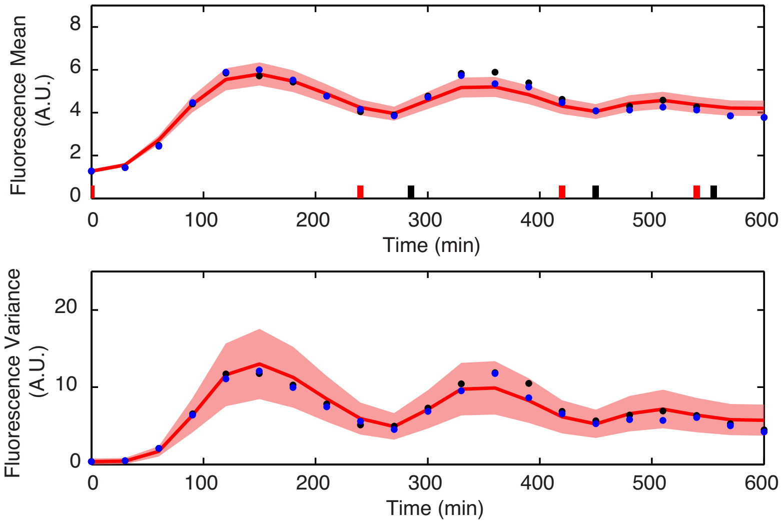

Figure S.2: Two replicates of the experiment in Figure 4 of the main text (data are shown as black and blue dots respectively). The red lines are the simulated mean and variance computed using the model with parameters $\hat{\gamma}^{F4}$. The shaded region shows the 99% confidence region coming from the finite sampling Gaussian error distribution computed in formula (19). The additive noise terms due to autofluorescence artifacts and technical errors appear in these plots as a time invariant additive mean μ_A and variance σ_A^2 which become evident at time 0 where the real process mean and variance should otherwise be zero.

We then performed a replicate of the experiment in Figure 4 and verified that the new collected data (blue dots), which were not used to identify $\hat{\gamma}^{F4}$, lie inside the predicted confidence region, hence validating our noise model. We note that in Figure S.2 it can be seen that the uncertainty in the predicted means and variances is not constant over time. This is due to the fact that the uncertainty in the estimates $\hat{\mu}(t_s) := [\hat{\mu}_1(t_s)\ \hat{\mu}_2(t_s)]^\top$ depends on the process distribution. For example, if we consider the estimate of the mean, we see from formula (19), that its variance is proportional to the variance of the fluorescence intensity. Therefore we can expect that whenever the process has high variance, the estimate of the mean will also have large uncertainty.

As a final note to this, we would like to point out that it has to be expected that the variance of the fluorescence distribution (and hence the uncertainty in the estimated mean) increases whenever the mean increases, because the variance is not a scale free measure. On the other hand, one would expect the coefficient of variation to decrease with increasing mean, because a larger mean corresponds to larger molecule numbers, and this should lower the stochasticity in the individual chemical reactions. To better illustrate this point we compare in Figure S.3 the measured mean, variance and coefficient of variation in the two optimal experiments, together with their predictions according to the model with parameters $\hat{\gamma}_2$ (see Table S.5). To compute the coefficient of variation, we used the moments of $I(t)$, that is we first subtracted mean and variance of A from the measured moments. This is necessary because otherwise differences in the coefficient of variation that stem from larger molecule numbers might be overshadowed by the autofluorescence and technical noise. As expected we can see in Figure S.3 that whenever the mean increases the variance also increases but at the same time the coefficient of variation decreases.

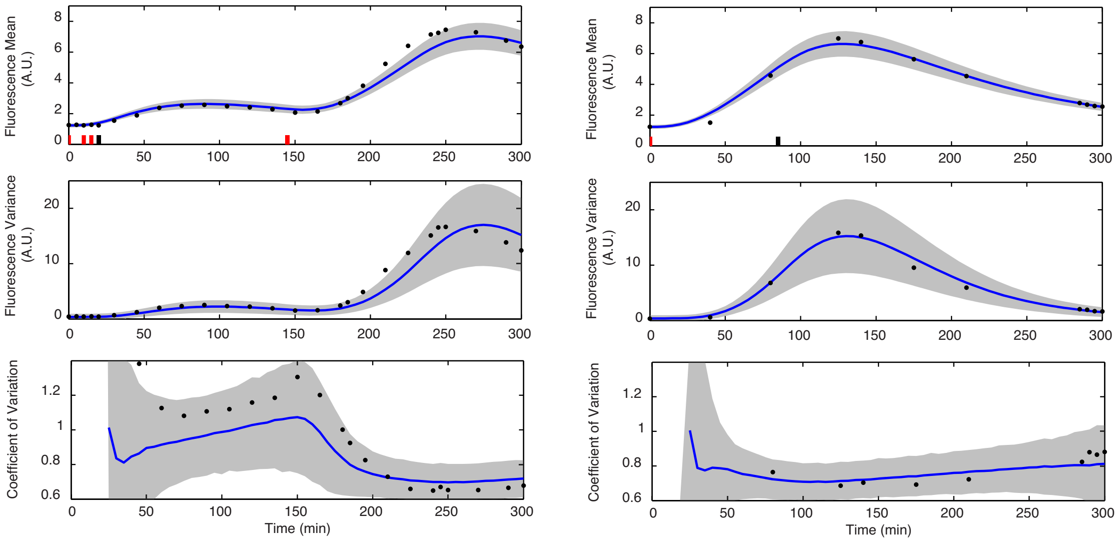

Figure S.3: Black dots: mean (top), variance (middle) and coefficient of variation (bottom) measured in the first (left) and second (right) optimally designed experiment. The solid blue line is the prediction obtained by simulating the model with parameters $\hat{\gamma}_2$ (see Table S.5). The grey shaded area is the confidence region due to finite sampling noise ($N = 2000$). Early measurement time points are not included for the coefficient of variation because at those time points the measured distribution is only marginally different from the noise distribution A, hence the computation of the coefficient of variation of the deconvoluted distribution $I(t)$ may be unreliable.

S.3 Optimal experiment design

S.3.1 Computation of the Fisher information matrix

The first thing that is required for the design of optimal experiments is a way of quantifying the utility of different experiments. The goal of our study is to estimate the model parameters γ with the highest possible precision. Accordingly, we quantify the utility in terms of the information that an experiment can provide about γ. One way of quantifying this information is through the computation of the Fisher information matrix $I(\gamma)$ [11]. The entries of the Fisher information matrix are given by

$$[I(\gamma)]_{k,l} = \mathbb{E}\left[\left(\frac{\partial}{\partial \gamma_k} \log p(\bar{Y};\gamma)\right)\left(\frac{\partial}{\partial \gamma_l} \log p(\bar{Y};\gamma)\right)\right], \tag{20}$$

where $p(\cdot\ ;\gamma)$ is the probability density of the data $\bar{Y}$ given that γ are the model parameters, and the expectation is taken with respect to all possible realizations of the data $\bar{Y}$. Thereby, the data $\bar{Y}$ is a collection of measurements taken from the process $Y(t)$ at multiple different time points and possibly even in different experiments. Accordingly, the expectation in (20) is taken with respect to a very high dimensional and complicated distribution.

In our case, an experiment $e = \{L_e, t_1, \dots, t_S\}$ is characterized by a light pattern L_e and measurement times $t_1, \dots, t_S$. At each of these time points, flow cytometry allows us to measure a realization of the random fluorescence intensities $Y_n(t_s),\ n = 1, \dots, N$ in N different cells of the population. The Fisher information matrix for such data can be approximately computed, using only the first four moments of the probability distribution, according to the formulas derived in references [7] and [12]:

$$I(\gamma, e) = \sum_{s=1}^{S} I^e_{t_s}(\gamma) \quad \text{where} \tag{21}$$

$$[I^e_{t_s}(\gamma)]_{k,l} = N \frac{\frac{\partial \mu^e_1(t_s)}{\partial \gamma_k} \frac{\partial \mu^e_1(t_s)}{\partial \gamma_l}}{\mu^e_2(t_s)} + N \frac{\left(\mu^e_2(t_s)\frac{\partial \mu^e_2(t_s)}{\partial \gamma_k} - \frac{\partial \mu^e_1(t_s)}{\partial \gamma_k}\mu^e_3(t_s)\right)\left(\mu^e_2(t_s)\frac{\partial \mu^e_2(t_s)}{\partial \gamma_l} - \frac{\partial \mu^e_1(t_s)}{\partial \gamma_l}\mu^e_3(t_s)\right)}{(\mu^e_2(t_s))^2\,(\mu^e_4(t_s) - (\mu^e_2(t_s))^2) - \mu^e_2(t_s)(\mu^e_3(t_s))^2},$$

where $\mu^e_1(t_s) = \mu^e_1(t_s;\gamma)$ is the mean and $\mu^e_i(t_s) = \mu^e_i(t_s;\gamma),\ i = 2, 3, 4$ are the centered moments of the distribution $p^Y_{t_s}(\cdot;\gamma)$ of $Y(t_s)$ for experiment e, as introduced in the previous section. To evaluate this formula, in addition to the moments themselves, partial derivatives of means $\mu^e_1(t_s;\gamma)$ and variances $\mu^e_2(t_s;\gamma)$ with respect to γ have to be computed from the model. For the parameters $\tilde{\gamma}$ the partial derivatives of the moments $\mu^{eF}_i(t_s;\tilde{\gamma}),\ i = 1, 2$ can be obtained by solving the population moment equations (16) with any solver for ordinary differential equations which also returns parameter sensitivities, such as CVODES of the SUNDIALS toolbox [13]. From this we can compute the partial derivatives of $\mu^e_i(t_s;\gamma),\ i = 1, 2$ because

$$\frac{\partial \mu^e_i(t_s;\gamma)}{\partial \gamma_k} = \frac{\partial (r^i \mu^{eF}_i(t_s;\tilde{\gamma}) + \mu^A_i)}{\partial \gamma_k} = r^i \frac{\partial \mu^{eF}_i(t_s;\tilde{\gamma})}{\partial \gamma_k}$$

for $i = 1, 2$ and $k = 1, \dots, 8$.

Because the moments $\mu^{eF}_i(t_s;\tilde{\gamma}),\ i = 1, 2$ do not depend on the parameter r, the partial derivatives of $\mu^e_i(t_s;\gamma),\ i = 1, 2$ with respect to $\gamma_9 = r$ can be computed as follows:

$$\frac{\partial \mu^e_1(t_s;\gamma)}{\partial r} = \frac{\partial (r \mu^{eF}_1(t_s;\tilde{\gamma}) + \mu^A_1)}{\partial r} = \mu^{eF}_1(t_s;\tilde{\gamma})$$

$$\frac{\partial \mu^e_2(t_s;\gamma)}{\partial r} = \frac{\partial (r^2 \mu^{eF}_2(t_s;\tilde{\gamma}) + \mu^A_2)}{\partial r} = 2r\mu^{eF}_2(t_s;\tilde{\gamma}).$$

S.3.2 Sequential experiment design

The expectation in (20) is taken with respect to all possible realizations of the data. Accordingly, the Fisher information matrix does not depend on any measurements and we can use it to evaluate the utility of different experiments before they are performed. This means that we can search among all possible experiments for the one which can be expected to provide the most information about the model parameters. In other words we can aim at solving the optimization problem

$$e^* = \underset{e \in \mathcal{E}}{\operatorname{argmax}} \{\det I(\gamma, e)\}, \tag{22}$$

where $I(\gamma, e)$ is the approximate Fisher information matrix defined in (21), $\mathcal{E}$ is the set of all possible experiments and an experiment $e = \{L_e, t_1, \ldots, t_S\} \in \mathcal{E}$ consists of a light-induction pattern L_e and S measurement times $t_1, \ldots, t_S \in [0, T_{max}]$ where T_{max} is the maximal duration of the experiment. The determinant $\det I(\gamma, e)$ in (22) provides one way of summarizing the information of an experiment in a scalar quantity that can be maximized. This is known as D-optimality. There exist many other optimality criteria, we refer the reader to [14, 15] for a detailed discussion.

It is important to underline that the Fisher information matrix depends on the values of the parameters γ which are to be estimated. These parameters are obviously unknown (otherwise performing an experiment for their identification would not be necessary). One way to overcome this problem is to invoke sequential experiment design. In sequential experiment design the parameters γ are replaced by their best currently available estimates $\hat{\gamma}$, so that a new experiment is designed using $\hat{\gamma}$ for the computation of the Fisher information matrix. The data collected in this experiment can then be used to improve the quality of the estimates and another experiment can be designed with the updated parameter estimates. In general, there is no guarantee that evaluating $I(\gamma, e)$ at estimated values $\hat{\gamma}$ will result in the design of informative experiments. Simulation studies [16], however, have shown that sequential experiment design often leads to good results, especially if many different experiments are needed to identify the model parameters.

Here, we follow such a sequential experiment design procedure. To be able to design the first experiment we used values found in [1] as initial estimates $\hat{\gamma}^0$ of the model parameters. Note that the index at the estimates refers to the number of experiments which have been performed to estimate the parameters and not to a component of the parameter vector. Specifically, we fixed $\hat{c}_M^0 = 0.03, \hat{c}_P^0 = 0.0066, \hat{k}_F^0 = 0.0419$ to the values found in [1] and $\hat{M}_{k_M}^0 = 0.9$ and $\hat{k}_P^0 = 0.22$ such that 30 mRNA molecules and 1000 protein molecules would on average be present at stationarity if the light signal $u(t)$ would be constantly at 1. Further, we chose $\hat{V}_{k_M}^0 = 0.27$ such that the coefficient of variation of k_M is 0.3 and the light parameters $\hat{d}_r^0 = 0.0155, \hat{h}^0 = 0.5$ such that the light signal drops slowly over time if no new light pulses are applied to the cells. Since no a priori knowledge of the scaling parameter r was available we set its estimate to $\hat{r}^0 = 10^3$.

The moments $\mu_i(t_s; \gamma),\ i = 1, \ldots, 4,\ s = 1, \ldots, S$ depend on the moments $\mu_i^A,\ i = 1, \ldots, 4$. Accordingly, also the Fisher information matrix depends on them and we have to choose initial estimates also for these moments if we want to design an experiment before any data is collected. We chose $\mu_1^A = \frac{4}{3}$ such that the application of a single red light pulse at time zero leads approximately to a 7-fold increase in the average fluorescence intensity according to the above parameters, $\mu_2^A = 1/2$ such that the coefficient of variation of $p^A(\cdot)$ is 0.5 and $\mu_3^A = 0$ and $\mu_4^A = 3(\mu_2^A)^2$ according to a Gaussian distribution. To design the first experiment we then have to solve the optimization problem

$$e^* = \underset{e \in \mathcal{E}}{\operatorname{argmax}} \left\{\det I(\hat{\gamma}^0, e)\right\}, \tag{23}$$

where the dependence of $I(\hat{\gamma}, e)$ on the moments of $p^A(\cdot)$ has been omitted in the notation.

It could be argued that a better strategy would be to start the sequential experiment design iteration with an experiment (e.g. randomly chosen) and not with computations which necessarily have to be performed with rather arbitrary initial guesses of the parameters. Another option would be to use a measure of information which does not depend on specific values of the parameters and instead evaluates the information with respect to a prior distribution on the parameters. For instance, Bayesian experiment design strategies [14] or the approach proposed in [7, 17] could be used. Such experiment design strategies, however, have a computational cost that is usually much larger than the cost of the approach used here. In addition, if no prior knowledge of the parameters is available, the choice of a prior distribution is no less arbitrary than the choice of initial point estimates. We do not investigate such strategies here, but in Section S.7.2.1 we show that the first experiment designed by locally evaluating the Fisher information at the parameter values $\hat{\gamma}^0$ leads to better results than a random or an experience-based experiment.

The second experiment is designed using the maximum a posteriori (MAP) estimates $\hat{\gamma}^1$ that are obtained from the data collected in the first experiment as described in Section S.4. An important thing to note is that for the design of the second experiment the joint Fisher information matrix of the already performed first experiment and the yet to be determined second experiment has to be computed. Since the data from the second experiment is statistically independent of the data from the first experiment, the joint information can be obtained by summing the Fisher information matrices of the two experiments, similar to the summation over the different measurement time points in (21). If only the Fisher information matrix of the second experiment alone would be used for the design, it would be likely that an experiment which is similar to the first would be designed (since the only difference to the design of the first experiment in (23) would be that the Fisher information matrix is evaluated at different parameter values). If, on the other hand, the joint information is used, one can expect that an experiment that adds new information and in some sense complements the already performed experiment is designed. Consequently, to design the second experiment we have to solve

$$e_2^* = \underset{e_2 \in \mathcal{E}}{\operatorname{argmax}} \Big\{ \det I(\hat{\gamma}^1, \{e_1^*, e_2\}) \Big\}, \tag{24}$$

where e_1^* is the first optimal experiment and $\hat{\gamma}^1$ are the maximum a posteriori estimates that are obtained from the data collected in the first optimal experiment.

According to the proposed sequential experiment design procedure one should design additional experiments, as described above, until the parameter posterior distribution and the predictive distributions for new experiments are sufficiently tight. We note that sequential experiment design allows one to distinguish whether poor model performances are due to lack of informativeness in the inference experiments or to structural deficiencies of the model. Specifically, poor performances of the model after several rounds of the proposed iterative scheme can be used as an indicator for the latter case.

S.3.3 Description of the algorithm for determining the most informative experiment

To search the set of possible light-induction patterns and measurement times we used a custom-designed algorithm similar to the one used in [7]. First, in order to simplify the optimization problem (22) we decoupled the search for the best light-induction pattern from the search for the best measurement times. Specifically, we fixed $T_{max} = 300$min as maximally allowed duration of the experiments and $S = 10$ as the maximal number of measurement times (see Section S.3.4). For a start, we placed these 10 measurement times equally spaced in the time interval $t \in [0, T_{max}]$ (i.e. every half hour) and then searched for the light-induction pattern L that maximizes the information for these measurement times. To perform this search we gridded the time axis into 5 minute intervals and only allowed light pulses to be placed on this grid. This simplifies the optimization problem and is also convenient for the experiments because the light pulses have to be manually administered to the cells and cannot be placed arbitrarily close to each other. Further, we added the constraint that each light-induction pattern must have a red light pulse at $t = 0$, because this initializes the gene expression and any experiment which does not have such a light pulse is effectively of a shorter duration than 300min and cannot be optimal. We then used a stochastic search in which, starting from an arbitrary initial light-induction pattern, new experiments are proposed randomly either by adding a new light pulse (randomly either red or far-red) at a random grid point, by removing an existing light pulse, or by shifting the existing light pulses to neighboring grid points as detailed in the following.

- If the light-induction pattern has only the one fixed red light pulse at $t = 0$, the algorithm adds a new light-pulse at a random grid point according to a uniform distribution and chooses it to be either red or far-red with equal probability.
- If the light-induction pattern already has more than the one light pulse, one of three possible actions is chosen.
    1. With probability 0.05 a new light pulse is added. This light pulse is placed on the grid according to a uniform distribution over all the grid points which do not already have a light pulse. Then, the new pulse is chosen either as red or far-red. Thereby, if the last previous light pulse in the light-induction pattern is far-red, the new light pulse is automatically chosen as red. This prevents the placements of subsequent far-red light pulses which do not change the experiment because far-red light sets the light signal $u(t)$ to zero and it does not have any effect to set the signal to zero if it is already at zero. If the last previous light pulse in the light-induction pattern is red, the new light pulse is chosen either as red or far-red with equal probability.
    2. With probability 0.05 an existing light pulse is removed at random.
    3. With probability 0.9 the existing light pulses are adjusted. Thereby, with equal probability each existing light pulse is either moved to a neighboring grid point that currently does not have a light pulse or left where it is.

The population moment equations (16) and formula (21) are then used to determine the Fisher information matrix for the new light-induction pattern and the new pattern is accepted with probability 1 if the information provided by the new experiment is larger than for the previous light-induction pattern and with a probability that depends on the difference in information if the information provided by the new experiment is smaller than for the previous light-induction pattern. We ran this search for 10000 iterations, recorded the information for each light-induction pattern and determined the light-induction pattern that leads to the highest information.

Subsequently, we used this light-induction pattern to sequentially place the measurement times. Specifically, for the first designed experiment we hand-placed two measurements, one at $t = 300$min and the other at $t = 150$min. This is necessary because with only two measurements the model parameters are so badly identifiable that the determinant of the Fisher information matrix is essentially zero for all possible placements of the measurement times, i.e. according to this measure of information all measurement times are equally bad. The remaining 8 measurement times in the first experiment were then added in a sequential fashion: first we gridded the time axis into one minute intervals, then for all the grid points we determined the information that would be obtained by adding a new measurement at this grid point to the already placed measurements and finally we placed the new measurement at the grid point which lead to the maximal information. Thereby, we added the constraint that no measurement is placed before time $t = 10$min because one cannot expect that the maturation of the fluorescent reporter is fast enough to allow one to detect protein production before that time. Further, we added the constraint that any new measurement must be placed at least 5min apart from all the other measurements to ensure that taking all measurements at the computed time points is practically feasible.

To conclude this section we note that the procedure for designing the first and second experiment, with this algorithm, is essentially equivalent with the difference that for the design of the second experiment the joint Fisher information of the already performed first experiment and the potential second experiment has to be computed and its determinant has to be maximized with respect to the second experiment only (see SI Section S.3.2). This also means that hand-placing the first two measurement times for the second experiment is not necessary because the joint Fisher information matrix always involves the 10 measurements of the first experiment and its determinant is always sufficiently far from zero.

S.3.4 A discussion of the design choices

There are two main design choices we made in the search for the experiment that maximizes (22). Firstly, to reduce the computational complexity of the optimization algorithm, we decoupled the search for the optimal light induction pattern from the placement of the measurement times. Specifically, in the first step of the optimization procedure we search for the optimal light induction pattern, assuming equally spaced measurements. The main motivation for this choice (apart from the computational simplification) was that we expected that measurements taken every half hour should suffice to capture the dynamics of each experiment. Therefore experiments that rank among the best ones when the measurement times are equally spaced should be nearly optimal also for the original optimization problem (22). Nonetheless, since we are restricting the class of possible experiments to $\mathcal{E}^1 := \{e \in \mathcal{E} \mid t_i = i \cdot 30 \text{ min}\}$, the light pattern of the optimal solution $\tilde{e}_1^* \in \mathcal{E}^1$ of this first step is not guaranteed to coincide with the one of the optimal experiment $e^* \in \mathcal{E}$, solution of (22). Since the optimization problem (22) restricted to the class $\mathcal{E}^1$ is still very difficult to solve, we opted for a randomized maximization algorithm. To ensure that our randomized search actually finds the optimum of the simplified problem, we ran the search for the light-induction pattern several times starting from different initial light induction patterns. In each case the same light induction pattern was returned confirming the effectiveness of our approach. Once the light pattern has been fixed, in the second step of our procedure, we search for the best measurement times in the restricted class $\mathcal{E}^2 := \{e \in \mathcal{E} \mid L_e = L_{\tilde{e}_1^*}\}$. This second optimization problem was solved sequentially by enumeration on a very fine time grid, which means that for any practical purposes the used measurement times are the best times that can be found using a sequential procedure for their placement. Overall, the final experiment $\tilde{e}_2^* \in \mathcal{E}^2$ is not guaranteed to exactly coincide with $e^* \in \mathcal{E}$, however for the reasons detailed above, we believe that the information provided by $\tilde{e}_2^*$ should be only marginally smaller than the information provided by e^*. Since a small difference in the provided information should not lead to a large difference in the precision of the parameter estimates, we believe that the computational advantages of the two step procedure outweight the loss of optimality. Nonetheless, it should be noted that our use of the term "optimal experiment" should not be interpreted as optimal in the set $\mathcal{E}$ but rather as optimal according to our simplified two step version of the problem. We conclude this discussion by noting that, for more complicated systems, even the proposed two step procedure might be computationally too expensive. If this is the case, available information about the system can be used to further restrict the class of experiments $\mathcal{E}$. For instance, one could restrict $\mathcal{E}$ by only allowing red light, far-red light and measurements to be placed in an alternating fashion. This would simplify the optimization problem but it might exclude more informative experiments, such as O1, that contain sequences of consecutive red light pulses and measurements placed both after red light pulses (giving information about parameters involved in protein production) and after far-red light pulses (giving information about parameters involved in protein degradation).

The second main design choice was to fix $T_{max} = 300$min as the maximal duration of the experiments and $S = 10$ as the maximal number of measurement times. The choice of measuring on average every half hour was made mainly for experimental convenience, but also because we believed that this would provide a sufficiently fine time resolution to observe the dynamics of the system. The choice of the experiment length, on the other hand, is more subtle. It is clear that longer experiments with more data samples lead to more information (and hence more precise parameter estimates), however since the design of the optimal experiments depends on the current estimates of the parameters it might be preferable to update these estimates as often as possible (hence voting for short experiments). Accordingly, we decided to fix a duration of the experiments which is as short as possible while still allowing us to excite the system properly. Based on our a priori knowledge of the system we estimated 5 hours to be a good compromise.

Note that *a priori*, without knowing the parameters of the system, there is no way to determine the optimal experiment length. It is, however, possible to test whether the choice made is justifiable *a posteriori*. For example, using the obtained parameter estimates, one can compare the informativeness of the performed two optimal 5h experiments to a single 10h experiment. More in detail, we can compare the information of the following set of experiments

1. $\{O_{10h}^{\hat{\gamma}^0}\}$: a single optimal 10 hours experiment, with 20 optimally placed measurement times, designed using the initial parameter estimates $\hat{\gamma}^0$;
2. $\{O1, O2\}$: the two performed 5 hours experiment designed respectively using $\hat{\gamma}^0$ and $\hat{\gamma}^1$;
3. $\{O_{10h}^{\hat{\gamma}^2}\}$: a single optimal 10 hours experiment, with 20 optimally placed measurement times, designed using the final parameter estimates $\hat{\gamma}^2$.

By using the final parameter estimates $\hat{\gamma}^2$ to compute the information of all these experiments, that is the best available estimate of the real parameters and hence of the real information of the experiments, we found that

$$\det\left(I(\hat{\gamma}^2, O_{10h}^{\hat{\gamma}^0})\right) < \det\left(I(\hat{\gamma}^2, \{O1, O2\})\right) < \det\left(I(\hat{\gamma}^2, O_{10h}^{\hat{\gamma}^2})\right).$$

This means that our 5h iterative experiment design procedure was a good strategy as it provided more information than what could have been obtained with the same experimental effort from one 10h experiment. From the second inequality we can see that the failure of the 10h approach is not due to the fact that such experiments cannot provide sufficient information, but rather to the fact that good 10h experiments cannot be found without good estimates of the model parameters.

Finally, it is important to notice that even though the time and number of measurements for a single experiment are fixed, the total combined duration and the total number of measurements of all identification experiments can be adjusted in the process of the iterative identification (in our case in multiples of 5h and 10 sampling times) by terminating the iterative procedure at the iteration where sufficient information is available.

S.4 Parameter Identification

S.4.1 Moment-based inference

Consider a characterization experiment e_d. To compute posterior distributions for the model parameters from the measured data we used a Bayesian moment-based inference scheme [8]. In particular, we used the means $\hat{\mu}_1^{ed} = \left[\hat{\mu}_1^{ed}(t_1)\dots\hat{\mu}_1^{ed}(t_S)\right]$ and variances $\hat{\mu}_2^{ed} = \left[\hat{\mu}_2^{ed}(t_1)\dots\hat{\mu}_2^{ed}(t_S)\right]$ of the measured fluorescence distributions as data. In flow cytometry experiments the cells are discarded after being measured. Consequently, measurements at different time points are statistically independent. Further, it is a direct consequence of the central limit theorem that the joint distributions of the measured means and variances $\hat{\mu}^{ed}(t_s) = \left[\hat{\mu}_1^{ed}(t_s)\ \hat{\mu}_2^{ed}(t_s)\right]^\top$, $s = 1, \dots, S$ at each time point of experiment e_d are approximately two-variate Gaussian distributions. Accordingly, the likelihood of the data $\mathcal{D} = \left(\hat{\mu}_1^{ed}, \hat{\mu}_2^{ed}\right)$ can be obtained as

$$p(\mathcal{D} \mid \gamma) = p\left(\hat{\mu}_1^{ed}, \hat{\mu}_2^{ed} \mid \gamma\right) = \prod_{s=1}^{S} p_{t_s}\left(\hat{\mu}_1^{ed}(t_s), \hat{\mu}_2^{ed}(t_s) \mid \gamma, L_{ed}\right), \quad \text{where} \tag{25}$$

$$p_{t_s}(\cdot, \cdot \mid \gamma, L_{ed}) = \mathcal{N}\left(\mu^{ed}(t_s;\gamma), \Sigma^{ed}(t_s;\gamma)\right), \quad \text{and}$$

$$\mu^{ed}(t_s;\gamma) = \left[\mu_1^{ed}(t_s;\gamma)\ \mu_2^{ed}(t_s;\gamma)\right]^\top,$$

$$\Sigma^{ed}(t_s;\gamma) = \frac{1}{N}\begin{bmatrix} \mu_2^{ed}(t_s;\gamma) & \mu_3^{ed}(t_s;\gamma) \\ \mu_3^{ed}(t_s;\gamma) & \mu_4^{ed}(t_s;\gamma) - \frac{N-3}{N-1}\left(\mu_2^{ed}(t_s;\gamma)\right)^2 \end{bmatrix}$$

where N is the number of measured cells [8]. It is important to note that the Gaussian distributions which define this likelihood only depend on moments of $p_t^Y(\cdot;\gamma)$ up to order four. Accordingly, the likelihood can be evaluated from the solution of the population moment equations, with light pattern L_{ed}, and the entire distribution $p_t^Y(\cdot;\gamma)$ is not required.

The posterior distribution can then be obtained using Bayes formula

$$p(\gamma \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \gamma)\, p(\gamma)}{p(\mathcal{D})}, \tag{26}$$

where $p(\gamma)$ is the parameter prior. We draw samples from this distribution using a Sequential Monte Carlo algorithm which is described in the next section. This gives a particle approximation of the posterior distribution which is the final result of the parameter inference. If point estimates $\hat{\gamma}$ of the model parameters are desired we can also extract the maximum a posteriori (MAP) estimates, i.e. the values of the parameters for which the probability density function of the posterior distribution attains its maximum, from this particle approximation.

Since the data collected in different experiments is statistically independent, it is straightforward to extend (25) and (26) to multiple experiments, $\mathcal{D} = \{(\hat{\mu}_1^{ed}, \hat{\mu}_2^{ed})\}_{d=1}^{D}$, by using the likelihood

$$p(\mathcal{D} \mid \gamma) = \prod_{d=1}^{D} \prod_{s=1}^{S} p_{t_s}\left(\hat{\mu}_1^{ed}(t_s), \hat{\mu}_2^{ed}(t_s) \mid \gamma\right), \tag{27}$$

where $\hat{\mu}_1^{ed}$ and $\hat{\mu}_2^{ed}$ are the means and variances measured in experiment $e_d,\ d = 1, \ldots, D$.

S.4.2 Description of the parameter inference algorithm

Analytic calculation of the parameter posterior distribution given the experimental data is infeasible in practise, as it depends on the solution of a large system of moment equations. We thus have to resort to approximation methods, such as Monte Carlo sampling, which can be used to generate a set of samples (termed *particles*) from the desired posterior. This particle approximation of the posterior can then be used to provide further approximations to posterior-related quantities, such as the posterior predictive distributions.

Sampling from the complex, high-dimensional parameter posteriors generated by a dynamic system is not a trivial computational problem [18], and therefore requires the use of sophisticated sampling algorithms. In this work we use a Sequential Monte Carlo (SMC) scheme [19], based on the idea of Annealed Importance Sampling [20], which has already been tested successfully in computationally complex parameter inference and model selection problems [21, 22].

Our SMC sampler is able to circumvent the problem of sampling from complex, high-dimensional posteriors by generating weighted samples from a sequence of distributions f_β which form a “bridge” from the (typically very diffuse) prior to the much more concentrated posterior. This sequence is generated according to the following “cooling” scheme:

$$f_{\beta_j}(\gamma) \propto p(\mathcal{D}|\gamma)^{\beta_j} p(\gamma), \text{ for } 0 = \beta_0 < \beta_1 < \cdots < \beta_J = 1.$$

At each ”cooling” step, the distribution f_{β_j} is used to perform importance sampling with $f_{\beta_{j+1}}$ as target distribution: starting from the sampled points of f_{β_j}, samples from $f_{\beta_{j+1}}$ are drawn using a Markov chain transition kernel K_j whose invariant distribution is $f_{\beta_{j+1}}$. This is the only requirement imposed on K_j, which can be selected to be arbitrarily complex.

Since the quality of the particle approximation of intermediate distributions tends to deteriorate with each cooling step, the Effective Sample Size (ESS) [19] of the particle population is monitored at each cooling step, and the particles are resampled when ESS $< 0.5Q$, where Q is the population size used to approximate the target distribution of each cooling step, that is the number of particles.

S.4.3 Parameter settings and implementation details

For all the posterior approximations in this work, we use a population of $Q = 8000$ particles, which are propagated using an adaptive annealing schedule [23]. That is, at each annealing step β_j, the next step β_{j+1} is chosen such that the effective sample size of the particle population at the next step, ESS_{j+1}, is equal to αESS_j, with $\alpha \in (0, 1)$, as suggested in [23]. Note that this is possible, thanks to the fact that ESS_{j+1} depends on the particle population at the j-th annealing step [19].

To sample from the distribution $f_{\beta_{j+1}}$, we use a Markov Chain Monte Carlo (MCMC) [24] kernel, consisting of 15 iterations of an Independent Metropolis-Hastings sampler [24], whose proposal distribution is defined through a Gaussian mixture density estimate obtained from the available particle population at step j.

For every parameter, we use a log-uniform prior distribution $p(\gamma_i)$ centered around the nominal parameter value and spanning m_i orders of magnitude, as given in Table S.4. Equivalently, we define the new parameters

$$\gamma_i' := \log_{10}\left(\frac{\gamma_i}{\gamma_{i,nom}}\right), \quad i = 1, \ldots, 9 \tag{28}$$

with uniform prior distribution $p(\gamma_i') \in \mathcal{U}\left[-m_i, m_i\right]$, and we use the above scheme to perform inference on γ'.

Table S.4: Nominal values and range m_i used in the inference algorithm

c_M	k_P	k_F	c_P	d_r	h	M_{k_M}	V_{k_M}	r^{-1}
0.0811	1868	0.0047	0.0134	0.0359	0.1067e-7	0.8230	0.01607	2500
2	2	2	2	2	2	2	2	1

S.5 Predictions of the outcome of new experiments

S.5.1 Computation of the posterior predictive distributions

The most straightforward approach for using the model to predict the response of the system to new light-induction patterns would be to solve the corresponding population moment equations with the MAP estimates $\hat{\gamma}$ as model parameters γ. Even though this is a valid approach, we will use it for example later on to determine the light-induction patterns for the control experiments (see Section S.8.2), it does not allow us to quantify how certain we can be about the model predictions. Consider, for instance, a case where the parameter posterior distribution is very flat such that its density function is only marginally larger at the MAP estimates than at other parameter values. In such a case predictions computed with parameters other than the MAP estimates would, according to our knowledge, be almost equally likely outcomes of the new experiment. In other words, predictions computed using only the MAP estimates do not take into account that the model parameters might be only badly identifiable from the data that was used to compute the MAP estimates. Accordingly, such predictions do not take into account that we might not be very certain that the MAP estimates are actually precise.

Due to the above considerations we decided to validate our model by using the complete posterior predictive distributions for the population means and variances instead of predictions computed with point estimates only. These distributions describe how likely different measurements of the moments $\hat{\mu}^{ev}(t_s) = [\hat{\mu}_1^{ev}(t_s)\ \hat{\mu}_2^{ev}(t_s)]^\top$, $s = 1, \ldots, S$, are in a new experiment e_v, given all the previously measured data $\mathcal{D}$. They can be computed according to

$$p_{t_s}^{pred}(\hat{\mu}^{ev}(t_s)) := p_{t_s}(\hat{\mu}^{ev}(t_s) \mid \mathcal{D}) = \int_\gamma p_{t_s}(\hat{\mu}^{ev}(t_s) \mid \gamma, L_{ev}) p(\gamma \mid \mathcal{D}) d\gamma, \tag{29}$$

where $p(\gamma \mid \mathcal{D})$ is the posterior distribution given data $\mathcal{D}$ and $p_{t_s}(\cdot \mid \gamma, L_{ev})$ is the distribution of the data $\hat{\mu}^{ev}(t_s)$ for the new experiment e_v given that γ are the model parameters and the light pattern

L_{ev} is applied. It is difficult to compute these predictive distributions exactly, but approximations can easily be obtained by replacing the integral over γ with a sum over samples $\gamma_q,\ q = 1, \ldots, Q$, drawn from the posterior distribution $p(\gamma \mid \mathcal{D})$. Recalling from Section S.4.1 that $p_{t_s}(\cdot \mid \gamma, L_{ev}),\ s = 1, \ldots, S$, are two-variate Gaussian distributions that can be computed from the solution of the population moment equations according to (25), we obtain Gaussian mixture approximations of the predictive distributions

$$p_{t_s}^{pred}(\cdot) \approx \frac{1}{Q}\sum_{q=1}^{Q} \mathcal{N}\left(\mu^{ev}(t_s;\gamma_q), \Sigma^{ev}(t_s;\gamma_q)\right),\ s = 1, \ldots, S. \tag{30}$$

It is straightforward to compute means and variances of these Gaussian mixture distributions. However, to compute the confidence regions shown in Figure 3 and Figure 4 in the main text, the entire predictive distributions are required. To obtain approximations of the entire predictive distributions, we took $Q = 4000$ samples $\gamma_q,\ q = 1, \ldots, Q$, from $p(\gamma \mid \mathcal{D})$. For each sample we solved the population moment equations and computed the Gaussian distributions $\mathcal{N}\left(\mu^{ev}(t_s;\gamma_q), \Sigma^{ev}(t_s;\gamma_q)\right)\ \ s = 1, \ldots, S,\ q = 1, \ldots, Q$. Finally, we took one sample from each of the Gaussian distributions. This gave us $Q = 4000$ samples for each of the predictive distributions $p_{t_s}^{pred}(\cdot),\ s = 1, \ldots, S$, which we used to compute the confidence regions shown in Figure 3 and Figure 4 in the main text.

S.5.2 Stochastic simulation of the model

The population moments equations can only be used to predict moments of the fluorescence distribution in new experiments. If predictions of the entire fluorescence distributions are required, other approaches have to be used. Directly computing these distributions using finite state projection-based approaches [25] is not possible due to the continuous distribution of the parameter k_M. A possibility would be to sample values of k_M from P_{k_M} and then approximately solve the conditional CME (Eq.15) for all the sampled values using finite state projection. This approach is, however, computationally expensive because the rather large amount of protein necessitates very large state truncations. An alternative is to sample values of k_M from P_{k_M} and then simulate the model with Gillespie's stochastic simulation algorithm [26] with the sampled values of k_M. While also this approach is computationally very expensive (simulating one 10h trajectory of the system may take up to 2min on a standard computer), it is the only approach that can give exact samples from the fluorescence distribution and, hence, we decided to use it to generate the results in Figure 3B in the main text.

Specifically, we generated 30000 trajectories of the system using the MAP estimates $\hat{\gamma}^2$, extracted the amount of fluorescent protein at the measurement time points $F(t_s),\ t_s = s \cdot 30\text{min},\ s = 1, \ldots, 20$ and converted the values to fluorescent units through multiplication with the final estimate of the scaling parameter $\hat{r} = (1.7171 \cdot 10^4)^{-1}$. This gives us 30000 samples from $I(t) = \hat{r}F(t)$ at the measurement times. To obtain samples from $Y(t) = I(t) + A$ that are comparable to the measured fluorescence intensity values, we randomly took 30000 samples from the measured fluorescence values at $t = 0$ in the two optimal experiments and added them to the samples from $I(t)$. The final comparison of simulated and measured distributions in Figure 3 in the main text was then performed using histograms with 50 bins.

S.6 Iterative characterization of the model

In this section we provide detailed results of the iterative identification procedure. Figure S.4, top row, shows some of the two-dimensional marginals of the posterior distribution that was obtained from the data collected in the first optimal experiment (Figure 2B in the main text). For this posterior distribution as well as for the following ones, the marginals according to all the possible pairs of parameters are reported for completeness at the end of the supplementary. It can be seen that for some parameters the marginals of the posterior distribution are tight, whereas for others they are very spread out. For instance, the parameters k_P and k_F are difficult to identify. The corresponding panel suggests that the reason for this is that the estimates of these parameters are highly correlated. Contrary to that, other parameters, for instance the degradation rates c_M and c_P, are identified with only small uncertainty. The MAP estimates $\hat{\gamma}^1$, which we extracted from this posterior distribution and used to design the second experiment, are reported in Table S.5.

Table S.5: Initial estimates $\hat{\gamma}^0$ of the model parameters and MAP estimates obtained from the first optimal experiment $\hat{\gamma}^1$ and from both optimal experiments $\hat{\gamma}^2$

	c_M	k_P	k_F	c_P	d_r	h	M_{k_M}	V_{k_M}	r^{-1}
$\hat{\gamma}^0$	0.03	0.22	0.0419	0.0066	0.0155	0.5	0.9	0.27	10^{-3}
$\hat{\gamma}^1$	0.0124	3159.9	0.0022	0.0465	0.2677	$1.3450 \cdot 10^{-10}$	0.1185	0.0045	$2.5831 \cdot 10^3$
$\hat{\gamma}^2$	0.0322	731.1953	0.0300	0.0114	0.2741	$1.7482 \cdot 10^{-10}$	0.1484	0.0080	$1.7171 \cdot 10^4$

Figure S.4, bottom row, shows how the marginals of the posterior distribution change if the data collected in the second optimal experiment (Figure 2C in the main text) are also used. It can be seen that for some parameters, for instance for k_F, the region of high posterior probability is significantly smaller than if only the first optimal experiment is used, hence additional certainty about the model parameters has been gained from the second experiment. The MAP estimates $\hat{\gamma}^2$ extracted from this posterior distribution are shown in Table S.5.

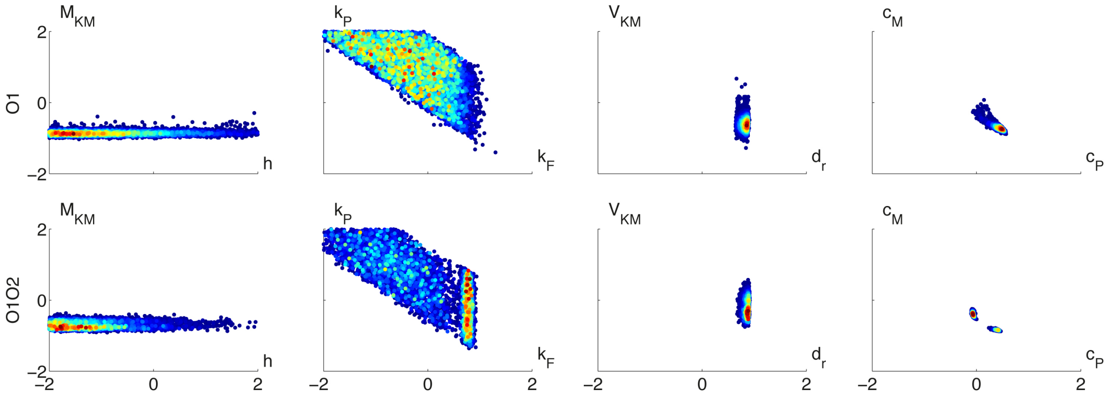

Figure S.4: Comparison of the posterior distributions computed from the data collected in the first optimal experiment O1 (top row) and from both the optimal experiments O1O2 (bottom row). The different panels show some of the two dimensional marginals of the full posterior distribution of the parameters γ', as defined in (28). The color is an index of the likelihood of each particle: blue for the particles with lower likelihood and red for the particles with higher likelihood.

S.7 Comparison of different experiments

S.7.1 Description of the performed random and experience-based experiments

In this section we describe the random and the experience-based experiments that are used in the main paper to quantify the advantage of the optimal characterization approach. To ensure that any difference in the posterior distributions and in the predictive distributions can really be attributed to the experimental approach we fixed the duration of these experiments and the number of measurement times to the same values that were used for the optimally designed experiments. Specifically, we chose

$T_{max} = 300$min and $S = 10$ for all experiments and placed the 10 measurements equally spaced in the time interval $[0, 300\text{min}]$, i.e. every half hour.

Random experiments. To obtain these experiments we randomly generated light-induction patterns. More in detail, we restricted the possible times to apply a light pulse to be every 5 min and, at every possible time, we applied a red/far-red pulse with probability 0.25 each. Note that since these experiments are designed without considering the model dynamics we allow consecutive far-red pulses, even though according to our model they are redundant.

The generated light-induction patterns and the data collected in these experiments are shown in Figure S.5. Initially, we generated two random experiments (panels R1 and R2) for a fair comparison to the two optimally planned experiments. Subsequently, we performed a third random experiment (panel R3) to investigate whether the lower information content of the random experiments can be compensated by performing more experiments.

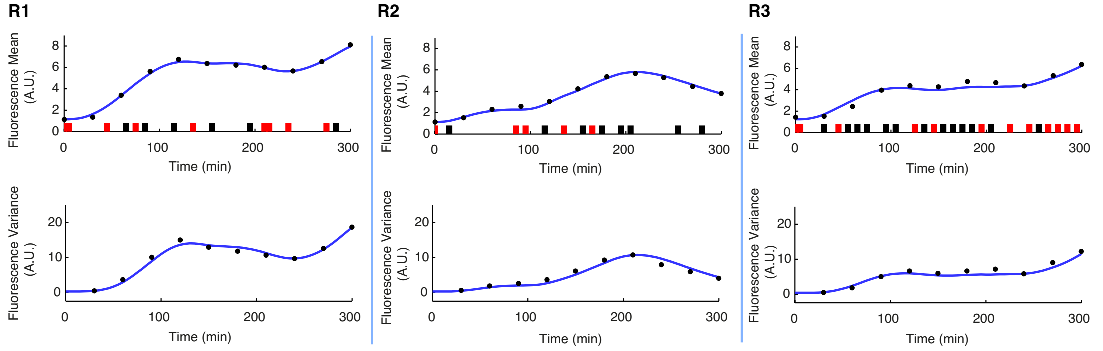

Figure S.5: Random characterization of the light-inducible gene expression circuit. (R1) Applied light-induction pattern and measured means and variances in the first random experiment (black dots). The blue line is the model output with the MAP estimates obtained from this first random experiment. (R2) Applied light-induction pattern and measured means and variances in the second random experiment. The blue line is the model output with the MAP estimates obtained from the first two random experiments. (R3) Applied light-induction pattern and measured means and variances in the third random experiment. The blue line is the model output with the MAP estimates obtained from the three random experiments.

Experience-based experiments. The model used in [1] was identified from the measured average fluorescence of 8 characterization experiments. Here, on the one hand, we exploit for each experiment the additional information provided by the measured variances. On the other hand, we use a model that contains more parameters than the simple one used in [1]. Hence, it is unclear whether more or less experiments are required for our study. The result that only two optimally designed experiments are enough to characterize the system, however, suggests that less than eight experiments might be sufficient.

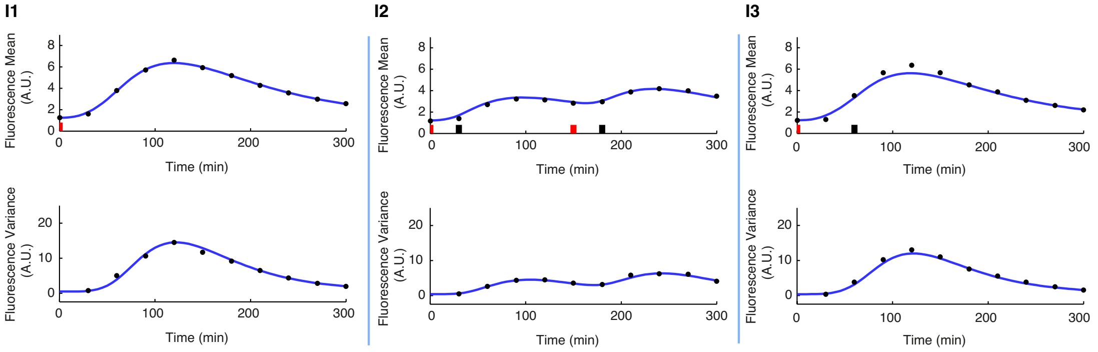

Figure S.6: Experience-based characterization of the light-inducible gene expression circuit. (I1) Applied light-induction pattern and measured means and variances in the first experience-based experiment (black dots). The blue line is the model output with the MAP estimates obtained from this first experience-based experiment. (I2) Applied light-induction pattern and measured means and variances in the second experience-based experiment. The blue line is the model output with the MAP estimates obtained from the first two experience-based experiments. (I3) Applied light-induction pattern and measured means and variances in the third experience-based experiment. The blue line is the model output with the MAP estimates obtained from the three experience-based experiments.

For a fair comparison to the two optimally planned experiments we had to choose two experiments out of the eight characterization experiments in [1]. There is no argument for preferring any two of the eight experiments. To reduce any possibly bias stemming from this choice we chose three experience-based experiments (one from each panel of Figure 1 in [1]) and compared the two optimally planned experiments to any combination of two of the three experience-based experiments. The light-induction pattens and the data collected in the three experience-based experiments are shown in Figure S.6.

S.7.2 Comparison of parameter posterior distributions and protein predictive distributions

S.7.2.1 After one performed experiment

In this section we compare the results from only the first optimal experiment to the results from one random experiment and to the results from one experience-based experiment. While our main comparison will be performed after two experiments, the comparison in this section is also educational because it allows us to determine whether it is reasonable to design an experiment before any real estimates of the model parameters are available.

Figure S.7 shows the comparison of some of all the possible two-dimensional marginals of the full posterior distribution computed from the first optimal experiment (Figure 2B in the main text), the random experiment shown in Figure S.5R1 and the experience-based experiment shown in Figure S.6I1.

It can be seen that the random experiment leads to tight marginals of the posterior distribution for some parameters but also to very wide distributions for others. Especially the parameters d_r and h that characterize the light signal cannot be identified very well from this experiment.

The differences between the first optimal experiment and the experience-based experiment are more subtle, but on a close inspection it can be noticed that some of the parameters (for instance h and V_{k_M}) can be identified with higher precision from the optimal experiment.

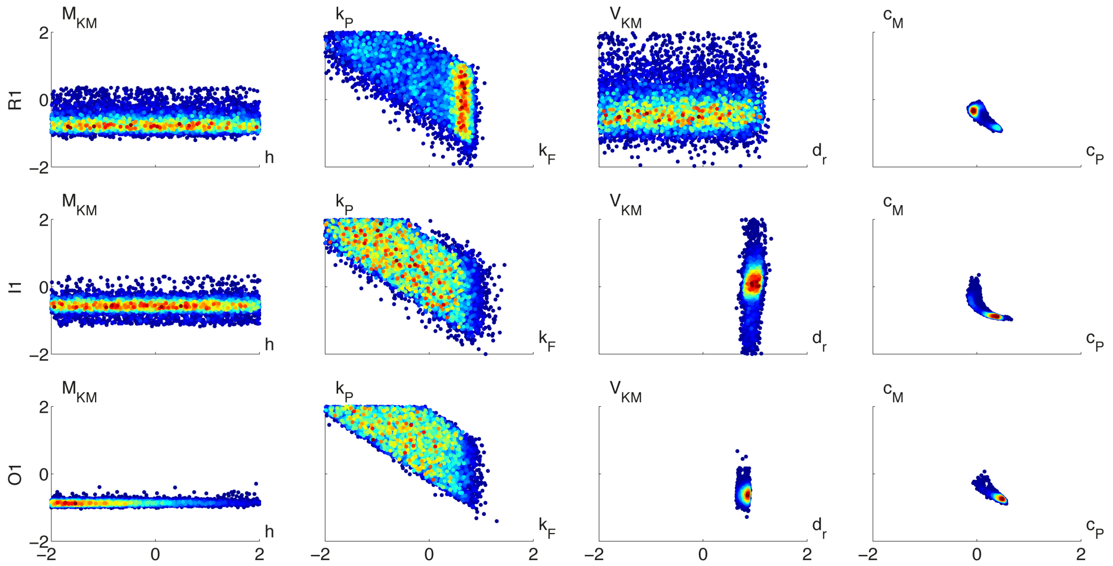

Figure S.7: Comparison of the posterior distributions computed from the data collected in the first random experiment R1 (top row), the first experience-based I1 (middle) and the first optimal experiment O1 (bottom row). The different panels show some of the two dimensional marginals of the full posterior distribution of the parameters γ', as defined in (28). The color is an index of the likelihood of each particle: blue for the particles with lower likelihood and red for the particles with higher likelihood.

Table S.6: **Comparison of optimal, random and experience-based after 1 experiment.** For each performed experiment the log of the mean likelihood of the measured data according to the three different posterior distributions is computed. The best model among O1, I1 and R1 is the one with highest expected likelihood and is denoted with a star.

	O1	I1	R1
Figure 5A	$-115.5^\star$	-232.3	-210.9
Figure 5B	$-224.3^\star$	-293.1	-262.1
Figure 5C	-223.1	-252.2	$-202.4^\star$
Figure 5D	-83.4	$-78.7^\star$	-127.5
Figure 3	$-111.4^\star$	-225.5	-193.6
Figure 4	-414.6	-11.5	$1.6^\star$

Because the differences in the posterior distributions are not easy to notice and also because we ultimately want to predict new experiments, we performed a second comparison using the posterior predictive distributions for all the control and validation experiments that we performed throughout our study. To summarize the quality of the predictions in a scalar quantity we computed the mean likelihood of the data collected in the control and validation experiments using the posterior predictive distributions of the first optimal, random and experience-based experiment. Table S.6 shows that the collected data was predicted best on average by the posterior predictive distribution of the first optimal experiment.

S.7.2.2 Comparison of two optimal and the random experiments

In this section we compare the results from the two optimal experiments to the results from two and three random experiments.

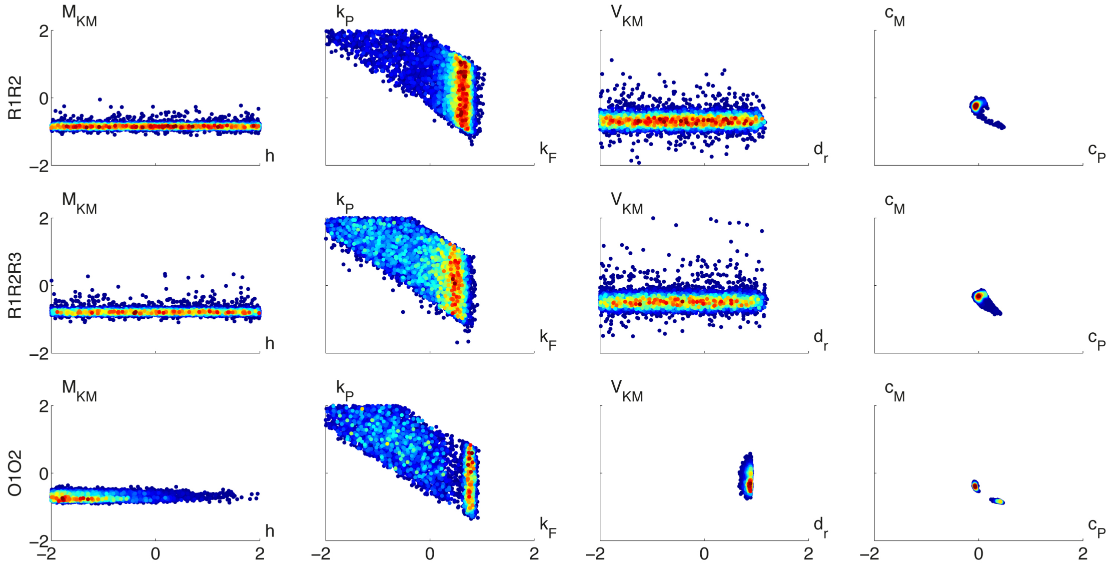

Figure S.8: Comparison of the posterior distributions computed from the data collected in the first two random experiment R1R2 (top row), in all the random experiments R1R2R3 (middle) and in the two optimal experiments O1O2 (bottom row).

Figure S.8 shows that the posterior distribution obtained from the two optimal experiments is much tighter, not only than the one obtained from the first two random experiments, but also than the one obtained from all three random experiments. In fact, almost no additional knowledge about the model parameters was gained from the third random experiment. The corresponding predictive distributions for the two validation experiments are shown in Figure S.9.

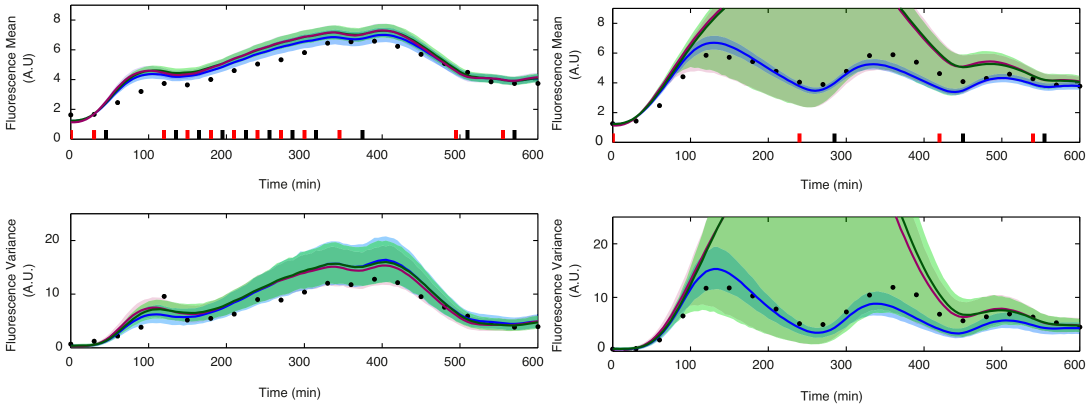

Figure S.9: Predictive distributions for the two validation experiments shown in the main text. The different colors correspond to the predictive distributions obtained with different experimental approaches. Green: with all three random experiments. Blue: with the two optimal experiments. Magenta: with two random experiments.

Figure S.9 shows that there is hardly any difference between the predictive distributions obtained from two and three random experiments. In particular, both are not informative for the second validation experiment, V2. Contrary to that, the predictive distributions obtained from the two optimal experiments are very tight and agree reasonably well with the measured data in both the validation experiments. A similar conclusion can be reached by investigating the differences of the mean likelihood of the data collected in all control and validation experiments, using the posterior predictive distributions identified from the two optimal experiments versus three randoms (Table S.7).

Table S.7: **Comparison of two optimal with three random experiments.** For each performed experiment the log of the mean likelihood of the measured data according to the different parameters posterior distributions is computed. The best model is the one with highest mean likelihood and is denoted by a star.

	O1O2	R1R2R3
Figure 5A	$-44.7^{\star}$	-161.3
Figure 5B	$-174.4^{\star}$	-252.6
Figure 5C	$-67.7^{\star}$	-164.0
Figure 5D	$-67.1^{\star}$	-68.2
Figure 3	$-28.8^{\star}$	-132.4
Figure 4	-69.6	$-59.9^{\star}$

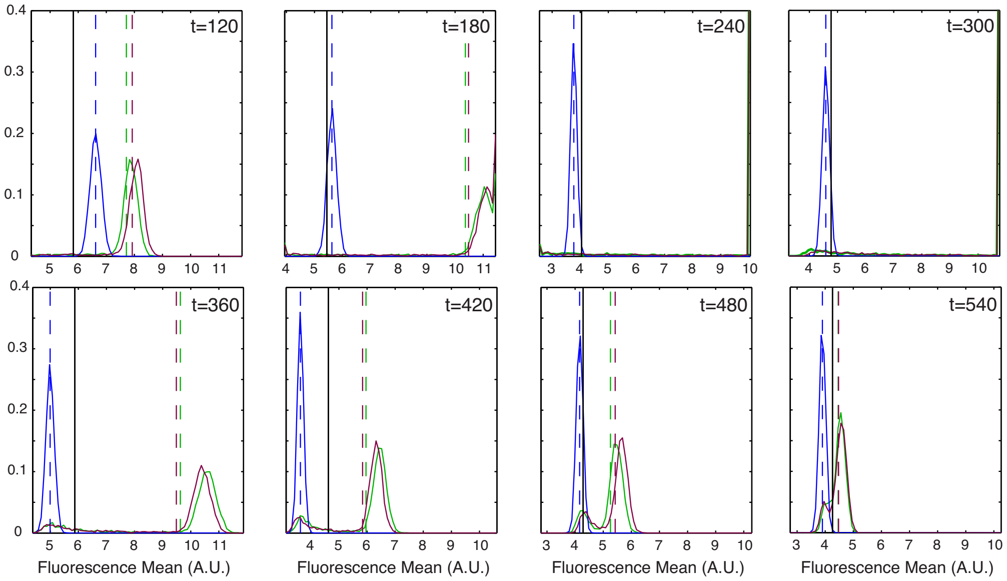

Figure S.10: Marginals of the predictive distribution for the validation experiment V2 at different time points. The different colors correspond to the predictive distributions obtained with different experimental approaches. Green: with all three random experiments. Blue: with the two optimal experiments. Magenta: with two random experiments.

One peculiar and seemingly unintuitive result in Table S.7 is that the mean likelihood for the validation experiment shown in Figure 4 is slightly larger for the parameter posterior distribution obtained from the three random experiments R1R2R3 than for the one from the two optimal experiments O1O2. In other words, the probability density of the predictive distribution from R1R2R3 evaluated at the measured data is larger than the corresponding quantity for O1O2. This seems to be in contrast to the confidence regions for the predicted means and variances given in Figure S.9, which show that the predictive distribution from the random experiments is more spread out than the one from the optimals.

A closer inspection of the shape of the predictive distributions at different time points (see Figure S.10), however, resolves this apparent contradiction. It can be seen that the predictive distributions obtained from the random experiments are bimodal with a large peak far away from the measured data and another smaller peak close to the measured data. Since these distributions are very spread, many different outcomes of the experiment can be explained with a small, but not negligible, probability. The predictive distributions obtained from O1O2, on the other hand, are unimodal with very small tails, i.e. there is almost no uncertainty about the predicted value. Consequently, for some time points, the measured data, while being quite close to the predicted values from O1O2, turned out to have a lower likelihood according to this distribution than according to the one from R1R2R3 (see Figure S.10 for example at 360 min).

Since the mean likelihood does not take into account the uncertainty about the prediction, looking only at this number one would therefore slightly prefer R1R2R3 to O1O2. It is however important to remember that the predictive distributions are used to predict new data: as can be seen in Figure S.9 and S.10 the most likely prediction from R1R2R3 is the one corresponding to the main peak which is not at all consistent with the measured data. For this reason the results of Table S.7 cannot be used alone but need to be complemented by other posterior predictive checks, such as the visual inspection of high-density regions in the predictive distributions.

S.7.2.3 Comparison of two optimal experiment and different pairs of experience-based experiments

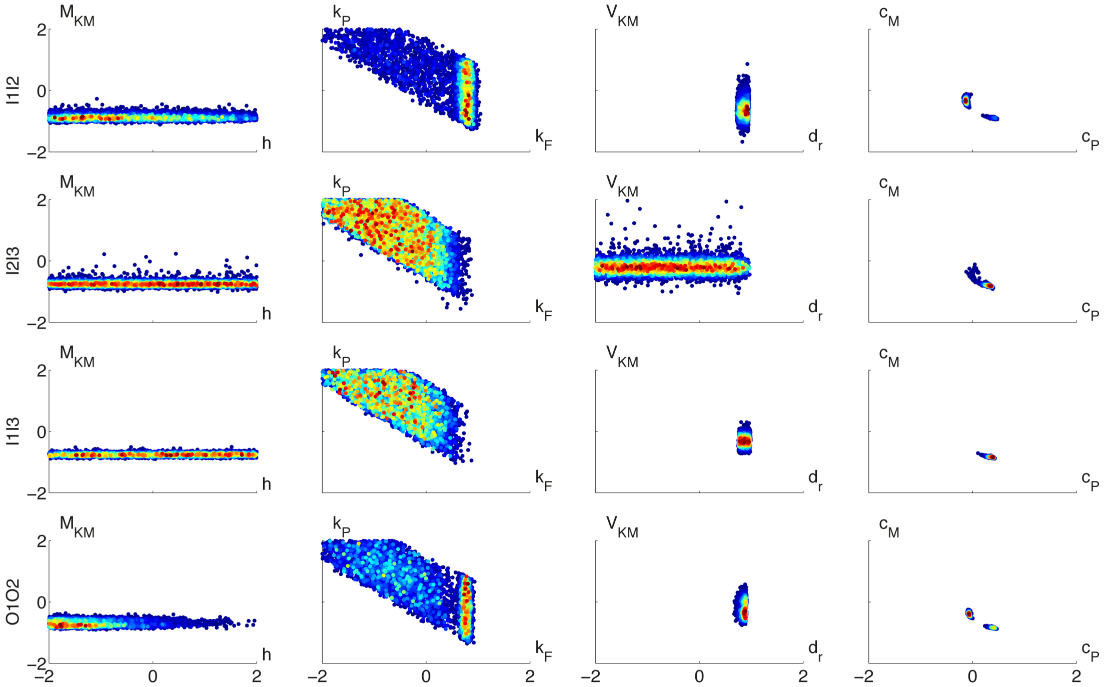

Figure S.11: Comparison of the posterior distributions computed from the data collected in different combinations of two experience-based experiments (first three rows) and from the two optimal experiments O1O2 (bottom row).

In this section we compare the results from the two optimal experiments to the results from different pairs of experience-based experiments. Figure S.11 show the posterior distributions obtained from the two optimal and all pairs of experience-based experiments. It is easy to notice that different pairs lead to significantly different posterior distributions. The corresponding predictive distributions are shown in Figure S.12.

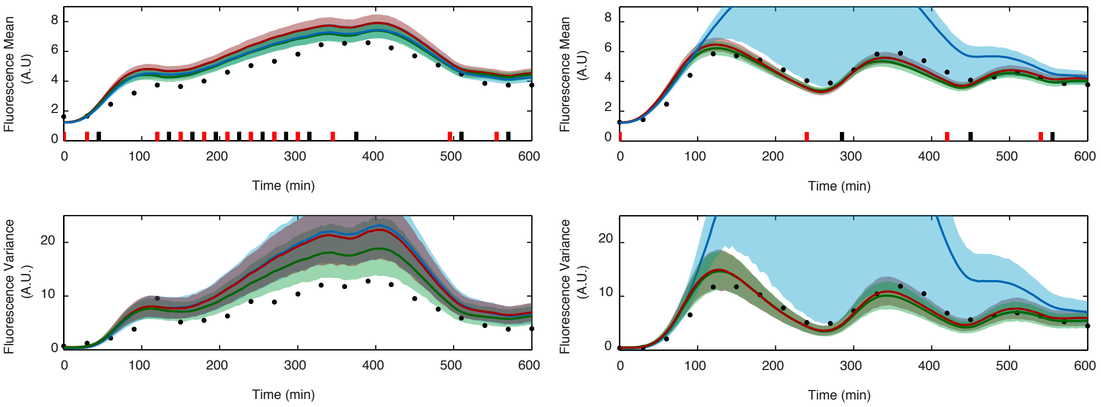

Figure S.12: Predictive distributions for the two validation experiments shown in the main text. The different colors correspond to the predictive distributions obtained with different combinations of two of the experience-based experiments in Figure S.6. Green: with the experiments from panel I1 and I2. Red: with the experiments from panel I1 and I3. Light blue: with the experiments from panel I2 and I3.

It can be seen in Figure S.12 that different combinations of two experience-based experiments lead to drastically different performances. The same conclusions can be reached from the differences of the mean likelihood of the data collected in all control and validation experiments, using the different posterior predictive distributions. Table S.8 shows the comparison of different combinations of experience-based experiments.

Table S.8: **Comparison of different pairs of two experience-based experiments.** For each performed experiment the log of the mean likelihood of the measured data according to the three different parameters posterior distributions is computed. The best model is the one with highest expected likelihood and is denoted by a star.

	I1I2	I2I3	I1I3
Figure 5A	$-125.1^\star$	-200.9	-356.5
Figure 5B	$-210.0^\star$	-235.5	-342.3
Figure 5C	$-135.9^\star$	-169.2	-315.1
Figure 5D	$-71.2^\star$	-123.3	-131.5
Figure 3	$-123.0^\star$	-186.3	-335.0
Figure 4	30.6	$68.0^\star$	20.2

By assuming uniform prior over the models, the likelihoods reported in Table 1 (in the main text) and Table S.8 can be used to statistically compare the models identified from the different pairs of experiments (optimal, random and experience-based) via Bayes factors. According to this metric, the model identified from the two optimal experiments outperforms all the other candidates obtained both with randoms and experience-based experiments (Bayes factor $\gg 1$ for each pair).

S.8 Control of population statistics

The model derived in the previous sections can be used not only to predict the behavior of the system in future experiments but also to design input sequences achieving a desired objective. For example, in biotechnological applications where genetically modified organisms are used to produce antibiotics or biofuels, one may require high protein production consistently in all cells of the population. On the other hand, to study effects related to heterogeneous cell populations, also the case of high variability between the expression levels in different cells could be of scientific interest. These properties can be enforced by regulating the mean and/or the variance of the protein across the population.

S.8.1 Reachability analysis

In order to determine whether the protein mean and variance can be regulated to follow a desired time-varying reference, the more basic question of what configurations of mean and variance can be achieved at a fixed time point using the available input must be addressed first. In the terminology of control theory such problems are known as reachability problems where the set of states that can be reached from the origin are determined for the system and input under consideration.

Definition 1 (Reachable set from the origin) *The reachable set $\mathcal{R}$ from the origin, for the control system*

$$\dot{x}(t) = f(x(t), u(t)),$$

with $x(t) \in \mathbb{R}^n, f(\cdot,\cdot) : \mathbb{R}^n \times \mathbb{R} \mapsto \mathbb{R}^n$ and bounded input $u(t) \in [0,1]$, is defined as the set of all states $\bar{x} \in \mathbb{R}^n$ that are reachable in finite time from $x(0) = 0$, using an admissible control law. That is

$$\mathcal{R} = \left\{ \bar{x} \in \mathbb{R}^n \mid \quad \exists\, \bar{t} \geq 0, \quad \exists\, u : [0,\bar{t}] \to [0,1] \quad \text{s.t} \quad \left\{ \begin{array}{rcl} x(0) & = & 0 \\ \dot{x}(t) & = & f(x(t),u(t))\ \forall t \in [0,\bar{t}\,] \\ x(\bar{t}) & = & \bar{x} \end{array} \right. \right\}.$$

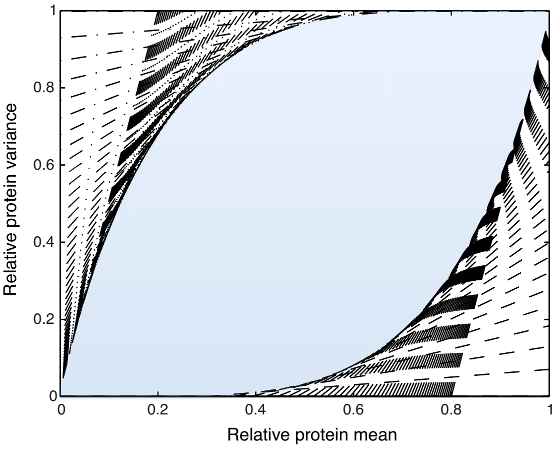

Figure S.13: The shaded region corresponds to the outer approximation of the reachable set for the mean and variance of the protein P, obtained using the approach suggested in [27] and the parameters identified from the two optimal experiments, $\hat{\gamma}^2$. For this computation, the population is assumed to be homogeneous, therefore V_{k_M} is set to zero. The protein mean and variance are scaled by the stationary values that would be obtained if $u(t) \equiv 1$ for all times, which are the maximum reachable values of protein mean and variance [27].

Generally, determining the set $\mathcal{R}$ can be very difficult for non-linear systems. For a simplified model of the light-inducible gene expression circuit, an outer approximation of $\mathcal{R}$ can be obtained using the approach proposed in [27]. Figure S.13 shows the approximation that is obtained with MAP estimates $\hat{\gamma}^2$. It can be seen that not all the configurations of protein mean and variance are achievable using the available input. In particular, it is not possible to drive the system to extreme configurations with high mean and low variance or viceversa. This suggests that also for the full model such configurations cannot be achieved.

S.8.2 Computing the light-induction patterns

Once a desired reference $z(t)$, compatible with the analysis of Section S.8.1, has been chosen it is possible to design a light sequence able to track it by solving the following optimization problem

$$\begin{aligned} L^\star_{[0,T]} = \operatorname*{argmin}_{L\in\mathcal{L}} & \quad \int_0^T J(\mu(\tau;\gamma), z(\tau))d\tau \\ \text{subject to:} & \quad \frac{d}{dt}\tilde{\mu}(t;\gamma) = A(\gamma, u(t;\gamma,L))\tilde{\mu}(t,\gamma) + B(\gamma, u(t;\gamma,L)),\ \forall t \in [0,T] \end{aligned} \tag{31}$$

where $\mu(\tau;\gamma) = [\mu_1(t;\gamma)\ \mu_1(t;\gamma)]$ is the vector of the mean and variance of the fluorescence intensity Y at time τ, $T > 0$ is the total duration of the experiment, $\mathcal{L}$ is the set of admissible light-induction pattern sequences and $J(\cdot,\cdot) : \mathbb{R}^2 \times \mathbb{R} \mapsto \mathbb{R}$ is a function that weights the deviation of the fluorescence mean $\mu_1(t;\gamma)$, variance $\mu_2(t;\gamma)$ or of the coefficient of variation, from the given scalar reference $z(\tau)$ at time τ. According to the objective that we wanted to achieve, we used the following cost functions

$$\begin{aligned} \text{mean tracking:} & \quad J(\mu,z) = \|\mu_1 - z\|^2 \\ \text{variance tracking:} & \quad J(\mu,z) = \|\mu_2 - z\|^2 \\ \text{coefficient of variation tracking:} & \quad J(\mu,z) = \|\tfrac{\sqrt{\mu_2-\mu_2^A}}{\mu_1-\mu_1^A} - z\|^2 \end{aligned} \tag{32}$$

To simplify the optimization problem (31), we approximate the integral of the cost function using the Euler method, with time step T_s, and we limit the choice of possible light sequences to the set $\mathcal{L}_{T_L}$ of light patterns obtainable if the pulses can be applied every T_L min. Therefore we have to solve the optimization problem

$$\begin{aligned} L^\star_{[0,T]} = \arg\min_{L\in\mathcal{L}_{T_L}} & \quad \sum_{k=1}^{\lfloor T/T_s \rfloor} J(\mu(kT_s;\gamma), z(kT_s)) \\ \text{subject to:} & \quad \frac{d}{dt}\tilde{\mu}(t;\gamma) = A(\gamma, u(t;\gamma,L))\tilde{\mu}(t;\gamma) + B(\gamma, u(t;\gamma,L)),\ \forall t \in [0,T] \end{aligned} \tag{33}$$

For all the experiments of Figure 5, we chose $T_s = 5$min and $T_L = 15$min. For an experiment of 10hrs this corresponds to $\#\left[\mathcal{L}_{T_L}\right] = 3^{41} = 3.6473\cdot 10^{19}$ different sequences. Since the solution of problem (33) using this set $\mathcal{L}_{T_L}$ is prohibitive, we decided to solve (33) in a receding horizon fashion. We set a time horizon $H := hT_s < T$ and solved (33) over this horizon by enumerating all the possible sequences and selecting the one with the lowest cost. Then we fixed the first designed pulse and repeated the same procedure for the following time interval, see Algorithm 1. We chose $H = 60$min for mean and variance tracking and $H = 90$min to track the coefficient of variation. Note that Algorithm 1 leads to a suboptimal solution of Problem (31), however we verified by simulation that the model output corresponding to the input pattern L^{tot} was able to track the given reference $z(t)$ with satisfactory precision.

```
Algorithm 1: Receding horizon light sequence design
Initialization. Choose an horizon $h > 0$, a time step size $T_s > 0$, an input grid step size $T_L > 0$ and a
                reference $z : [0, T + hT_s] \mapsto \mathbb{R}$;
                Let $K := \lfloor T/T_L \rfloor$ and initialize $L^{tot}(t) = 0 \ \forall t \in \{kT_L\}_{k=0}^{K}$;
for $k = 0 : K$ do
    Compute $L^\star_{[0,(k+h)T_L]}$ according to (33) with
        $\mathcal{L}_{T_L} := \left\{ L \ \middle| \ \begin{array}{ccll} L(iT_L) & = & L^{tot}(iT_L) & \text{for } i \in \{0, \ldots, k-1\} \\ L(iT_L) & \in & \{R, FR, 0\} & \text{for } i \in \{k, \ldots, k+h\} \end{array} \right\};$
    Set $L^{tot}(kT_L) = L^\star(kT_L)$;
end
```

S.9 Appendix

S.9.1 One dimensional marginals of the parameters posterior distribution

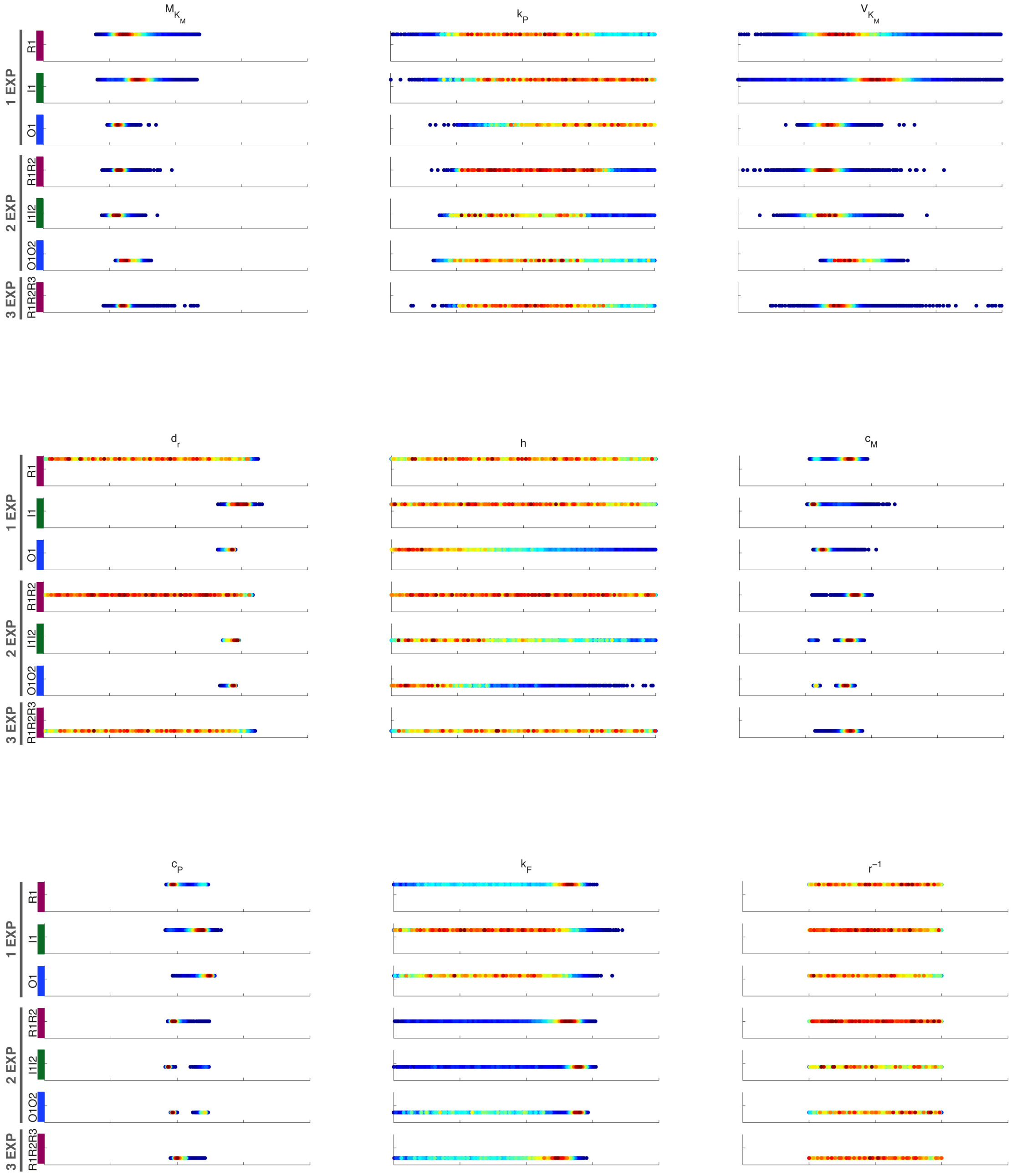

Figure S.14: One dimensional marginal posterior distribution of the parameters. The color legend on the left identifies combinations of random (magenta), experience-based (green) and optimal (blue) experiments.

S.9.2 Two dimensional marginals of the parameters posterior distribution

S.9.2.1 Optimal Design

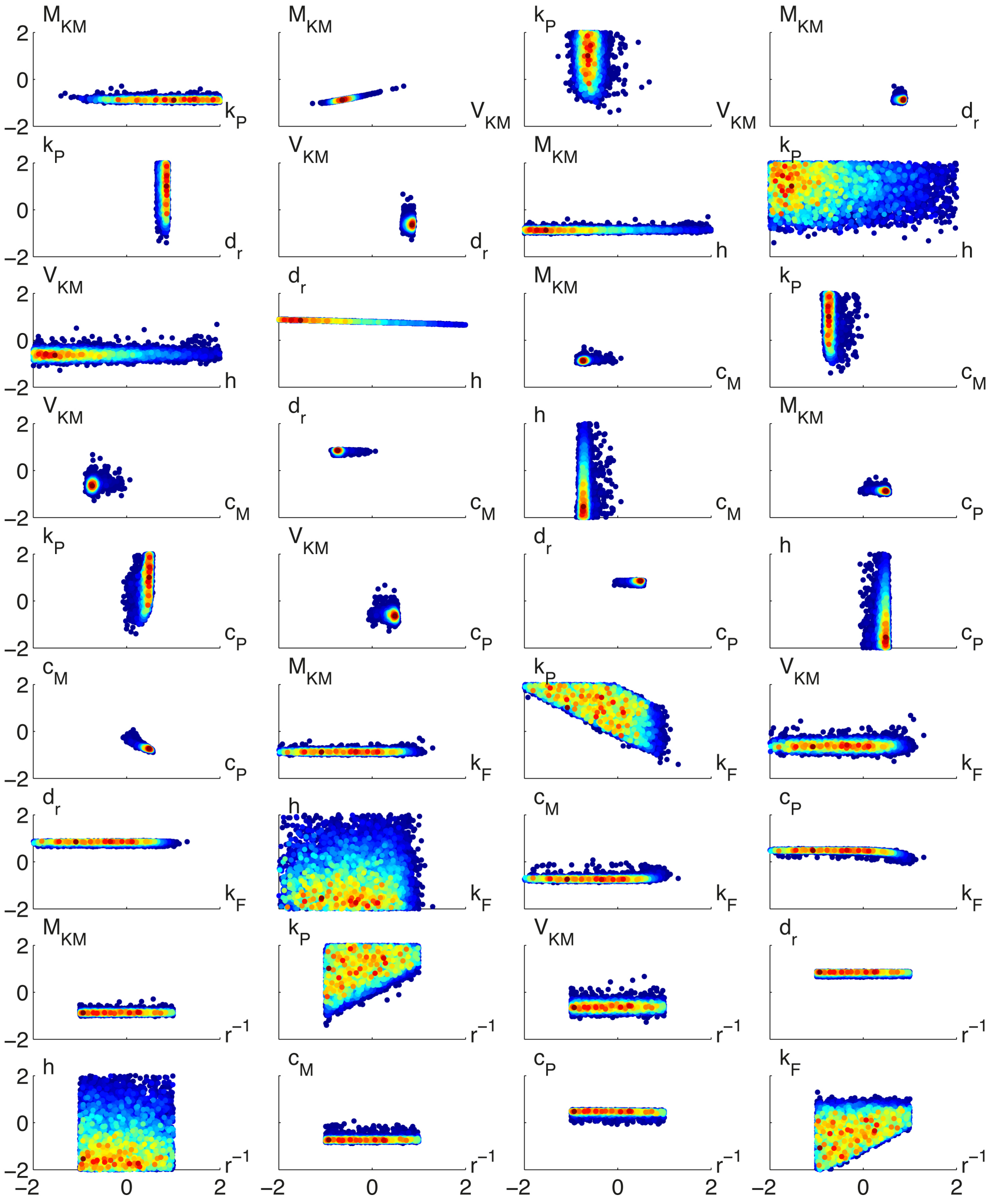

Figure S.15: Two dimensional marginals of the posterior distribution, obtained from O1, according to all the possible pairs of parameters.

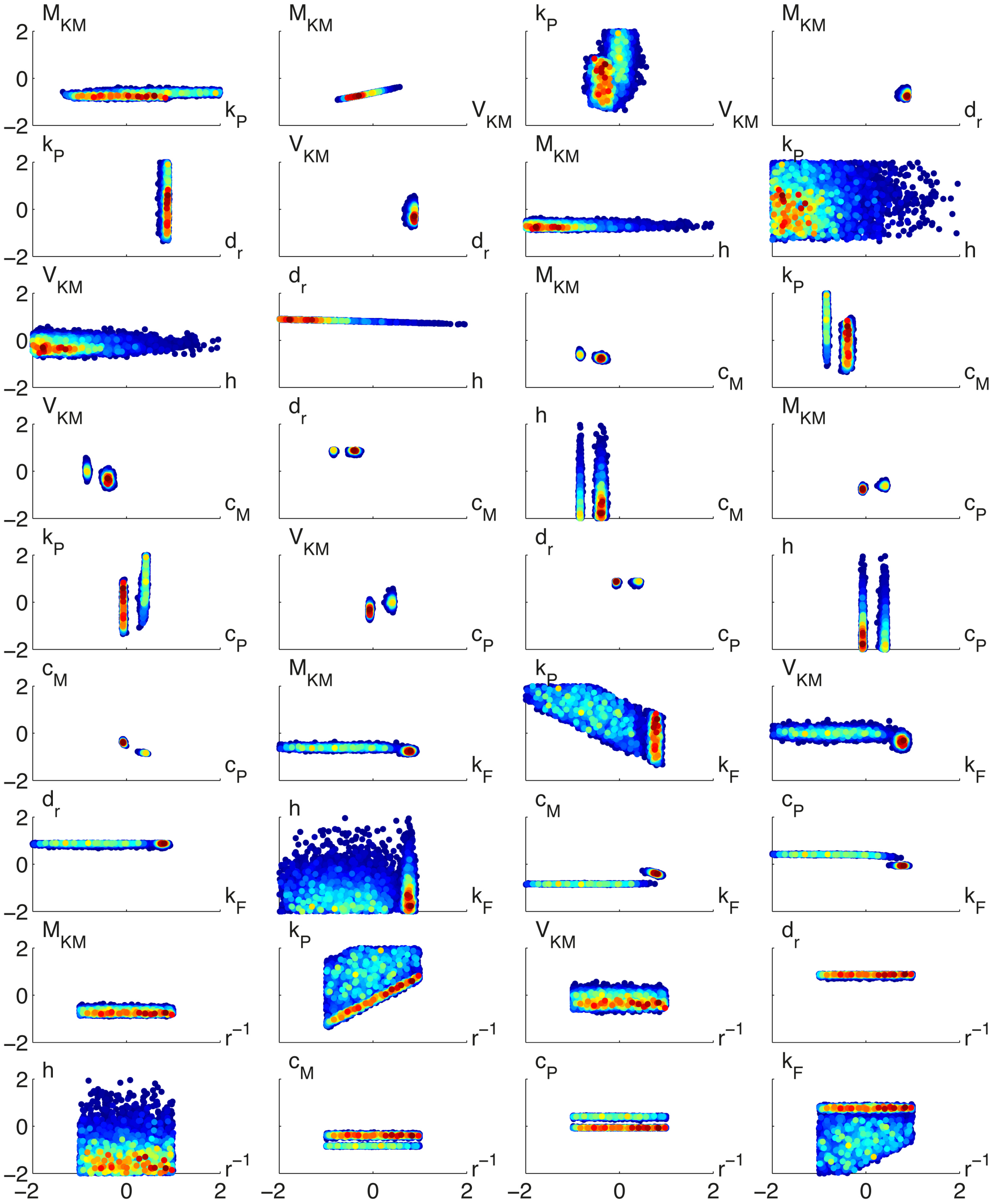

Figure S.16: Two dimensional marginals of the posterior distribution, obtained from O1O2, according to all the possible pairs of parameters.

S.9.2.2 Experience-based Experiments

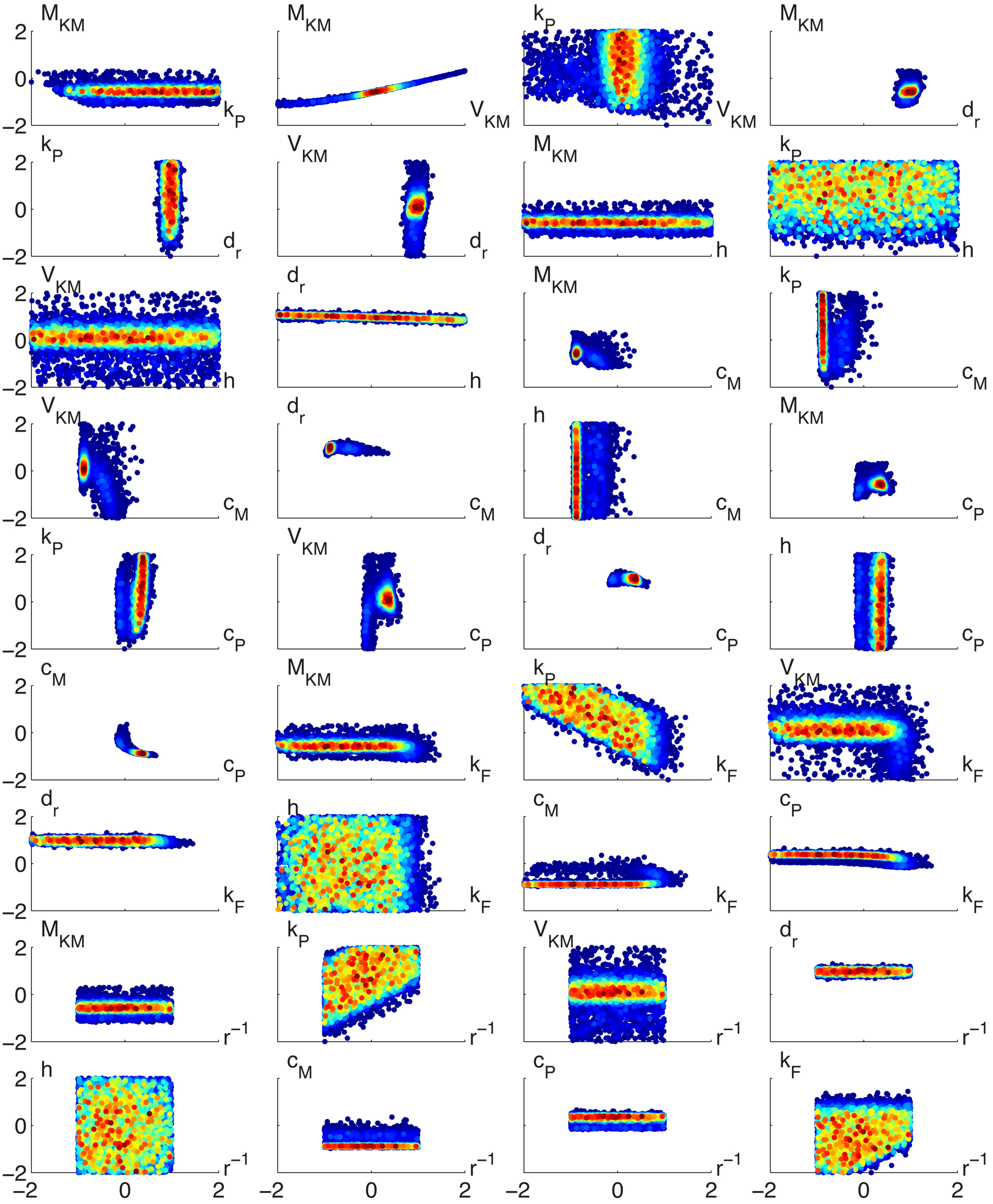

Figure S.17: Two dimensional marginals of the posterior distribution, obtained from the experience-based experiment I1, according to all the possible pairs of parameters.

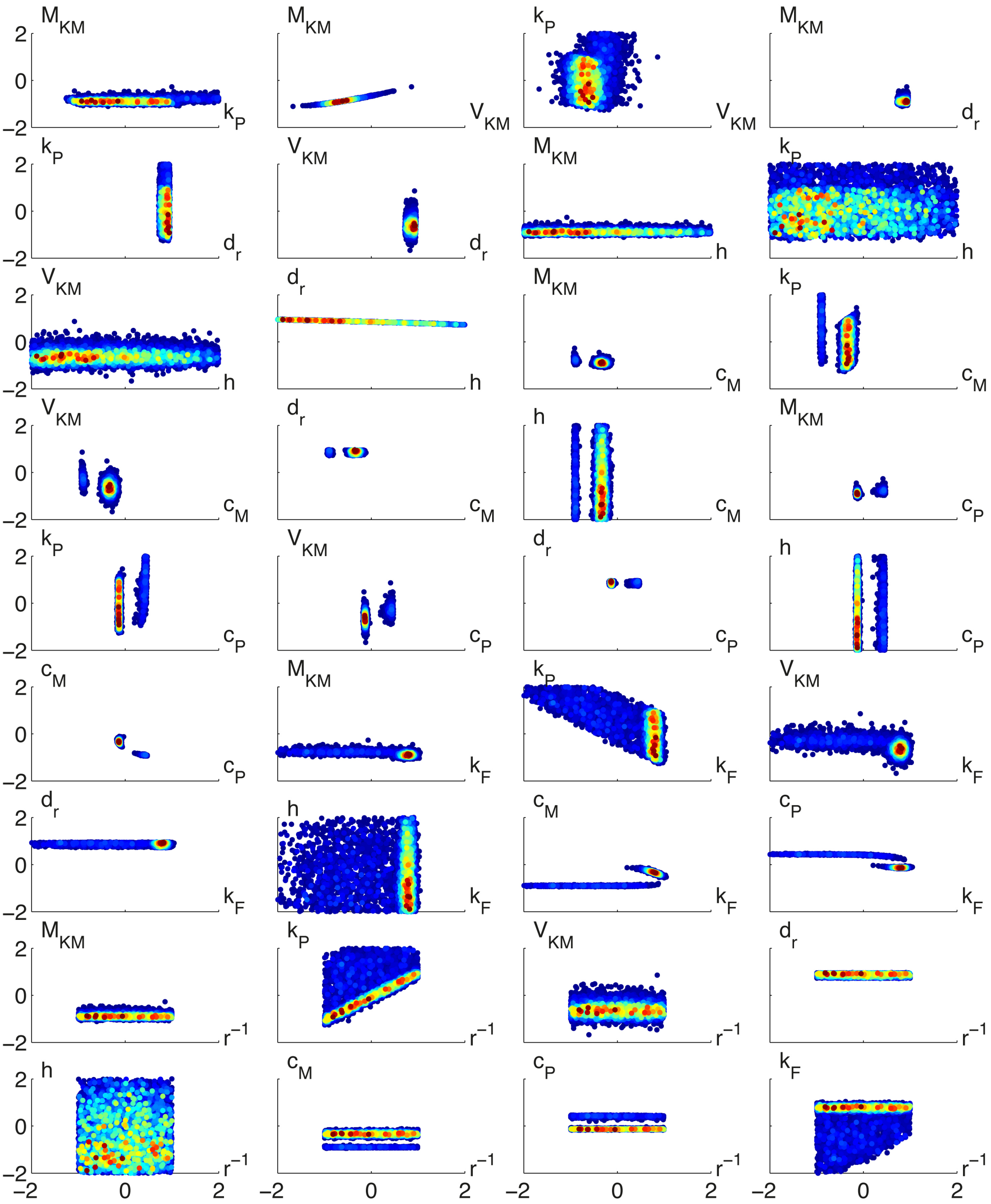

Figure S.18: Two dimensional marginals of the posterior distribution, obtained from the two best experience-based experiments I1I2, according to all the possible pairs of parameters.

S.9.2.3 Random Experiments

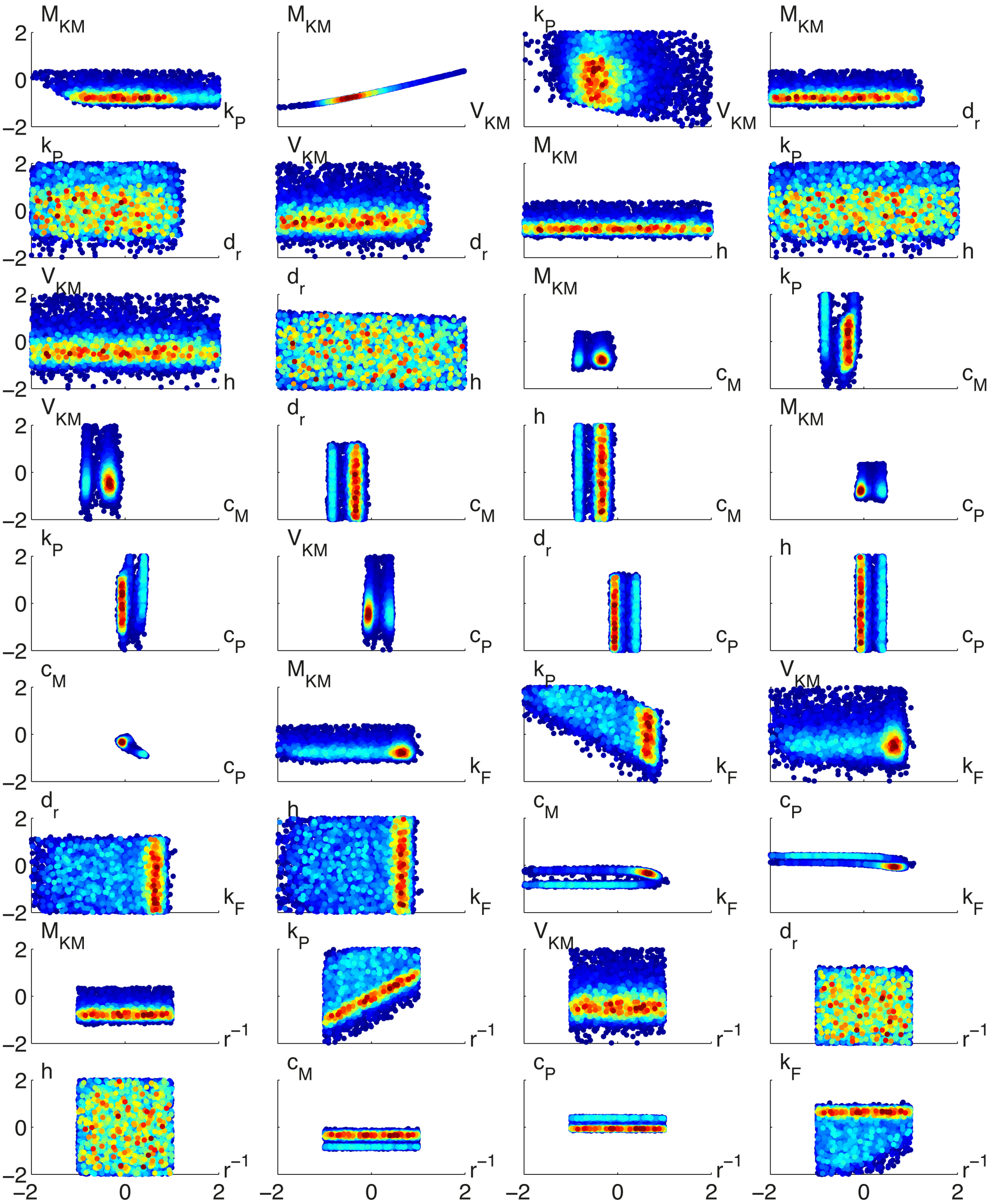

Figure S.19: Two dimensional marginals of the posterior distribution, obtained from one random experiment R1, according to all the possible pairs of parameters.

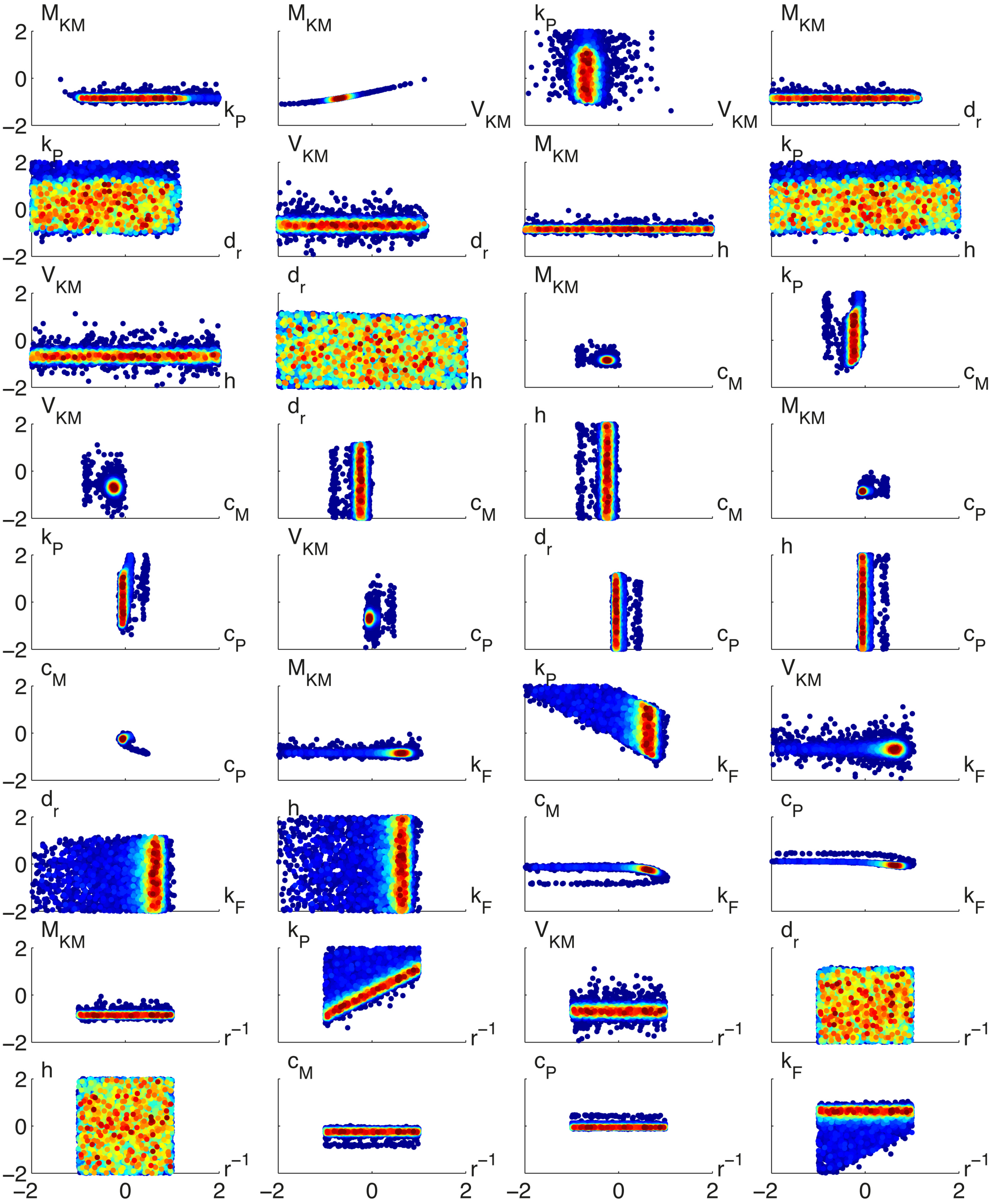

Figure S.20: Two dimensional marginals of the posterior distribution, obtained from two random experiments R1R2, according to all the possible pairs of parameters.

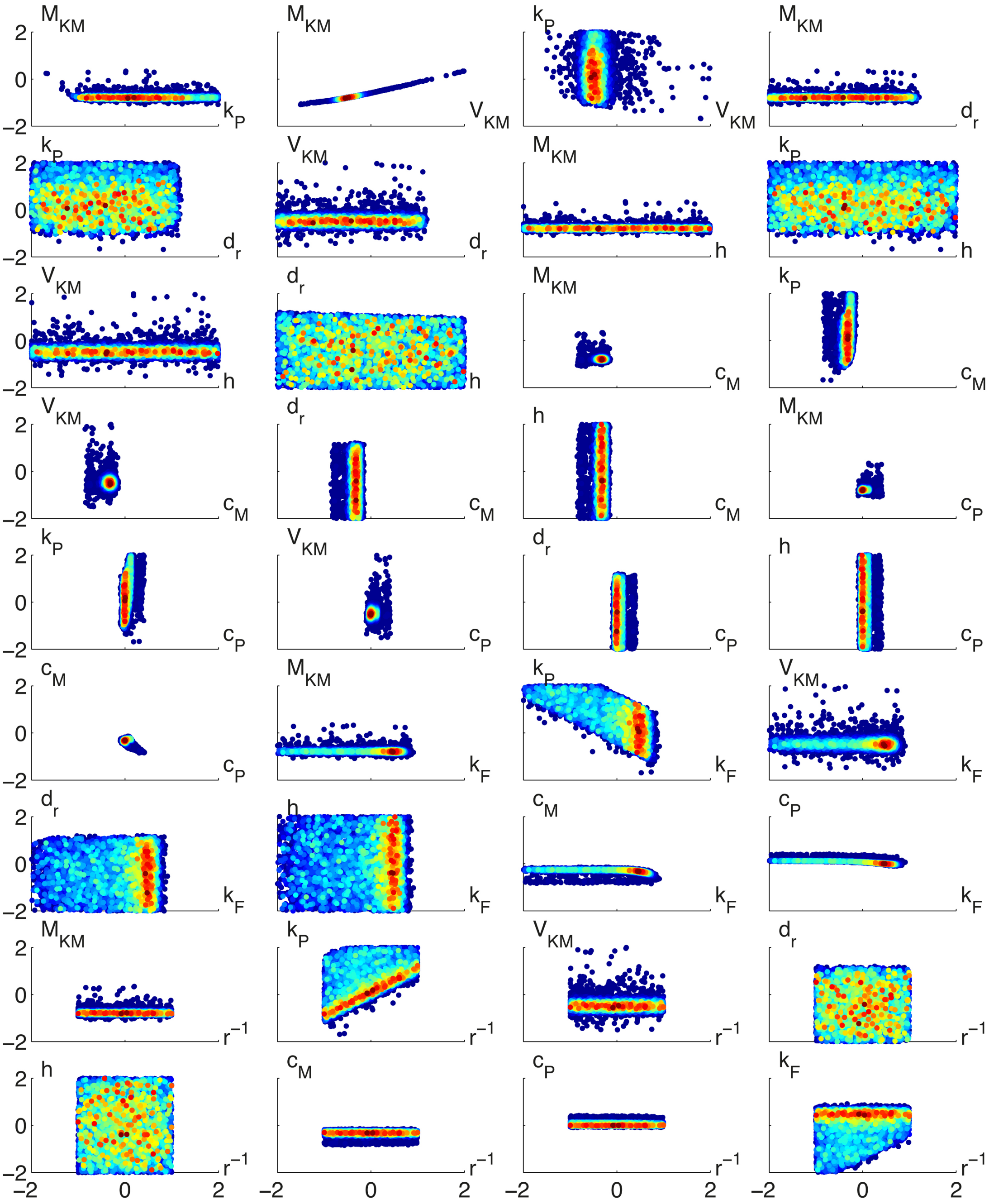

Figure S.21: Two dimensional marginals of the posterior distribution, obtained from three random experiments R1R2R3, according to all the possible pairs of parameters.

References

[1] Milias-Argeitis A, et al. (2011) In silico feedback for in vivo regulation of a gene expression circuit. *Nature Biotechnology* 29:1114–1116.

[2] Shimizu-Sato S, Huq E, Tepperman J, Quail P (2002) A light-switchable gene promoter system. *Nature Biotechnology* 20:1041–1044.

[3] Kendrick RE, Kronenberg G (1994) *Photomorphogenesis in plants* (Kluwer academic publishers).

[4] Levskaya A, Weiner OD, Lim WA, Voigt CA (2009) Spatiotemporal control of cell signalling using a light-switchable protein interaction. *Nature* 461:997–1001.

[5] Gillespie DT (1992) A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications* 188:404–425.

[6] Bretó C, Ionides E (2011) Compound markov counting processes and their applications to modeling infinitesimally over-dispersed systems. *Stochastic Processes and their Applications* 121:2571–2591.

[7] Ruess J, Milias-Argeitis A, Lygeros J (2013) Designing experiments to understand the variability in biochemical reaction networks. *Journal of the Royal Society Interface* 10:20130588.

[8] Zechner C, et al. (2012) Moment-based inference predicts bimodality in transient gene expression. *Proceedings of the National Academy of Sciences of the USA* 109:8340–8345.

[9] Singh A, Hespanha J (2011) Approximate moment dynamics for chemically reacting systems. *IEEE Transactions on Automatic Control* 56:414–418.

[10] Ruess J, Milias-Argeitis A, Summers S, Lygeros J (2011) Moment estimation for chemically reacting systems by extended kalman filtering. *The Journal of Chemical Physics* 135:165102.

[11] Komorowski M, Costa M, Rand D, Stumpf M (2011) Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. *Proceedings of the National Academy of Sciences of the USA* 108:8645–8650.

[12] Ruess J, Lygeros J (2013) Identifying stochastic biochemical networks from single-cell population experiments: a comparison of approaches based on the Fisher information. *IEEE 52nd Annual Conference on Decision and Control (CDC). Florence, Italy.* pp 2703–2708.

[13] Hindmarsh A, et al. (2005) SUNDIALS: Suite of Nonlinear and Differential/Algebraic Equation Solvers. *ACM Transactions on Mathematical Software (TOMS)* 31:363–396.

[14] Chaloner K, Verdinelli I (1995) Bayesian experimental design: a review. *Statistical Science* 10:273–304.

[15] Franceschini G, Macchietto S (2008) Model-based design of experiments for parameter precision: State of the art. *Chemical Engineering Science* 63:4846–4872.

[16] Hagen D, White J, Tidor B (2013) Convergence in parameters and predictions using computational experimental design. *Interface Focus* 3:20130008.

[17] Pronzato L, Walter E (1985) Robust experimental design via stochastic approximation. *Mathematical Biosciences* 75:103–120.

[18] Lawrence ND, Girolami M, Rattray M (2010) *Learning and inference in computational systems biology* (The MIT Press).

[19] Del Moral P, Doucet A, Jasra A (2006) Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68:411–436.

[20] Neal RM (2001) Annealed importance sampling. *Statistics and Computing* 11:125–139.

[21] Milias-Argeitis A, Porreca R, Summers S, Lygeros J (2010) Bayesian model selection for the yeast GATA-factor network: a comparison of computational approaches. *IEEE 49th Annual Conference on Decision and Control (CDC). Atlanta, GA, USA.* pp 3379–3384.

[22] Milias-Argeitis A (2013) Computational methods for simulation, identification and model selection in systems biology. *Doctoral dissertation, ETH Zurich.*

[23] Del Moral P, Doucet A, Jasra A (2012) An adaptive sequential monte carlo method for approximate bayesian computation. *Statistics and Computing* 22:1009–1020.

[24] Robert CP, Casella G (2004) *Monte Carlo statistical methods* (New York: Springer) Vol. 319.

[25] Munsky B, Khammash M (2006) The finite state projection algorithm for the solution of the chemical master equation. *The Journal of Chemical Physics* 124:044104.

[26] Gillespie D (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics* 22:403–434.

[27] Parise F, Valcher M, Lygeros J (2014) On the reachable set of the controlled gene expression system. *IEEE 53rd Annual Conference on Decision and Control (CDC). Los Angeles, USA.*