Bayesian design of synthetic biological systems

Chris Barnes, Daniel Silk, Xia Sheng and Michael P.H. Stumpf

Theoretical Systems Biology Group, Division of Molecular Biosciences,
Imperial College London, London SW7 2AZ, UK

http://www.theosysbio.bio.ic.ac.uk

Email: christopher.barnes@imperial.ac.uk, m.stumpf@imperial.ac.uk

Abstract

Here we introduce a new design framework for synthetic biology that exploits the advantages of Bayesian model selection. We will argue that the difference between inference and design is that in the former we try to reconstruct the system that has given rise to the data that we observe, while in the latter, we seek to construct the system that produces the data that we would like to observe, i.e. the desired behavior. Our approach allows us to exploit methods from Bayesian statistics, including efficient exploration of models spaces and high-dimensional parameter spaces, and the ability to rank models with respect to their ability to generate certain types of data. Bayesian model selection furthermore automatically strikes a balance between complexity and (predictive or explanatory) performance of mathematical models. In order to deal with the complexities of molecular systems we employ an approximate Bayesian computation scheme which only requires us to simulate from different competing models in order to arrive at rational criteria for choosing between them. We illustrate the advantages resulting from combining the design and modeling (or *in-silico* prototyping) stages currently seen as separate in synthetic biology by reference to deterministic and stochastic model systems exhibiting adaptive and switch-like behavior, as well as bacterial two-component signaling systems.

1 Introduction

As we are beginning to understand the mechanisms governing biological systems we are starting to identify potential ways of guiding or controlling the behavior of cellular and molecular systems. Rationally reengineering organisms for biomedical or biotechnological purposes has become the central aim of the fledgling discipline of synthetic biology. By redirecting regulatory and physical interactions or by altering molecular binding affinities we may, for example, control metabolic processes [1, 2] or alter intra and inter cellular communication and decision making processes [3, 4]. The range of potential applications of such engineered systems is vast: designing microbes for biofuel production [5, 6] and bioremediation [7]; developing control strategies which drive stem cells through the various decisions to become terminally differentiated (or back) [8, 9], with the aim of developing novel therapeutics [10, 11]; construction of new drug-delivery systems with homing microbes delivering molecular medicines directly to the site where they are needed [12]; use of bacteria or bacterial populations (employing swarming and quorum sensing) as biosensors [13]; and gaining better understanding of all manner of biological systems by systematically probing their underlying molecular machinery.

A range of tools and building blocks for such engineered biological systems are now available which allow us to, at least in principle, build such systems from simple and reusable biological components [14]. In electronic systems, such modularity has been crucial and has allowed the cost-effective production of reliable components that can be combined to produce desired outputs. Biology, however, poses

different and novel challenges that are intimately linked to the biophysical and biochemical properties of biomolecules and the media in which they are suspended. Especially in crowded environments such as found inside living cells the lack of insulation between different components, i.e. the very real possibility of undesired cross talk, can create problems; with increasing miniaturization similar, albeit quantum effects, are now also surfacing in electronic circuits [15].

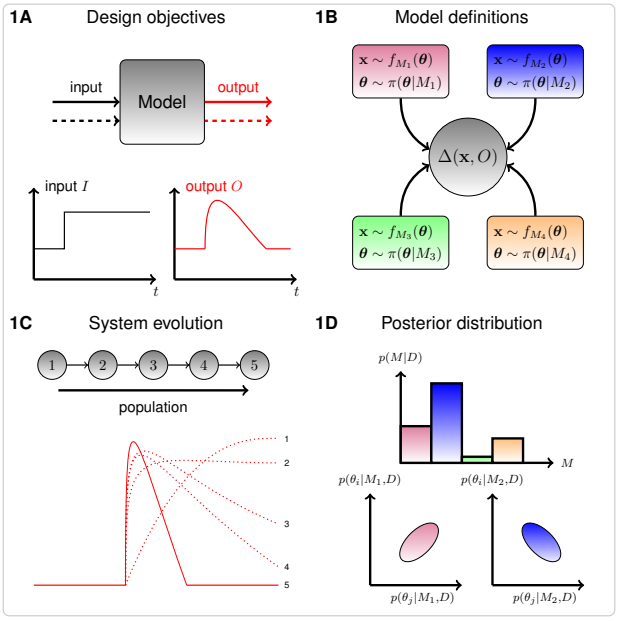

Figure 1: Bayesian approach to system design. A) The design objectives are encoded by the specification of input and output characteristics. B) One or more competing designs for the system are specified together with priors on the parameters. A distance function, $\Delta(\mathbf{x}, O)$, relates model output to the desired output characteristic. C) The system is evolved using sequential Monte Carlo. Each population more accurately approximates the desired behavior. D) The model posterior probability encodes the ability of each design to achieve the desired behavior. The parameter posterior shows parameters that are sensitive or insensitive to the input-output specification.

As synthetic biology gears up to bring engineering methods and tools to bear on biological problems the way in which we manipulate biological systems and processes is likely to change. Historically, each new branch of engineering has gone through a phase of what can be described as tinkering before rationally planned and executed designs became common place. Arguably, this is the current state of synthetic biology and it has indeed been suggested that the complexity of synthetic biological systems over the past decade has reached a plateau [16]. From the earliest days, explicit quantitative modeling of systems has been integral to the vision and practice of synthetic biology and it will become increasingly important in the future. The ability to model how a natural or synthetic system will perform under controlled conditions must in fact be seen as one of the hallmarks of success of the integrative approaches taken by systems and synthetic biology.

Here we present a statistical approach to the design of synthetic biological systems that utilizes methods from Bayesian statistics to train models according to specified input-output characteristics. It incorporates modeling and automated design and is general in the sense that it can be applied to any system that can be described by a mathematical model which can be simulated. Because of the statistical nature of this approach, previously challenging problems such as handling stochastic

models, accounting for kinetic parameter uncertainty and incorporating environmental stochasticity can all be handled in a straightforward and consistent manner.

2 Bayesian approach to system design

The question of how to design a system to perform a specified task can be viewed as an analogue to reverse engineering. In design we want to elucidate the most appropriate system to achieve our design objectives; in reverse engineering we aim to infer the most probable system structure and dynamics that can give rise to some observational data. In this respect, the design question can be viewed as statistical inference on data *we wish to observe*.

In the Bayesian approach to statistical inference the posterior distribution is the quantity of interest and this is given by the normalized product of the *likelihood* and the *prior*. In most practical applications the posterior distribution cannot be derived analytically but if the likelihood (and prior) can be expressed mathematically we can use Monte Carlo methods to sample from the posterior. In many cases where the model structure is complex the likelihood cannot be written in closed form and traditional Monte Carlo techniques cannot be applied. These include inference for the types of stochastic processes encountered in systems and synthetic biology. In these cases a family of techniques known collectively as approximate Bayesian computation (ABC) can be applied: these use model simulations to approximate the posterior distribution directly. Here we use a sequential Monte Carlo ABC algorithm known as ABC SMC to move from the prior to the approximate posterior via a series of intermediate distributions [17]. This framework can also be used to perform Bayesian model selection [18] and has been implemented in the software package ABC-SysBio [19].

Figure 1 depicts the approach presented here. The design objectives are first specified through input-output characteristics. Here these have been depicted as a single time series, though the method can be applied in a much broader sense with multiple inputs and outputs. A set of competing designs is then specified through deterministic or stochastic models each containing a set of kinetic parameters and associated prior distributions. The distance function measures the discrepancy between the model output and the objective. In principal it is possible to specify a distribution over the objective and each model could also contain experimental error. The ABC SMC algorithm then automatically evolves the set of models towards the desired design objectives. The results are a set of posterior probabilities representing the probability for each design to achieve the specified design objectives in addition to the posterior probability distribution of the associated kinetic parameters. This approach is similar in spirit to some existing methods for the automated design of genetic networks such as those adopting evolutionary algorithms [20, 21], Monte Carlo methods [22, 23] or optimization [24, 25, 26] but the advantages of our method over traditional ones are that we can utilize powerful concepts from Bayesian statistics in the design of complex biological systems, including

- the rational comparison of models under parameter uncertainty using Bayesian model selection which automatically accounts for model complexity (number of parameters) and robustness to parameter uncertainty
- a posterior *distribution* over possible design parameter values that can be analyzed for parameter sensitivity and robustness and provide credible limits on design parameters
- the treatment of stochastic systems at the *design* stage including the design of systems with required *probability distributions* on system components.
- methods for the efficient exploration of high dimensional parameter space

In the following we demonstrate the power of this approach by examining, from this new perspective, systems that have been of interest in the recent literature. First we consider systems that are

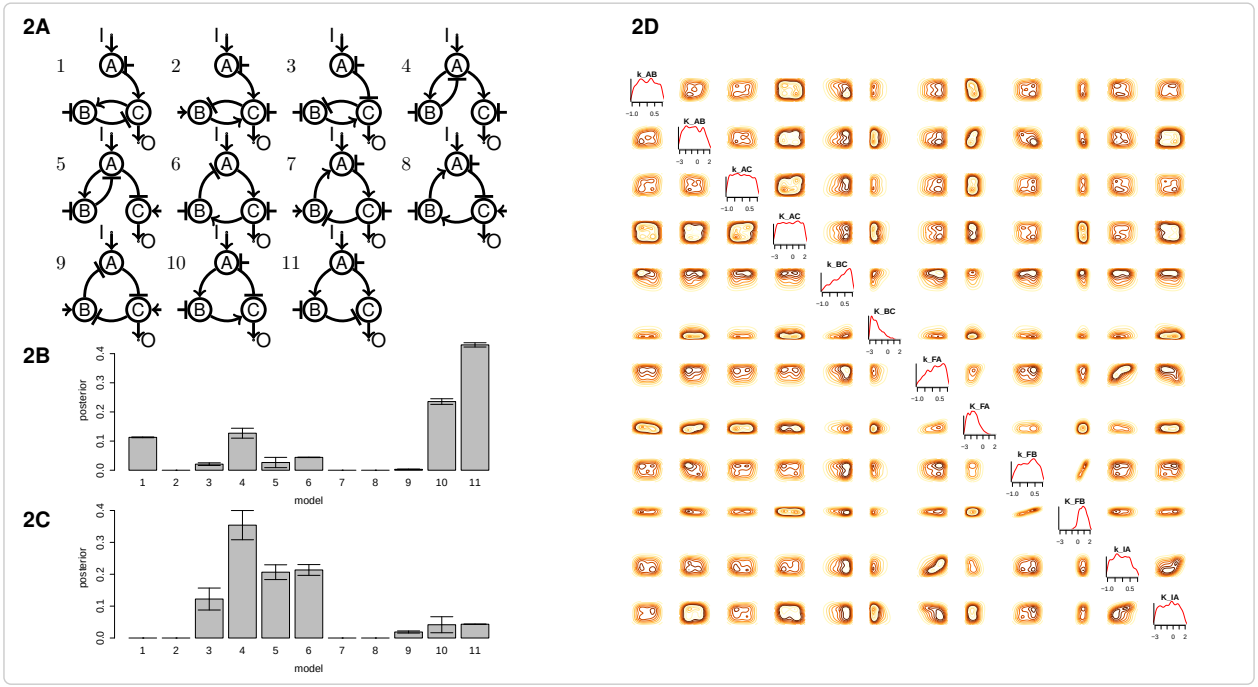

Figure 2: Biochemical adaptation. A) 11 networks capable of biochemical adaptation. A, B, C represent enzymes that catalyze reactions in their active state. For example $A \rightarrow B$ indicates that A converts B from its inactive to active state and $A \dashv B$ indicates that A converts B from its active to inactive state. The input is applied to species A and the output is taken to be the concentration of the active form of C. The concentrations of active and inactive forms sum to 1. Reactions with no origin node refer to background activating/deactivations enzymes. B) Posterior probability for achieving biochemical adaptation when there is no cooperativity. C) Posterior probability for achieving biochemical adaptation when cooperativity is included. The error bars in panels B and C indicate the variability in the marginal model posteriors over three separate runs. D) Parameter posterior distribution, represented by univariate and bivariate marginal distributions, for model 11 in the case of no cooperativity.

capable of biochemical adaptation [27]. We then look at the ability of two bacterial two component system (TCS) topologies to achieve particular input-output behaviors; and finally we finish with an analysis of designs for a stochastic toggle switch with no cooperative binding at the promoter.

3 Biochemical adaptation

Biochemical adaptation refers to the ability of a system to respond to an input signal and return to the pre-stimulus steady state (Figure 1A). Ma *et al* [27] identified two three-node network topologies that are necessary for biochemical adaptation: a negative feedback loop with a buffering node (NFBLB) and an incoherent feedforward loop with a proportioner node (IFFLP). Within these categories they identified eleven simple networks that were capable of adaptation (Figure 2A). We applied the Bayesian design approach to these eleven networks using Michaelis-Menten kinetic models with and without cooperativity (see appendix B for the ordinary differential equations (ODEs) describing these models). The desired output characteristics were defined through the adaptation efficiency, E, and sensitivity,

S, given by

$$E = \left| \frac{(O_2 - O_1)/O_1}{(I_2 - I_1)/I_1} \right| \; S = \left| \frac{(O_{peak} - O_1)/O_1}{(I_2 - I_1)/I_1} \right|,$$

where I_1, I_2 are the input values (here fixed at 0.5 and 0.6 respectively), O_1, O_2 are the output steady state levels before and after the input change and O_{peak} is the maximal transient output level. We defined the two component distance to be $\epsilon = \{E, S^{-1}\}$ such that as ϵ decreases the behavior approaches the desired behavior. The final population was defined to obey the toletances $\epsilon = \{0.1, 1.0\}$, which defines close to perfect adaptation (when $O_1 - O_2 \leq O_1/50$) and a fractional response equal to the fractional change in input.

The results of the model selection are shown in Figure 2 (B and C). When cooperativity is not included the most robust designs for producing the desired input-output characteristics are the incoherent feedforward loops, but when cooperativity is added the posterior shifts significantly towards the negative feedback topology. If a system with these requirements were to be implemented then not only would designs 11 and 4 be clear candidates for further study, but many of the designs can be effectively ruled out and the ranking of the models provides a clear strategy for an experimental programme. These results also illustrate how small changes in context or incomplete understanding of a system can produce a large change in the most robust design. The Bayesian framework allows us to incorporate such uncertainty — or safeguard against our ignorance — naturally into the design process.

The posterior distribution provides information on which parameters are correlated and which are the most sensitive to the desired behavior. The posterior for model 11 under no cooperativity is shown in Figure 2D, where the ODE model is given by

$$\begin{aligned}
\frac{dA}{dt} &= Ik_{IA}\frac{(1-A)}{(1-A)+K_{IA}} - F_A k_{FA}\frac{A}{A+K_{FA}} \\
\frac{dB}{dt} &= Ak_{AB}\frac{(1-B)}{(1-B)+K_{AB}} - F_B k_{FB}\frac{B}{B+K_{FB}} \\
\frac{dC}{dt} &= Ak_{AC}\frac{(1-C)}{(1-C)+K_{AC}} - Bk_{BC}\frac{C}{C+K_{BC}},
\end{aligned}$$

where $X = \{A, B, C\}$ corresponds to the concentrations of the active forms of proteins ($1 - X$ corresponds to the concentrations of the inactive form). I represents the input signal and the k and K represent the reaction rate parameters (of which there are 12 in total). F_A and F_B represent background deactivating enzymes with concentration fixed to 0.5.

The posterior shows in particular that the parameters for the background deactivating enzyme (k_{FB} and K_{FB}) on node B are correlated, and that K_{FB} should be large; this is exactly the requirement for the linear regime necessary for the IFFLP system to achieve adaptation [27]. A principal component analysis of the posterior (Figure S1) shows other correlated parameters on the first few principal components. The last principal component describes the direction of least variance and therefore the most sensitive parameters. From this we can deduce for example that the reaction between nodes A and C is relatively unimportant. A similar analysis for model 4 in the case of cooperativity (Figures S2 and S3) shows for example that the behavior is insensitive to the values of the Hill coefficients and the details of the reaction between nodes A and C.

4 Robust oscillator design

Biochemical oscillations are increasingly being implemented in various synthetic systems [28, 29, 30, 31]. A recent study by Tsai *et al* [32] compared the ability of five small networks to achieve oscillations.

The five designs are shown in Figure 3A where each node represents the active form of a protein, edges represent enzymatic reactions and thicker edges represent increased feedback strength. We applied our Bayesian design methodology to the original problem and further investigated the ability of these designs, provided in detail in appendix C, to achieve particular amplitude-frequency values.

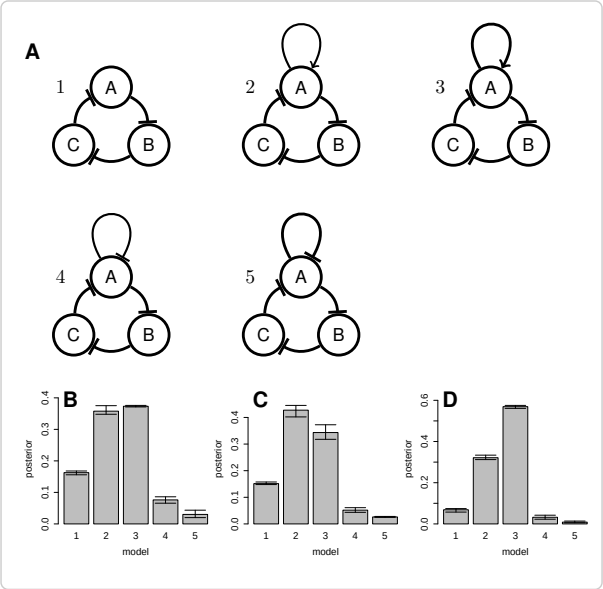

Figure 3: Robust oscillator models. A) 5 oscillator models. Model 1 is a loop of repressive enzymatic reactions. Models 2 and 3 have an additional positive feedback loop on node A with the feedback strength stronger in model 3 (represented by the thicker loop). Models 4 and 5 have an additional negative feedback loop on node A with the feedback strength stronger in model 5. B) Posterior probability for achieving Hopf bifurcation type limit cycle oscillations. C) Posterior probabilities for species A achieving oscillations with amplitude of 0.1 and a frequency of 1Hz. D) Posterior probabilities for species C achieving oscillations with amplitude of 0.1 and a frequency of 1Hz. The error bars in panels B, C and D indicate the variability in the marginal model posteriors over three separate runs.

Figure 3B shows the posterior probability for each model to achieve limit cycle behavior induced by a Hopf bifurcation. The addition of the negative feedback loop in models 4 and 5 does not improve the ability to achieve oscillations. We find that the addition of a positive feedback loop on species A in models 2 and 3 increases the ability of the system to achieve limit cycle behavior, but no significant increase in the posterior probability is provided by increasing the feedback strength. This is in conflict with the original study that found that model 3 outperformed model 2 [32]. Our approach does sample parameter space predominantly in regions where the desired behavior is more likely, rather than entirely at random as was done in the previous study; on balance this suggests that the posterior probability for delivering robust oscillations is approximately the same for models 2 and 3.

More insight can be gained into this discrepancy by specifying a particular frequency and amplitude of the oscillator as the desired output behavior. Figures 3C and D show the model posterior probability after requiring an amplitude of 0.1 and a frequency of 1.0 Hz on species A and C respectively. The first thing to note is that the model posterior is significantly different in the two cases. When the constraints are applied to species A, model 3 is favored with the increase in feedback strength *decreasing* the ability to reach the specified behavior. When the conditions are applied to species C (and species B

by symmetry) we get a posterior that more resembles the original findings; that the increase in feedback strength does indeed increase the ability to reach the specified oscillations. Thus the posterior for the Hopf bifurcation behavior represents a sum over all possible oscillator characteristics; in a manner that is reminiscent of Bayesian model averaging.

If we examine the posterior distribution and the principal component analysis for model 2 to achieve Hopf bifurcations (Figures S4 and S5), we can see that the parameters k_1 and k_3, which are the strengths of the deactivating reactions on nodes A and B, are constrained to be similar in magnitude to k_5. We also see that within this model the feedback strength, k_7, does not affect the dynamics significantly. Here, and elsewhere, we can use the posterior distributions in order to gain insights into the sensitivity and robustness of the system to variations in parameters, irrespective of whether the system's dynamics are deterministic or stochastic: our ABC SMC framework allows us to extract such information on the fly as part of the sequential design process.

5 Bacterial two component systems

Two component systems (TCS) allow bacteria to sense external environmental stimuli and relay information into the cell, e.g. to the gene expression apparatus. They consist of a histidine kinase (HK) that autophosphorylates upon interaction with a specific stimulus. The phosphate group is then passed on to a cognate response regulator protein (RR) which, once phosphorylated, can regulate transcription [33].

Naturally occurring TCS differ in the number of phosphate binding domains. In the most common form there are two phosphate binding domains (Figure 4A) but an alternative form exists that consists of a phosphorelay mechanism with four phosphate binding domains (Figure 4B). These shall be referred to as the orthodox and unorthodox TCSs, respectively [34]. The reason for the existence of two forms of TCS remains largely unknown but it has been demonstrated that the phosphorelay is robust to noise and can provide an ultrasensitive response to stimuli [34, 35], whereas the orthodox system can provide behavior that is independent of the concentrations of its components [36]. Here we have applied the Bayesian approach to directly compare the ability of orthodox and unorthodox designs to achieve various input-output behaviors, using ODE models similar to ones described previously [34] (see appendix D).

Figures 4C-F show four types of behaviors that may be desired in synthetic TCS systems (e.g. for bioremediation or biopharmaceutical applications), and the corresponding posterior probabilities of the orthodox and unorthodox models to achieve them. In Figure 4C the specified behavior is that of a fast response to a square pulse input signal. That is the output should show a maximum within 0.1 seconds after the pulse starts and a minimum 0.1 seconds after the pulse ends. As can be seen from the posterior probabilities, both models achieve this behavior easily, as one would expect from a signaling system, with the orthodox system slightly outperforming the unorthodox system. In Figure 4D the ability of the two systems to achieve a steady output state for $t > 2$ seconds under a constant input signal is examined, and again both systems perform comparably with the orthodox system appearing slightly more favorable.

In Figure 4E the input signal is a high frequency sinusoid with a mean of 0.6, and the desired output is a constant signal with the same mean and root mean square < 0.3; this would mimic a system that is robust to high-frequency noise. The output trajectories at some intermediate and final populations are shown in Figure S6. In this example the unorthodox system clearly outperforms the orthodox system, which indicates the increased robust to noisy signals that comes with the relay architecture [34]. But the direct comparison of the two models' ability to cope with noise, which is becoming possible in this approach, also reveals some unexpected characteristics: inspection of the posterior distribution (Figure S7) shows that all the dephosphorylation reaction rates, k_6, k_8, k_9, are

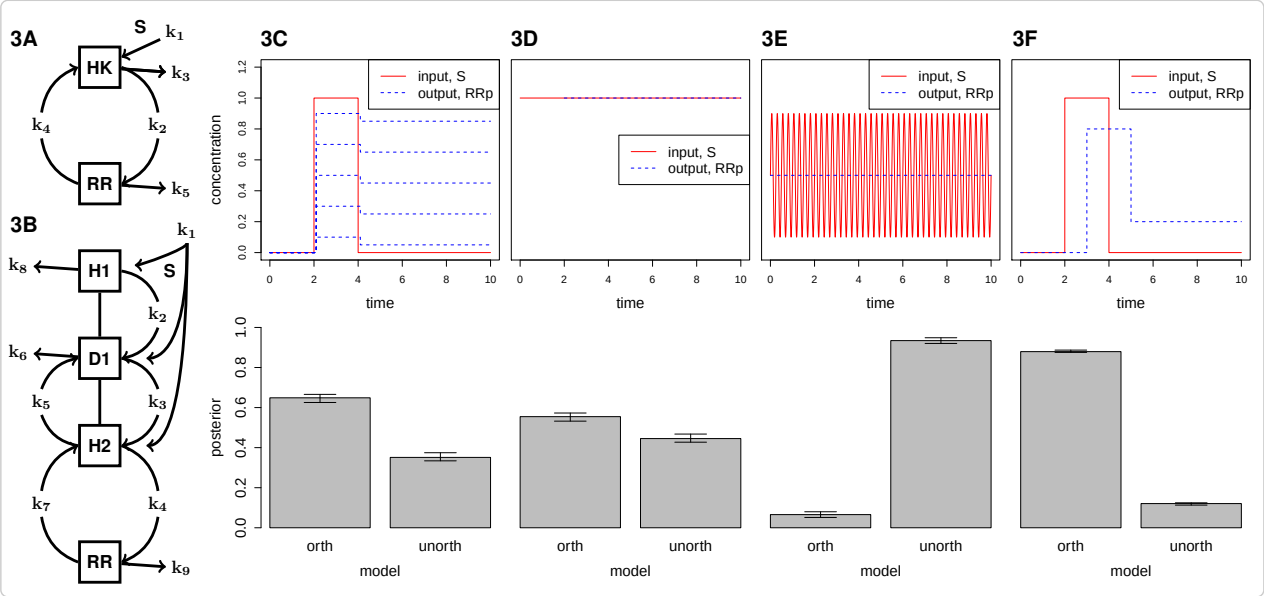

Figure 4: Bacterial two component systems. A) Orthodox system where HK denotes the histidine kinase and RR the response regulator, both of which have phosphorylated forms, HKp and RRp. Arrows represent reactions involving phosphate groups and the k_i represent the rate parameters. S is the input stimulus signal that causes autophosphorylation of the histidine kinase. B) Phosphorelay system with three phosphate-binding domains, where H, D refer to histidine and aspartate domains respectively. C-F) Specified input-output behavior (above) and posterior probabilities for the two designs to achieve it (below). The input signal corresponds to the stimulus, S, and the output signal is represented by the concentration of phosphorylated response regulator, RRp. The error bars indicate the variability in the marginal model posteriors over three separate runs.

minimized while the rate of the signal induced autophosphorylation (k_1) is large. Thus the noise reduction mechanism in the unorthodox system works by saturating the system.

In Figure 4F the input is again a step function but the output is more specific; it much reach to > 0.8 and drop to < 0.2 within 0.5 seconds of the pulse start and end, respectively, thus approximately reproducing the input. Figure S8 shows the evolution of the system in this case. Here the orthodox model clearly outperforms the unorthodox model. Inspection of the posterior distribution (Figure S9) shows that both the rate of the signal induced autophosphorylation and the rate of phosphorylation of the response regulator by the histidine kinase, k_1, k_2, are large while the rate of dephosphorylation of the response regulator, k_5, is small. This ensures that the shape of the signal is transferred faithfully through the system.

6 Stochastic genetic toggle switch without cooperativity

The genetic toggle switch is a synthetic realization of a bistable switch that forms the basis of cellular memory [37]. It is formed by two genes A and B, whose respective proteins repress the production of the other protein; protein A represses the production of protein B and *vice versa* (Figure 5). The presence of an interaction with inducer molecules allows the system to switch between steady states

with the probability of spontaneous switching low enough such that, in the absence of an interaction, the system will effectively remain in its appropriate steady state indefinitely.

Here we consider four versions of the stochastic genetic toggle switch that are all bistable without the requirement for cooperative binding of the proteins to the gene promoter [38]. Note that the deterministic models are not necessarily bistable; these are shown in Figure 5A and consist of the basic toggle switch, an exclusive version containing only one promoter, a version that includes bound repressor degradation (BRD), and a version containing a protein-protein interaction between A and B with the resulting complex nonfunctional. The additional reactions are are always to reduce the probability of the 'deadlock' state where both A and B are bound to the promotors of B and A respectively [38]. We modeled the switches using a continuous time Markov jump process which obeys the chemical master equation. Only protein level reactions were modeled which makes the models simpler (and faster to simulate) while retaining the important behavior. The stochastic models for all four switches are given in appendix E. For such complicated stochastic dynamical systems the advantages of a Bayesian perspective over conventional model design strategies (based e.g. on optimization) come to the fore: without an appreciation of the whole distribution choice of the best model would be subject to considerable uncertainty.

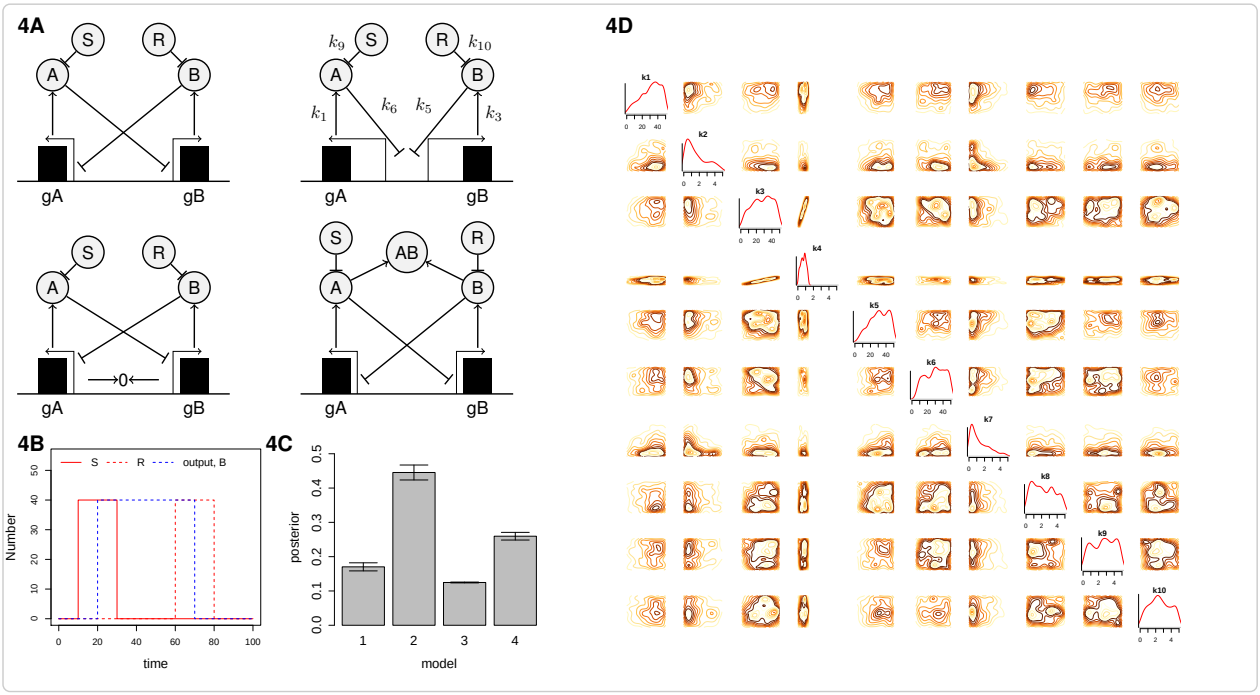

Figure 5: Stochastic toggle switch. A) Four different designs for a toggle switch without cooperative binding. Going clockwise from top left we have the basic switch, the exclusive switch where there is only one repressor site, the basic switch with bound repressor degradation (BRD), and the basic switch with a protein-protein interaction. Genes gA, gB express proteins A, B and S, R represent inducer molecules. The species AB represents complex formation. The rate constants for the reactions are shown for the exclusive switch (protein degradation and transcription factor dissociation from the promoter are not shown). B) The specified input-output behavior. C) The posterior probabilities for each model to achieve the toggle switch behavior. D) Parameter posterior distribution, represented by univariate and bivariate marginal distributions, for model 2 (exclusive switch).

Figure 5B shows the desired toggle switch behavior. The inducer S is added between $t = 10$ and $t = 30$, after which the level of protein B should reach a steady state with a mean number of 40. Between $t = 60$ and $t = 80$ the inducer R is added after which the level of protein B should drop to zero. The inducer numbers are both assumed to be 40 molecules which is fixed in the specified range. The desired output behavior was specified via the two component distance metric, defined to be $\epsilon = \{d_1, d_2\}$,

$$d_1 = \sqrt{\frac{\sum_{t\in\alpha}(x_t - y)^2}{n_\alpha}} \quad d_2 = \sqrt{\frac{\sum_{t\in\beta}(x_t - 0)^2}{n_\beta}},$$

where x_t is the number of protein B at time t, y is the target (here fixed at 40), $\alpha = \{t : 30 < t \leq 60\}$, $\beta = \{t : 0 < t \leq 10 \text{ and } 80 < t \leq 100\}$, $n_\alpha = \#\alpha$ and $n_\beta = \#\beta$. The final population was defined to be $\epsilon = \{7.0, 0.05\}$. Here d_1 represents the distance in the "on" region and d_2 represents the distance in the "off" region.

Figure S10 shows the evolution of the stochastic simulations towards the desired behavior. Figure 5C shows the posterior probabilities for each system to achieve the specified behavior, and in particular demonstrates that the exclusive toggle switch outperforms all the others. This chimes with intuition, since the exclusive switch removes the possibility of the deadlock state without the addition of any extra reactions. The fact that the BRD switch performs worse than the original toggle shows that the addition of the two extra degradation reactions does not offer a great enough performance increase for the extra parameters, a manifestation of the parsimony principle or Occam's razor which is inherent to the Bayesian model selection framework used here.

The reactions that comprise the exclusive switch are given by

$$\begin{aligned}
gA &\xrightarrow{k_1} gA + A & A &\xrightarrow{k_2} \emptyset \\
gB &\xrightarrow{k_3} gB + B & B &\xrightarrow{k_4} \emptyset \\
A + gB + P &\xrightarrow{k_5} A - gB & B + gA + P &\xrightarrow{k_6} B - gA \\
A - gB &\xrightarrow{k_7} A + gB + P & B - gA &\xrightarrow{k_8} B + gA + P \\
S + A &\xrightarrow{k_9} S - A & R + B &\xrightarrow{k_{10}} R - B,
\end{aligned}$$

where gA, gB represent the gene promotors for protein A, B respectively and are fixed at one copy, $A - gB, B - gA$ represent the bound transcription factors and S, R represent the inducer molecules. The P species, fixed to be one copy, ensures only A or B can be bound at any one time. Examination of the posterior distribution for this model, Figure 5D, clearly shows a large correlation between k_3 and k_4, which are the production and decay rates of protein B, respectively. This is clearly seen in the principal components (Figure S11) through the combination $k_3 + k_4$ dominating the first PC and the combination $k_4 - k_3$ dominating the last PC. Thus the system is sensitive to only the difference in these rates which is typical in birth-death processes.

7 Discussion

In this paper we have presented a new method for the design of synthetic biological systems employing ideas from Bayesian statistics. We have demonstrated its utility and generality on three different systems spanning biochemical, signaling and genetic networks, as well as oscillatory systems. This method has advantages over traditional design approaches in that the modeling is incorporated directly into the design stage. The statistical nature of the method has many attractive features including the handling of stochastic systems, the ability to perform model selection and the handling of parameter uncertainty in a well defined manner. We used the ABC-SysBio software [19] which takes as input a

set of SBML files and as such can be used by bioengineers and experimentalists to rationally compare their competing designs for a system. By using this method we hope that the implementation time of synthetic systems can be reduced by defining a program of experimental work based on the posterior probabilities of each design.

Monte Carlo sampling of parameter spaces has been used to assess the robustness of engineered and biological systems in the past. But like in the statistical case, simple Monte Carlo sampling tends to waste too much effort and time on those regions which are of no real interest for reverse-engineering or design purposes. Our statistically based sequential approach homes in onto those regions where the probability of observing the desired behavior is appreciable. This allows us a more nuanced comparative assessment of different design proposals, especially when dynamics are expected (or indeed desired) to exhibit elements of stochasticity. And the Bayesian model selection approach automatically strikes a balance between the systems' abilities to generate the desired behavior effectively but also robustly.

Further developments will include the incorporation of methods for model abstraction to reduce computation time [39] and to handle a database of standard parts as in other existing design software systems [40, 41]. Moreover, it is also possible to include the generation of novel structures (by e.g. using stochastic-context free grammars [42] to propose alterations to a reaction/interaction network) as part of the design process. Just like in the case where ideas from control engineering and statistics can gainfully be combined in order to reverse-engineer the structure of naturally evolved biological systems, we feel that in the design of synthetic systems such a union will also be fruitful.

A Approximate Bayesian Computation : ABC SMC

Here we outline the background behind approximate Bayesian computation (ABC) and describe the ABC SMC algorithm [17], which is implemented in the software package ABC-SysBio [19]. ABC methods have been developed to infer posterior distributions in cases where likelihood functions are computationally intractable or too costly to evaluate. They replace the calculation of the likelihood with a comparison between observed and simulated data.

A.1 Background

Let $\theta \in \Theta$ be a parameter vector with prior $\pi(\theta)$ and $f(y|\theta)$ be the likelihood of the data $y \in \mathcal{D}$. In Bayesian inference we are interested in the posterior density

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int_\Theta f(y|\theta)\pi(\theta)d\theta}.$$

Now imagine the case where we cannot write down the likelihood in closed form but we can simulate from the data generating model. We can proceed by first sampling a parameter vector from the prior, $\theta^* \sim \pi(\theta)$, and then sampling a data vector, x^*, from the model conditional on θ^*, ie $x^* \sim f(x|\theta^*)$. This alone gives the joint density $\pi(\theta, x)$. To obtain samples from the posterior distribution we must condition on the data y and this is done via an indicator function, i.e.

$$\pi(\theta, x|y) = \frac{\pi(\theta)f(x|\theta)\mathbb{I}_{\mathcal{A}_y}(x)}{\int_{\mathcal{A}_y \times \Theta} \pi(\theta)f(x|\theta)dxd\theta},$$

where $\mathbb{I}_{\mathcal{B}}(z)$ denotes the indicator function and is equal to 1 for $z \in \mathcal{B}$. Here $\mathcal{A}_y = \{x \in \mathcal{D} : x = y\}$, so the indicator is equal to one when the simulated data and the observed data are identical. This forms a rejection algorithm, and in this instance the accepted θ^* are from the true posterior density $\pi(\theta|y)$.

For most models it is impossible to achieve simulations with outputs in the subset $\mathcal{A}_y$ and so an approximation must be made. This is the basis for ABC. In the first instance we can replace $\mathcal{A}_y$ by $\mathcal{A}_{y,\epsilon} = \{x \in \mathcal{D} : \rho(x, y) \leq \epsilon\}$ where $\rho : \mathcal{D} \times \mathcal{D} \to \mathbb{R}^+$ is a distance function comparing the simulated data to the observed data. We then have

$$\pi_\epsilon(\theta, x|y) = \frac{\pi(\theta) f(x|\theta) \mathbb{I}_{\mathcal{A}_{y,\epsilon}}(x)}{\int_{\mathcal{A}_{y,\epsilon} \times \Theta} \pi(\theta) f(x|\theta) dx d\theta},$$

where π_ϵ is an approximation to the true posterior distribution. The rationale behind ABC is that if ϵ is small then the resulting approximate posterior, π_ϵ, is close to the true posterior. Often, for complex models or stochastic systems, the subset $\mathcal{A}_{y,\epsilon}$ is still too restrictive. In these cases we can resort to comparisons of summary statistics. We now specify the subset $\mathcal{A}_{y,\eta,\epsilon} = \{x \in \mathcal{D} : \rho_S(x, y) \leq \epsilon\}$ where $\eta : \mathcal{D} \to \mathcal{S}$ is a summary statistic and the distance function now takes the form $\rho_S : \mathcal{S} \times \mathcal{S} \to \mathbb{R}^+$. We often write the marginal posterior distribution as $\pi(\theta|\rho(x^*, y) \leq \epsilon)$.

A.2 ABC SMC

The simplest ABC algorithm is known as the ABC rejection algorithm [43] and proceeds as follows

R1 Sample θ^* from $\pi(\theta)$.
R2 Simulate a dataset x^* from $f(x|\theta^*)$.
R3 If $\rho(x^*, y) \leq \epsilon$ accept θ^*, otherwise reject.
R4 Return to R1.

This gives draws from π_ϵ but can be very inefficient in high dimensional models or when the overlap between the prior and posterior distributions is small. One way to improve the efficiency of the rejection algorithm is to perform sequential importance sampling (SIS) [44]. In SIS, instead of sampling directly from the posterior distribution, sampling proceeds via a series of intermediate distributions. The importance distribution at each stage is constructed from a perturbed version of the previous population. This approach can be used in ABC and the resultant algorithm is known as ABC SMC [17]. Described here is a slightly modified version that automatically calculates the ϵ schedule and as such, only the final value, ϵ_T, needs be specified. To obtain N samples $\{\theta^1, \theta^2, \theta^3 ...., \theta^N\}$ (known as particles) from the posterior, defined as, $\pi(\theta|\rho(x^*, y) \leq \epsilon_T)$, proceed as follows

S1 Initialize $\epsilon = \infty$
Set the population indicator $t = 0$
S2.0 Set the particle indicator $i = 1$
S2.1 If $t = 0$, sample θ^{**} independently from $\pi(\theta)$
If $t > 0$, sample θ^* from the previous population $\{\theta^i_{t-1}\}$ with weights w_{t-1}.
Perturb the particle, $\theta^{**} \sim K_t(\theta|\theta^*)$ where K_t is the perturbation kernel.
If $\pi(\theta^{**}) = 0$, return to S2.1
Simulate a candidate dataset $x^* \sim f(x|\theta^{**})$.
If $\rho(x^*, y) > \epsilon$ return to S2.1
S2.2 Set $\theta^i_t = \theta^{**}$ and $d^i_t = \rho(x^*, y)$, calculate the weight as

$$w^i_t = \begin{cases} 1 & \text{if } t = 0 \\ \frac{\pi(\theta^i_t)}{\sum_{j=1}^N w^j_{t-1} K_t(\theta^i_t|\theta^j_{t-1})} & \text{if } t > 0 \end{cases}$$

If $i < N$, set $i = i + 1$, go to S2.1
S3 Normalize the weights.
Determine ϵ such that $Pr(d_t \leq \epsilon) = 0.9$.
If $\epsilon > \epsilon_T$, set $t = t + 1$, go to S2.0.

Here $K_t(\theta|\theta^*)$ is the component-wise random walk perturbation kernel that, in this study, takes the form $K_t(\theta^*|\theta) = \theta + U(-\delta, \delta)$ where $\delta = \frac{1}{2}\text{range}\{\theta_{t-1}\}$. The denominator in the weight calculation can be seen as the probability of observing the current particle given the previous population.

A.3 Model selection

In Bayesian inference comparison of a discrete set of models can be be performed using the marginal posterior. Consider the joint space defined by $(M, \theta) \in \mathcal{M} \times \Theta_{\mathcal{M}}$; Bayes theorem can then be written

$$\pi(M|y) = \frac{f(y|M)\pi(M)}{\int_{\mathcal{M}} f(y|M')\pi(M')dM'} = \frac{f(y|M)\pi(M)}{\sum_{\mathcal{M}} f(y|M')\pi(M')},$$

where $f(y|M)$, the marginal likelihood, can be written

$$f(y|M) = \int_{\Theta_{\mathcal{M}}} \pi(\theta|M) f(y|\theta, M) d\theta.$$

Therefore the posterior probability of a model is given by the normalized marginal likelihood which may or may not be weighted depending on whether the prior over models is informative or uniform respectively. It has recently been noted that model selection using summary statistics can be problematic because the summary statistic must be sufficient for the joint space, $\{M, \theta\}$, rather than just θ [45]. This is not a concern here since in all our examples we use the full data set with no summary or we *define* our posterior distributions through the summary statistics.

Model selection can be incorporated into the ABC framework by introducing the model indicator M and proceeding with inference on the joint space. For example, the ABC rejection algorithm with model selection [46] proceeds as follows

MR1 Sample M^* from $\pi(M)$.
MR2 Sample θ^* from $\pi(\theta|M^*)$.
MR3 Simulate a dataset x^* from $f(x|\theta^*, M^*)$.
MR4 If $\rho(x^*, y) \leq \epsilon$ accept (M^*, θ^*), otherwise reject.
MR5 Return to R1.

Once N samples have been accepted an approximation to the marginal posterior, $\pi(M = m|y)$, is given by

$$\pi(M = m|y) = \frac{\#\text{accepted m}}{N}.$$

Model selection can also be incorporated into the ABC SMC algorithm [18]. To obtain N samples $\{(M, \theta)^1, (M, \theta)^2, (M, \theta)^3 ...., (M, \theta)^N\}$ from the posterior, defined as, $\pi(M, \theta|\rho(x^*, y) \leq \epsilon_T)$, proceed as follows

MS1 Initialize $\epsilon = \infty$
Set the population indicator $t = 0$
MS2.0 Set the particle indicator $i = 1$
MS2.1 If $t = 0$, sample (M^{**}, θ^{**}) from the prior $\pi(M, \theta) = \pi(M)\pi(\theta|M)$.
If $t > 0$, sample M^* with probability $P_{t-1}(M^*)$ and perturb $M^{**} \sim KM_t(M|M^*)$.
Sample θ^* from the previous population $\{\theta(M^{**})_{t-1}\}$ with weights w_{t-1}.
Perturb the particle, $\theta^{**} \sim K_{t,M^{**}}(\theta|\theta^*)$ where $K_{t,M}$ is the perturbation kernel.
If $\pi(M^{**}, \theta^{**}) = 0$, return to MS2.1
Simulate a candidate dataset $x^* \sim f(x|M^{**}, \theta^{**})$.
If $\rho(x^*, y) > \epsilon$ return to MS2.1
MS2.2 Set $(M, \theta)_t^i = (M^{**}, \theta^{**})$ and $d_t^i = \rho(x^*, y)$, calculate the weight as

$$w_t^i(M_t^i, \theta_t^i) = \begin{cases} 1 & \text{if } t = 0 \\ \frac{\pi(M_t^i, \theta_t^i)}{S_1 S_2} & \text{if } t > 0 \end{cases}$$

where
$S_1 = \sum_{j \in \mathcal{M}} P_{t-1}(M_{t-1}^j) KM_t(M_t^i|M_{t-1}^j)$
and

$$S_2 = \sum_{k \in M_t^i = M_{t-1}} \frac{w_{t-1}^k K_{t,M^i}(\theta_t^i|\theta_{t-1}^k)}{P_{t-1}(M_t^i = M_{t-1})}$$

If $i < N$, set $i = i + 1$, go to MS2.1
S3 Normalize the weights.
Obtain the marginal model probabilities given by

$$P_t(M_t = m) = \sum_{k \in M_t^i = M_{t-1}} w_t^i(M_t^i, \theta_t^i)$$

Determine ϵ such that $Pr(d_t \leq \epsilon) = 0.9$.
If $\epsilon > \epsilon_T$, set $t = t + 1$, go to MS2.0.

There are two obvious additions to the algorithm when compared to parameter inference. The model kernel, KM_t, perturbs the resampled models using a multinomial distribution, and the additional term in the weight denominator accounts for the probability of observing the current model given the previous population.

A.4 Prior distribution

The prior distribution encodes our knowledge of the system and should be set according to known biochemical properties. However, often the kinetic parameters are not well known and can be very difficult or even impossible to measure *in vivo*. In these cases we make the prior distribution non informative by specifying a large range over possible, biophysically and biochemically plausible values. As more information becomes available, through experimental studies or otherwise, the prior can be updated to reflect our increased knowledge of the system. Interestingly, for some systems, our design method could help to constrain kinematic parameters where experimental data are unavailable.

A.5 The distance function and output tolerance

In system design we would rarely insist on achieving the true posterior distribution corresponding to $\epsilon = 0$, but would like to reach the objective within some tolerance. A theorem due to Wilkinson (2008) [47] states that if we assume that the data can be considered as

$$y = \eta(\hat{\theta}) + e,$$

where $\eta(\hat{\theta})$ is a draw from the model at the 'best' input and e is an additive, independent error, then the approximate posterior distribution, $\pi(\theta|\rho(x^*, y) \leq \epsilon)$ can be interpreted as the 'true' posterior $\pi(\hat{\theta}|y)$. While the independence assumption is not always true, this theorem provides some insight into the relationship between the final ϵ value and the tolerance on our specified behavior. For example when using uniform kernels, as in this study, if our desired output behavior is a constant of 0.5 and we finish the inference at $\epsilon = 0.05$ our final trajectories will be distributed $U(0.45, 0.55)$ giving a tolerance of $\pm 10\%$ on the output behavior. This can be used when considering our desired output objectives. To achieve other error distributions, such as Gaussian errors, we can always explicitly specify the error model in the design objectives.

A.6 Deterministic models

Inference for deterministic models such as ordinary differential equations can be problematic since there is a one to one relationship between the parameter vector θ and the data set x. Therefore, in the absence of observational error, the posterior distribution resembles a delta function, $\delta(\theta - \hat{\theta})$ where $\hat{x} = f(\hat{\theta})$ is data 'closest' to y. An additional problem for ABC methods is that the minimum distance, $\rho(\hat{x}, y)$, is greater than zero [49]. However, in practice, observational data have associated experimental errors and when this is included explicitly in the model, the problem is resolved. In the case of systems design, we omit the explicit error model for clarity, but note that it could be included with assumptions on the form of the distribution.

B Biochemical adaptation

B.1 Models

We used the same models as those used in [27], which are enzymatic reactions assuming Michaelis-Menten kinetics. Below we give the full models including cooperativity but the more specific case of no cooperativity is when the exponents, n_i, are set to one. Here A, B, C denote the concentrations of the active form of the species and $(1-A), (1-B), (1-C)$ the concentrations of the inactive form. Species E_i and F_i refer to background activating and deactivating enzymes respectively and are assumed to have a constant concentration of 0.5. The models were simulated in the range $0 \leq t \leq 200$.

Design 1

$$\begin{aligned}
\frac{dA}{dt} &= Ik_{IA}\frac{(1-A)^{n_{IA}}}{(1-A)^{n_{IA}} + K_{IA}^{n_{IA}}} - F_A k_{FA}\frac{A^{n_{FA}}}{A^{n_{FA}} + K_{FA}^{n_{FA}}} \\
\frac{dB}{dt} &= Ck_{CB}\frac{(1-B)^{n_{CB}}}{(1-B)^{n_{CB}} + K_{CB}^{n_{CB}}} - F_B k_{FB}\frac{B^{n_{FB}}}{B^{n_{FB}} + K_{FB}^{n_{FB}}} \\
\frac{dC}{dt} &= Ak_{AC}\frac{(1-C)^{n_{AC}}}{(1-C)^{n_{AC}} + K_{AC}^{n_{AC}}} - Bk_{BC}\frac{C^{n_{BC}}}{C^{n_{BC}} + K_{BC}^{n_{BC}}}
\end{aligned}$$

Design 2

$$\begin{aligned}
\frac{dA}{dt} &= Ik_{IA}\frac{(1-A)^{n_{IA}}}{(1-A)^{n_{IA}} + K_{IA}^{n_{IA}}} - F_A k_{FA}\frac{A^{n_{FA}}}{A^{n_{FA}} + K_{FA}^{n_{FA}}} \\
\frac{dB}{dt} &= E_B k_{EB}\frac{(1-B)^{n_{EB}}}{(1-B)^{n_{EB}} + K_{EB}^{n_{EB}}} - Ck_{CB}\frac{B^{n_{CB}}}{B^{n_{CB}} + K_{CB}^{n_{CB}}} \\
\frac{dC}{dt} &= Ak_{AC}\frac{(1-C)^{n_{AC}}}{(1-C)^{n_{AC}} + K_{AC}^{n_{AC}}} - Bk_{BC}\frac{C^{n_{BC}}}{C^{n_{BC}} + K_{BC}^{n_{BC}}} - F_C k_{FC}\frac{C^{n_{FC}}}{C^{n_{FC}} + K_{FC}^{n_{FC}}}
\end{aligned}$$

Design 3

$$\frac{dA}{dt} = Ik_{IA}\frac{(1-A)^{n_{IA}}}{(1-A)^{n_{IA}} + K_{IA}^{n_{IA}}} - F_A k_{FA}\frac{A^{n_{FA}}}{A^{n_{FA}} + K_{FA}^{n_{FA}}}$$
$$\frac{dB}{dt} = E_B k_{EB}\frac{(1-B)^{n_{EB}}}{(1-B)^{n_{EB}} + K_{EB}^{n_{EB}}} - Ck_{CB}\frac{B^{n_{CB}}}{B^{n_{CB}} + K_{CB}^{n_{CB}}}$$
$$\frac{dC}{dt} = Bk_{BC}\frac{(1-C)^{n_{BC}}}{(1-C)^{n_{BC}} + K_{BC}^{n_{BC}}} - Ak_{AC}\frac{C^{n_{AC}}}{C^{n_{AC}} + K_{AC}^{n_{AC}}}$$

Design 4

$$\frac{dA}{dt} = Ik_{IA}\frac{(1-A)^{n_{IA}}}{(1-A)^{n_{IA}} + K_{IA}^{n_{IA}}} - Bk_{BA}\frac{A^{n_{BA}}}{A^{n_{BA}} + K_{BA}^{n_{BA}}}$$
$$\frac{dB}{dt} = Ak_{AB}\frac{(1-B)^{n_{AB}}}{(1-B)^{n_{AB}} + K_{AB}^{n_{AB}}} - F_B k_{FB}\frac{B^{n_{FB}}}{B^{n_{FB}} + K_{FB}^{n_{FB}}}$$
$$\frac{dC}{dt} = Ak_{AC}\frac{(1-C)^{n_{AC}}}{(1-C)^{n_{AC}} + K_{AC}^{n_{AC}}} - F_C k_{FC}\frac{C^{n_{FC}}}{C^{n_{FC}} + K_{FC}^{n_{FC}}}$$

Design 5

$$\frac{dA}{dt} = Ik_{IA}\frac{(1-A)^{n_{IA}}}{(1-A)^{n_{IA}} + K_{IA}^{n_{IA}}} - Bk_{BA}\frac{A^{n_{BA}}}{A^{n_{BA}} + K_{BA}^{n_{BA}}}$$
$$\frac{dB}{dt} = Ak_{AB}\frac{(1-B)^{n_{AB}}}{(1-B)^{n_{AB}} + K_{AB}^{n_{AB}}} - F_B k_{FB}\frac{B^{n_{FB}}}{B^{n_{FB}} + K_{FB}^{n_{FB}}}$$
$$\frac{dC}{dt} = E_C k_{EC}\frac{(1-C)^{n_{EC}}}{(1-C)^{n_{EC}} + K_{EC}^{n_{EC}}} - Ak_{AC}\frac{C^{n_{AC}}}{C^{n_{AC}} + K_{AC}^{n_{AC}}}$$

Design 6

$$\frac{dA}{dt} = Ik_{IA}\frac{(1-A)^{n_{IA}}}{(1-A)^{n_{IA}} + K_{IA}^{n_{IA}}} - Bk_{BA}\frac{A^{n_{BA}}}{A^{n_{BA}} + K_{BA}^{n_{BA}}}$$
$$\frac{dB}{dt} = Ck_{CB}\frac{(1-B)^{n_{CB}}}{(1-B)^{n_{CB}} + K_{CB}^{n_{CB}}} - F_B k_{FB}\frac{B^{n_{FB}}}{B^{n_{FB}} + K_{FB}^{n_{FB}}}$$
$$\frac{dC}{dt} = Ak_{AC}\frac{(1-C)^{n_{AC}}}{(1-C)^{n_{AC}} + K_{AC}^{n_{AC}}} - F_C k_{FC}\frac{C^{n_{FC}}}{C^{n_{FC}} + K_{FC}^{n_{FC}}}$$

Design 7

$$\frac{dA}{dt} = Ik_{IA}\frac{(1-A)^{n_{IA}}}{(1-A)^{n_{IA}} + K_{IA}^{n_{IA}}} - Bk_{BA}\frac{A^{n_{BA}}}{A^{n_{BA}} + K_{BA}^{n_{BA}}} - F_A k_{FA}\frac{A^{n_{FA}}}{A^{n_{FA}} + K_{FA}^{n_{FA}}}$$
$$\frac{dB}{dt} = E_B k_{EB}\frac{(1-B)^{n_{EB}}}{(1-B)^{n_{EB}} + K_{EB}^{n_{EB}}} - Ck_{CB}\frac{B^{n_{CB}}}{B^{n_{CB}} + K_{CB}^{n_{CB}}}$$
$$\frac{dC}{dt} = Ak_{AC}\frac{(1-C)^{n_{AC}}}{(1-C)^{n_{AC}} + K_{AC}^{n_{AC}}} - F_C k_{FC}\frac{C^{n_{FC}}}{C^{n_{FC}} + K_{FC}^{n_{FC}}}$$

Design 8

$$\frac{dA}{dt} = Ik_{IA}\frac{(1-A)^{n_{IA}}}{(1-A)^{n_{IA}} + K_{IA}^{n_{IA}}} - Bk_{BA}\frac{A^{n_{BA}}}{A^{n_{BA}} + K_{BA}^{n_{BA}}} - F_A k_{FA}\frac{A^{n_{FA}}}{A^{n_{FA}} + K_{FA}^{n_{FA}}}$$
$$\frac{dB}{dt} = Ck_{CB}\frac{(1-B)^{n_{CB}}}{(1-B)^{n_{CB}} + K_{CB}^{n_{CB}}} - F_B k_{FB}\frac{B^{n_{FB}}}{B^{n_{FB}} + K_{FB}^{n_{FB}}}$$
$$\frac{dC}{dt} = E_C k_{EC}\frac{(1-C)^{n_{EC}}}{(1-C)^{n_{EC}} + K_{EC}^{n_{EC}}} - Ak_{AC}\frac{C^{n_{AC}}}{C^{n_{AC}} + K_{AC}^{n_{AC}}}$$

Design 9

$$\begin{aligned}
\frac{dA}{dt} &= Ik_{IA}\frac{(1-A)^{n_{IA}}}{(1-A)^{n_{IA}}+K_{IA}^{n_{IA}}} - Bk_{BA}\frac{A^{n_{BA}}}{A^{n_{BA}}+K_{BA}^{n_{BA}}} \\
\frac{dB}{dt} &= E_B k_{EB}\frac{(1-B)^{n_{EB}}}{(1-B)^{n_{EB}}+K_{EB}^{n_{EB}}} - Ck_{CB}\frac{B^{n_{CB}}}{B^{n_{CB}}+K_{CB}^{n_{CB}}} \\
\frac{dC}{dt} &= E_C k_{EC}\frac{(1-C)^{n_{EC}}}{(1-C)^{n_{EC}}+K_{EC}^{n_{EC}}} - Ak_{AC}\frac{C^{n_{AC}}}{C^{n_{AC}}+K_{AC}^{n_{AC}}}
\end{aligned}$$

Design 10

$$\begin{aligned}
\frac{dA}{dt} &= Ik_{IA}\frac{(1-A)^{n_{IA}}}{(1-A)^{n_{IA}}+K_{IA}^{n_{IA}}} - F_A k_{FA}\frac{A^{n_{FA}}}{A^{n_{FA}}+K_{FA}^{n_{FA}}} \\
\frac{dB}{dt} &= Ak_{AB}\frac{(1-B)^{n_{AB}}}{(1-B)^{n_{AB}}+K_{AB}^{n_{AB}}} - F_B k_{FB}\frac{B^{n_{FB}}}{B^{n_{FB}}+K_{FB}^{n_{FB}}} \\
\frac{dC}{dt} &= Bk_{BC}\frac{(1-C)^{n_{BC}}}{(1-C)^{n_{BC}}+K_{BC}^{n_{BC}}} - Ak_{AC}\frac{C^{n_{AC}}}{C^{n_{AC}}+K_{AC}^{n_{AC}}}
\end{aligned}$$

Design 11

$$\begin{aligned}
\frac{dA}{dt} &= Ik_{IA}\frac{(1-A)^{n_{IA}}}{(1-A)^{n_{IA}}+K_{IA}^{n_{IA}}} - F_A k_{FA}\frac{A^{n_{FA}}}{A^{n_{FA}}+K_{FA}^{n_{FA}}} \\
\frac{dB}{dt} &= Ak_{AB}\frac{(1-B)^{n_{AB}}}{(1-B)^{n_{AB}}+K_{AB}^{n_{AB}}} - F_B k_{FB}\frac{B^{n_{FB}}}{B^{n_{FB}}+K_{FB}^{n_{FB}}} \\
\frac{dC}{dt} &= Ak_{AC}\frac{(1-C)^{n_{AC}}}{(1-C)^{n_{AC}}+K_{AC}^{n_{AC}}} - Bk_{BC}\frac{C^{n_{BC}}}{C^{n_{BC}}+K_{BC}^{n_{BC}}}
\end{aligned}$$

B.2 Distance

The two component distance metric was defined to be $\epsilon = \{E, S^{-1}\}$, where E and S are the adaptation efficiency and sensitivity defined by

$$\begin{aligned}
E &= \left|\frac{(O_2 - O_1)/O_1}{(I_2 - I_1)/I_1}\right| \\
S &= \left|\frac{(O_{peak} - O_1)/O_1}{(I_2 - I_1)/I_1}\right|,
\end{aligned}$$

where I_1, I_2 are the input values (here fixed at 0.5 and 0.6 respectively), O_1, O_2 are the output steady state levels before and after the input change and and O_{peak} is the maximal transient output level. The final population was defined to be $\epsilon = \{0.1, 1.0\}$.

B.3 Priors

The priors on the Michaelis-Menten rates were chosen to correspond to the parameter ranges used in the original study; $\log k \sim U(-1, 1)$ and $\log K \sim U(-3, 2)$ [27].

C Robust oscillator design

C.1 Models

We used the same models as those used in [32], simulated in the range $0 \leq t \leq 10$. Again A, B, C denote the concentrations of the active form of the species and $(1-A), (1-B), (1-C)$ the concentrations of the inactive form. The feedback is modeled using Michaelis-Menten kinetics but the conversion of inactive form into active form is assumed to have a constant rate.

Design 1

$$\begin{aligned}
\frac{dA}{dt} &= k_1(1-A) - \frac{k_2 C^{n_1}}{K_1^{n_1} + C^{n_1}} A \\
\frac{dB}{dt} &= k_3(1-B) - \frac{k_4 A^{n_2}}{K_2^{n_2} + A^{n_2}} B \\
\frac{dC}{dt} &= k_5(1-C) - \frac{k_6 B^{n_3}}{K_3^{n_3} + B^{n_3}} C
\end{aligned}$$

Designs 2 and 3

$$\begin{aligned}
\frac{dA}{dt} &= k_1(1-A) - \frac{k_2 C^{n_1}}{K_1^{n_1} + C^{n_1}} A + k_7(1-A)\frac{A^{n_4}}{K_4^{n_4} + A^{n_4}} \\
\frac{dB}{dt} &= k_3(1-B) - \frac{k_4 A^{n_2}}{K_2^{n_2} + A^{n_2}} B \\
\frac{dC}{dt} &= k_5(1-C) - \frac{k_6 B^{n_3}}{K_3^{n_3} + B^{n_3}} C
\end{aligned}$$

Designs 4 and 5

$$\begin{aligned}
\frac{dA}{dt} &= k_1(1-A) - \frac{k_2 C^{n_1}}{K_1^{n_1} + C^{n_1}} A - k_7 A\frac{A^{n_4}}{K_4^{n_4} + A^{n_4}} \\
\frac{dB}{dt} &= k_3(1-B) - \frac{k_4 A^{n_2}}{K_2^{n_2} + A^{n_2}} B \\
\frac{dC}{dt} &= k_5(1-C) - \frac{k_6 B^{n_3}}{K_3^{n_3} + B^{n_3}} C
\end{aligned}$$

C.2 Distance

For the direct Hopf bifurcation detection the distance metric was defined to be

$$\epsilon = \frac{\prod_i \mathrm{Re}[\lambda_i]}{\prod_i (1 - 0.99\exp(-|\,\mathrm{Im}[\lambda_i]|))}$$

where λ_i is the i^{th} complex eigenvalue of the linearized system in the steady state. Here $\epsilon = 0$ represents the location in parameter space where a limit cycle emerges through a Hopf bifurcation [48]. The final population was at $\epsilon = 0.001$.

To investigate the ability to achieve particular amplitude-frequency values, the distance was defined as $\epsilon = \{d_1, d_2, d_3\}$, where

$$\begin{aligned}
d_1 &= \sum_n |x_{t_0+nT} - x_{t_0+(n-1)T}| \\
d_2 &= |f_t - f| \\
d_3 &= |\max x_{t>t_0} - \min x_{t>t_0} - A_t|,
\end{aligned}$$

and n is an integer, f_t is the target frequency, f is the frequency determined from the largest component of the Fourier spectrum, A_t is the target amplitude and t_0 is a cut to remove initial transients (= 2s). The final population was defined to be $\epsilon = \{0.05, 0.05, 0.05\}$.

C.3 Priors

The priors were chosen to correspond to parameter ranges used in the original study; $k_1 \sim U(0, 10)$, $k_2 \sim U(0, 1000)$, $k_3 \sim U(0, 10)$, $k_4 \sim U(0, 1000)$, $k_6 \sim U(0, 1000)$, $k_7 \sim U(0, 100)$, $k_7^{strong} \sim U(500, 600)$, $n_i \sim U(1, 4)$ and $K_i \sim U(0, 4)$ [32].

D Bacterial two component systems

D.1 Models

The models we used were based on the ones found in [34]. All simulations were performed in the range $0 \leq t \leq 10$.

Orthodox system

We modeled the following reactions

$$HK + S \xrightarrow{k_1} HKp + S$$
$$HKp + RR \xrightarrow{k_2} HK + RRp$$
$$HKp \xrightarrow{k_3} HK$$
$$HK + RRp \xrightarrow{k_4} HKp + RR$$
$$RRp \xrightarrow{k_5} RR.$$

Additionally we assumed that the total concentration of $HK_{tot} = HK + HKp$ and $RR_{tot} = RR + RRp$ were equal to one. This resulted in the following ordinary differential equations

$$\frac{d[HK]}{dt} = k_2[HKp][RR] + k_3[HKp] - k_4[HK][RRp] - k_1[HK][S]$$
$$\frac{d[RRp]}{dt} = k_2[HKp][RR] - k_4[HK][RRp] - k_5[RRp].$$

Orthodox system

We labelled the occupied states of the phosphorelay as

	H1	D1	H2
HK_1	x	x	x
HK_2	o	x	x
HK_3	x	o	x
HK_4	x	x	o
HK_5	o	o	x
HK_6	o	x	o
HK_7	x	o	o
HK_8	o	o	o

where H1, D1 and H2 are the binding domains on the Histidine Kinase and x, o represent an empty, occupied domain respectively. We modeled the following reactions

$$HK + S \xrightarrow{k_1} HKp + S$$

Again we assumed that the total concentration of $HK_{tot} = \sum HK_i$ and $RR_{tot} = RR + RRp$ were equal to one. This resulted in the following ordinary differential equations

$$\frac{dHK_1}{dt} = k_4[HK_4][RR] + k_6[HK_3] - k_7[HK_1][RRp] + k_8[HK_2] - k1[HK_1][S]$$
$$\frac{dHK_2}{dt} = k_4[HK_6][RR] + k_6[HK_5] - k_7[HK_2][RRp] - k_8[HK_2] + k_1[HK_1][S] - k_2[HK_2]$$
$$\frac{dHK_3}{dt} = -k_3[HK_3] + k_4[HK_7][RR] + k_5[HK_4] - k_6[HK_3] - k_7[HK_3][RRp] + k_8[HK_5] - k_1[HK_3][S] + k_2[HK_2]$$
$$\frac{dHK_4}{dt} = k_3[HK_3] - k_4[HK_4][RR] - k_5[HK_4] + k_6[HK_7] + k_7[HK_1][RRp] + k_8[HK_6] - k_1[HK_4][S]$$
$$\frac{dHK_5}{dt} = -k_3[HK_3] + k_4[HK_8][RR] + k_5[HK_6] - k_6[HK_5] - k_7[HK_5][RRp] - k_8[HK_5] + k_1[HK_3][S]$$
$$\frac{dHK_6}{dt} = k_3[HK_5] - k_2[HK_6] - k_4[HK_6][RR] - k_5[HK_6] + k_6[HK_8] + k_7[HK_2][RRp] - k_8[HK_6] + k_1[HK_4][S]$$
$$\frac{dHK_7}{dt} = k_2[HK_6] - k_4[HK_7][RR] - k_6[HK_7] + k_7[HK_3][RRp] + k_8[HK_8] - k_1[HK_7][S]$$
$$\frac{dRRp}{dt} = k_4[RR]([HK_4] + [HK_6] + [HK_7] + [HK_8]) - k_7[RRp]([HK_1] + [HK_2] + [HK_3] + [HK_5]) - k_9[RRp]$$

D.2 Distance

The distance functions for input-output behaviors ϵ_{1-4} were defined to be

$$\begin{aligned}
\epsilon_1 &= \Big\{\mathbb{H}(0)(\text{argmax}_t\, x_t - 2.0), \mathbb{H}(0)(\text{argmin}_t\, x_t - 4.0)\Big\} \\
\epsilon_2 &= \sqrt{\sum_t (x_t - 1.0)^2} \\
\epsilon_3 &= \sqrt{\sum_t (x_t - 0.5)^2} \\
\epsilon_4 &= \Big\{\epsilon_1, \mathbb{H}(0)(1 - \max x_t - 0.2), \mathbb{H}(0)(\min x_t - 0.2)\Big\}
\end{aligned}$$

where $\mathbb{H}(0)$ is the Heaviside function ensuring the distance is positive.

D.3 Priors

The priors on all variables were distributed as $U(0, 1000)$.

E Stochastic genetic toggle switch

E.1 Models

We modeled each toggle switch using a continuous time Markov jump process which obeys the chemical master equation. We neglected processes at the RNA level and just modeled at the protein level. This makes the models simpler while retaining all the relevant behavior. In all the following gA, gB represent the gene promotor for protein A, B respectively and are fixed at one copy. $A - gB, B - gA$ represent the bound transcription factors and S, R represent the switch and reset signals. Because the concentration of these are fixed they have the effect of removing A and B from the system respectively.

The models were simulated in the range $0 \leq t \leq 200$

Design 1

$$
\begin{aligned}
gA &\xrightarrow{k_1} gA + A \\
A &\xrightarrow{k_2} \emptyset \\
gB &\xrightarrow{k_3} gB + B \\
B &\xrightarrow{k_4} \emptyset \\
A + gB &\xrightarrow{k_5} A - gB \\
B + gA &\xrightarrow{k_6} B - gA \\
A - gB &\xrightarrow{k_7} A + gB \\
B - gA &\xrightarrow{k_8} B + gA \\
S + A &\xrightarrow{k_9} S - A \\
R + B &\xrightarrow{k_{10}} R - B
\end{aligned}
$$

Design 2

Here, in addition to the species in design 1, we have introduced the P species, fixed to be one copy, which ensures only A or B can be bound at any one time

$$
\begin{aligned}
gA &\xrightarrow{k_1} gA + A \\
A &\xrightarrow{k_2} \emptyset \\
gB &\xrightarrow{k_3} gB + B \\
B &\xrightarrow{k_4} \emptyset \\
A + gB + P &\xrightarrow{k_5} A - gB \\
B + gA + P &\xrightarrow{k_6} B - gA \\
A - gB &\xrightarrow{k_7} A + gB + P \\
B - gA &\xrightarrow{k_8} B + gA + P \\
S + A &\xrightarrow{k_9} S - A \\
R + B &\xrightarrow{k_{10}} R - B.
\end{aligned}
$$

Design 3

Here we have the same reactions in design 1 but include two extra reactions for the decay of the bound proteins

$$
\begin{aligned}
A - gB &\xrightarrow{k_{11}} gB \\
B - gA &\xrightarrow{k_{12}} gA.
\end{aligned}
$$

Design 4

Here we have the same reactions in design 1 but include two extra reactions for the binding /unbinding

of the proteins A and B

$$A + B \xrightarrow{k_{11}} A - B$$
$$A - B \xrightarrow{k_{12}} A + B.$$

E.2 Distance

The two component distance metric was defined to be $\epsilon = \{d_1, d_2\}$,

$$\begin{aligned} d_1 &= \sqrt{\frac{\sum_{t\in\alpha}(x_t - y)^2}{n_\alpha}} \\ d_2 &= \sqrt{\frac{\sum_{t\in\beta}(x_t - 0)^2}{n_\beta}}, \end{aligned}$$

where x_t is the number of protein B at time t, y is the target (here fixed at 40), $\alpha = \{t : 30 < t \leq 60\}$, $\beta = \{t : 0 < t \leq 10 \text{ and } 80 < t \leq 100\}$, $n_\alpha = \#\alpha$ and $n_\beta = \#\beta$. The final population was defined to be $\epsilon = \{7.0, 0.05\}$.

E.3 Priors

The priors for production, binding and interaction rates were distributed as $U(0, 50)$ and the priors for the degradation rates were given $U(0, 5)$ distributions.

References

[1] Martin VJJ, Pitera DJ, Withers ST, Newman JD, Keasling JD (2003) Engineering a mevalonate pathway in escherichia coli for production of terpenoids. *Nat Biotechnol* 21:796–802.

[2] Ro DK, *et al.* (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* 440:940–3.

[3] You L, Cox RS, Weiss R, Arnold FH (2004) Programmed population control by cell-cell communication and regulated killing. *Nature* 428:868–71.

[4] Kobayashi H, *et al.* (2004) Programmable cells: interfacing natural and engineered gene networks. *Proc Natl Acad Sci USA* 101:8414–9.

[5] Fortman JL, *et al.* (2008) Biofuel alternatives to ethanol: pumping the microbial well. *Trends Biotechnol* 26:375–81.

[6] Savage DF, Way J, Silver PA (2008) Defossiling fuel: how synthetic biology can transform biofuel production. *ACS Chem Biol* 3:13–6.

[7] Cases I, de Lorenzo V (2005) Genetically modified organisms for the environment: stories of success and failure and what we have learned from them. *Int Microbiol* 8:213–22.

[8] Takahashi K, *et al.* (2007) Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* 131:861–72.

[9] Hanna JH, Saha K, Jaenisch R (2010) Pluripotency and cellular reprogramming: facts, hypotheses, unresolved issues. *Cell* 143:508–25.

[10] Lu TK, Khalil AS, Collins JJ (2009) Next-generation synthetic gene networks. *Nat Biotechnol* 27:1139–50.

[11] Macarthur BD, Ma'ayan A, Lemischka IR (2009) Systems biology of stem cell fate and cellular reprogramming. *Nat Rev Mol Cell Biol* 10:672–81.

[12] Anderson JC, Clarke EJ, Arkin AP, Voigt CA (2006) Environmentally controlled invasion of cancer cells by engineered bacteria. *J Mol Biol* 355:619–27.

[13] Rajendran M, Ellington AD (2008) Selection of fluorescent aptamer beacons that light up in the presence of zinc. *Anal Bioanal Chem* 390:1067–75.

[14] Canton B, Labno A, Endy D (2008) Refinement and standardization of synthetic biological parts and devices. *Nat Biotechnol* 26:787–93.

[15] Liang W, Shores MP, Bockrath M, Long JR, Park H (2002) Kondo resonance in a single-molecule transistor. *Nature* 417:725–9.

[16] Purnick PEM, Weiss R (2009) The second wave of synthetic biology: from modules to systems. *Nat Rev Mol Cell Biol* 10:410–22.

[17] Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH (2009) Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface* 6:187–202.

[18] Toni T, Stumpf MPH (2010) Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics* 26:104–10.

[19] Liepe J, *et al.* (2010) ABC-SysBio–Approximate Bayesian computation in Python with GPU support. *Bioinformatics* 26:1797–9.

[20] Bray D, Lay S (1994) Computer simulated evolution of a network of cell-signaling molecules. *Biophysical Journal* 66:972–7.

[21] François P, Hakim V (2004) Design of genetic networks with specified functions by evolution in silico. *Proc Natl Acad Sci USA* 101:580–5.

[22] Battogtokh D, Asch DK, Case ME, Arnold J, Schuttler HB (2002) An ensemble method for identifying regulatory circuits with special reference to the qa gene cluster of neurospora crassa. *Proc Natl Acad Sci USA* 99:16904–9.

[23] Feng XJ, *et al.* (2004) Optimizing genetic circuits by global sensitivity analysis. *Biophysical Jour-*

nal 87:2195–202.
[24] Rodrigo G, Carrera J, Jaramillo A (2007) Genetdes: automatic design of transcriptional networks. *Bioinformatics* 23:1857–8.
[25] Dasika MS, Maranas CD (2008) Optcircuit: an optimization based method for computational design of genetic circuits. *BMC systems biology* 2:24.
[26] Batt G, Yordanov B, Weiss R, Belta C (2007) Robustness analysis and tuning of synthetic gene networks. *Bioinformatics* 23:2415–22.
[27] Ma W, Trusina A, El-Samad H, Lim WA, Tang C (2009) Defining network topologies that can achieve biochemical adaptation. *Cell* 138:760–73.
[28] Elowitz MB, Leibler S (2000) A synthetic oscillatory network of transcriptional regulators. *Nature* 403:335–8.
[29] Stricker J, Cookson S, Bennett MR, Mather WH, Tsimring LS, Hasty J (2008) A fast, robust and tunable synthetic gene oscillator. *Nature* 456:516–9.
[30] Tigges M, Marquez-Lago TT, Stelling J, Fussenegger M (2009) A tunable synthetic mammalian oscillator. *Nature* 457:309–12.
[31] Purcell O, Savery NJ, Grierson CS, di Bernardo M (2010) A comparative analysis of synthetic genetic oscillators. *Journal of the Royal Society Interface* 7:1503–1524.
[32] Tsai TYC, Choi YS, Ma W, Pomerening JR, Tang C, Ferrell JE (2008) Robust, Tunable Biological Oscillations from Interlinked Positive and Negative Feedback Loops. *Science* 321:126–129.
[33] Stock AM, Robinson VL, Goudreau PN (2000) Two-component signal transduction. *Annu Rev Biochem* 69:183–215.
[34] Kim JR, Cho KH (2006) The multi-step phosphorelay mechanism of unorthodox two-component systems in e. coli realizes ultrasensitivity to stimuli while maintaining robustness to noises. *Comput Biol Chem* 30:438–44.
[35] Csikász-Nagy A, Cardelli L, Soyer OS (2010) Response dynamics of phosphorelays suggest their potential utility in cell signalling. *Journal of the Royal Society Interface* :Epub ahead of print.
[36] Shinar G, Milo R, Martínez MR, Alon U (2007) Input output robustness in simple bacterial signaling systems. *Proc Natl Acad Sci USA* 104:19931–5.
[37] Gardner TS, Cantor CR, Collins JJ (2000) Construction of a genetic toggle switch in escherichia coli. *Nature* 403:339–42.
[38] Lipshtat A, Loinger A, Balaban NQ, Biham O (2006) Genetic toggle switch without cooperative binding. *Phys Rev Lett* 96:188101.
[39] Myers CJ, *et al.* (2009) ibiosim: a tool for the analysis and design of genetic circuits. *Bioinformatics* 25:2848–9.
[40] Hill AD, Tomshine JR, Weeding EMB, Sotiropoulos V, Kaznessis YN (2008) Synbioss: the synthetic biology modeling suite. *Bioinformatics* 24:2551–3.
[41] Marchisio MA, Stelling J (2008) Computational design of synthetic gene circuits1 with composable parts. *Bioinformatics* 24:1903–10.
[42] Baldi P, Brunak Sa (2001) *Bioinformatics: The Machine Learning Approach, Second Edition (Adaptive Computation and Machine Learning)* (The MIT Press), 2 edn.
[43] Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol* 16:1791–8.
[44] Del Moral P, Doucet A, Jasra A (2006) Sequential Monte Carlo samplers. *J. Roy. Stat. Soc. B* 68:411–436.
[45] Robert CP, Cornuet J-M, Marin J-M, Pillai NS (2011) Lack of confidence in ABC model choice. *arXiv* 1102.4432v1.
[46] Grelaud A, Robert CP, Marin, JM (2009) ABC methods for model choice in Gibbs random fields. *Cr Math* 347:205–210.
[47] Wilkinson RD (2008) Approximate Bayesian computation (ABC) gives exact results under the

assumption of model error. *arXiv* 0811.3355v1.

[48] Chickarmane V, Paladugu SR, Bergmann F, Sauro, HM (2005) Bifurcation discovery tool. *Bioinformatics* 21:3688–90.

[49] Toni T. (2010) Approximate Bayesian computation for parameter inference and model selection in systems biology. *PhD Thesis, University of London, UK*.

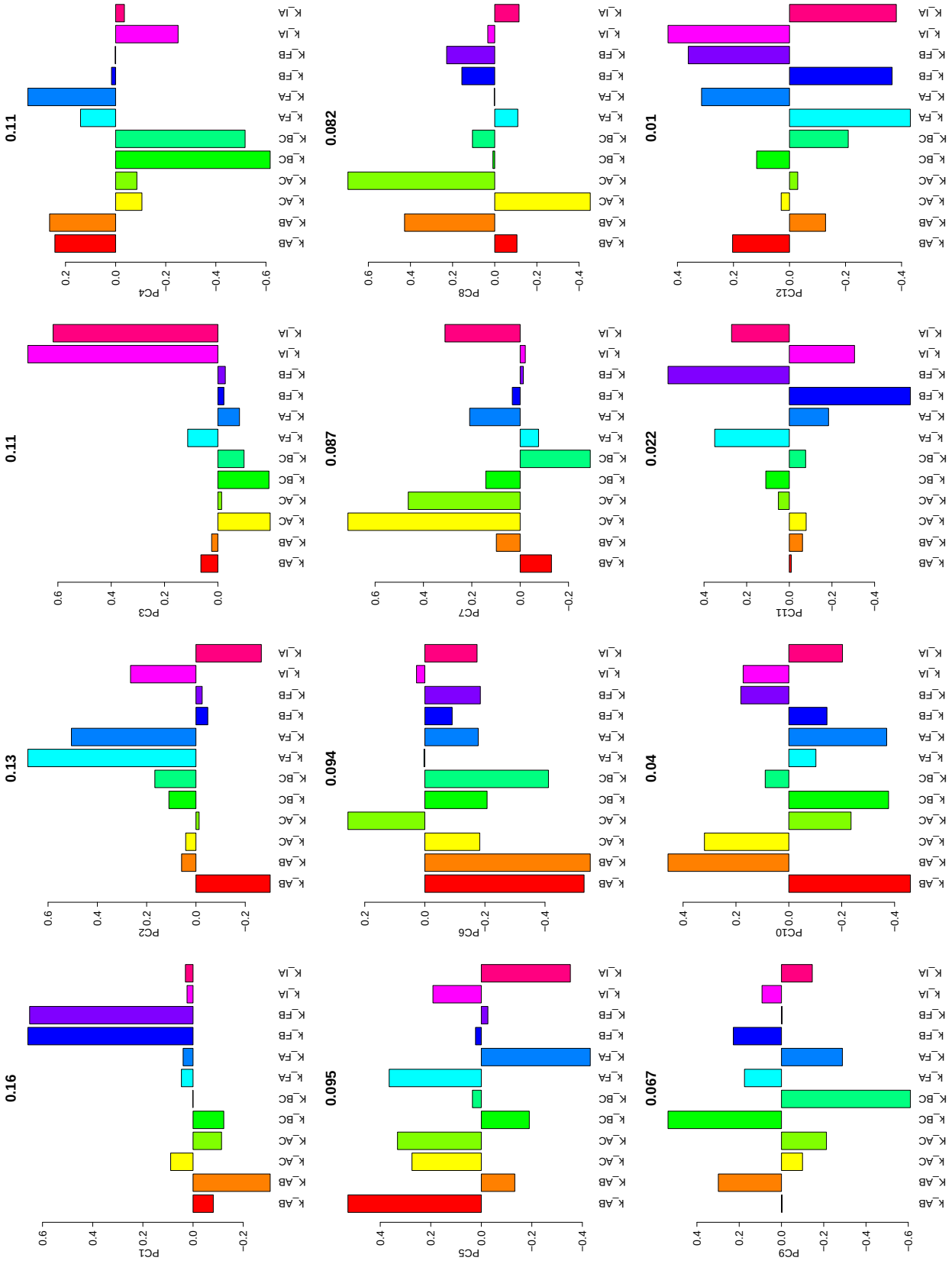

Figure S1: Biochemical adaptation: principal component analysis of the posterior distribution for model 11 in the case of no cooperativity.

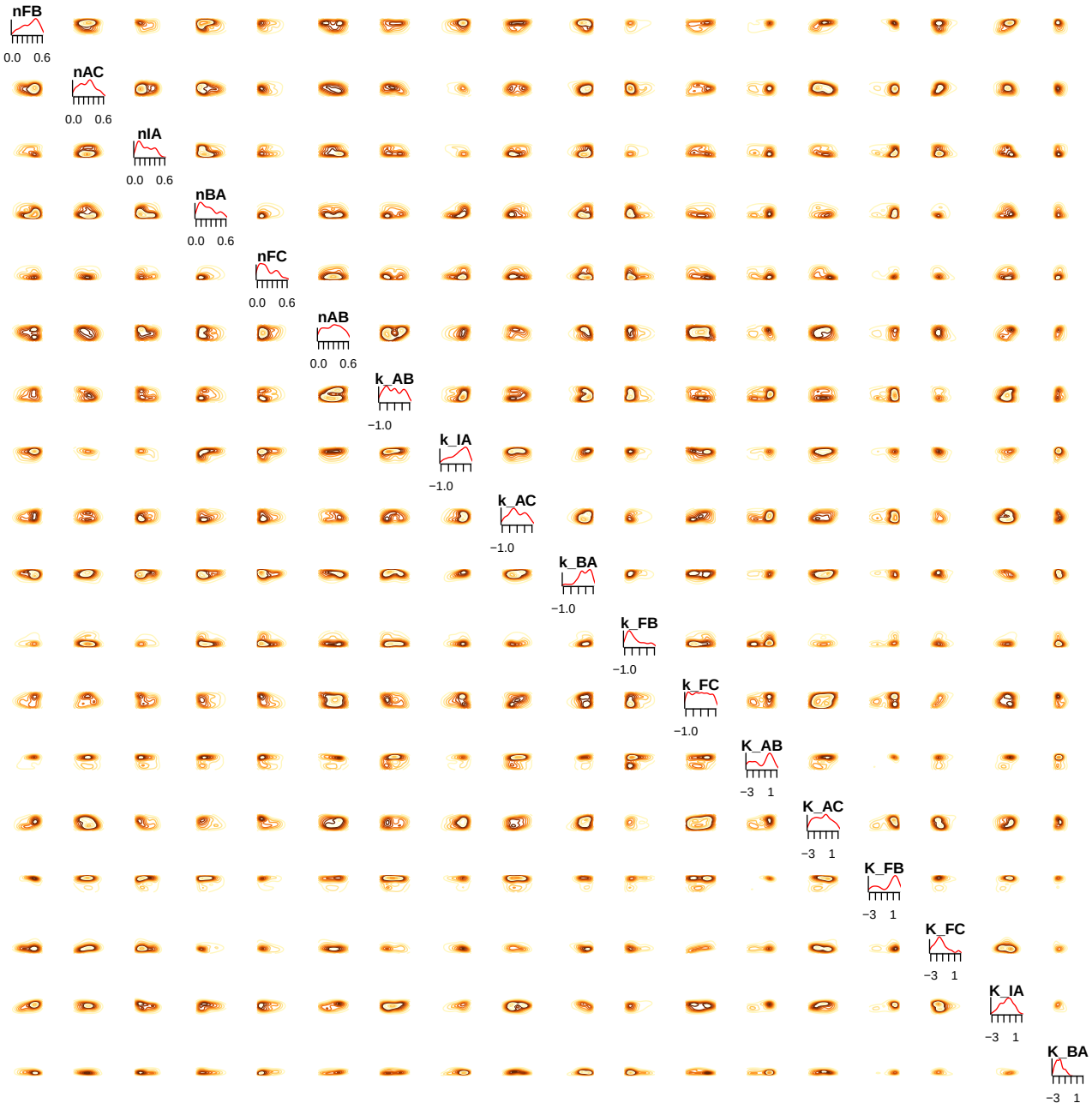

Figure S2: Biochemical adaptation: posterior distribution for model 4 in the case when cooperativity is included.

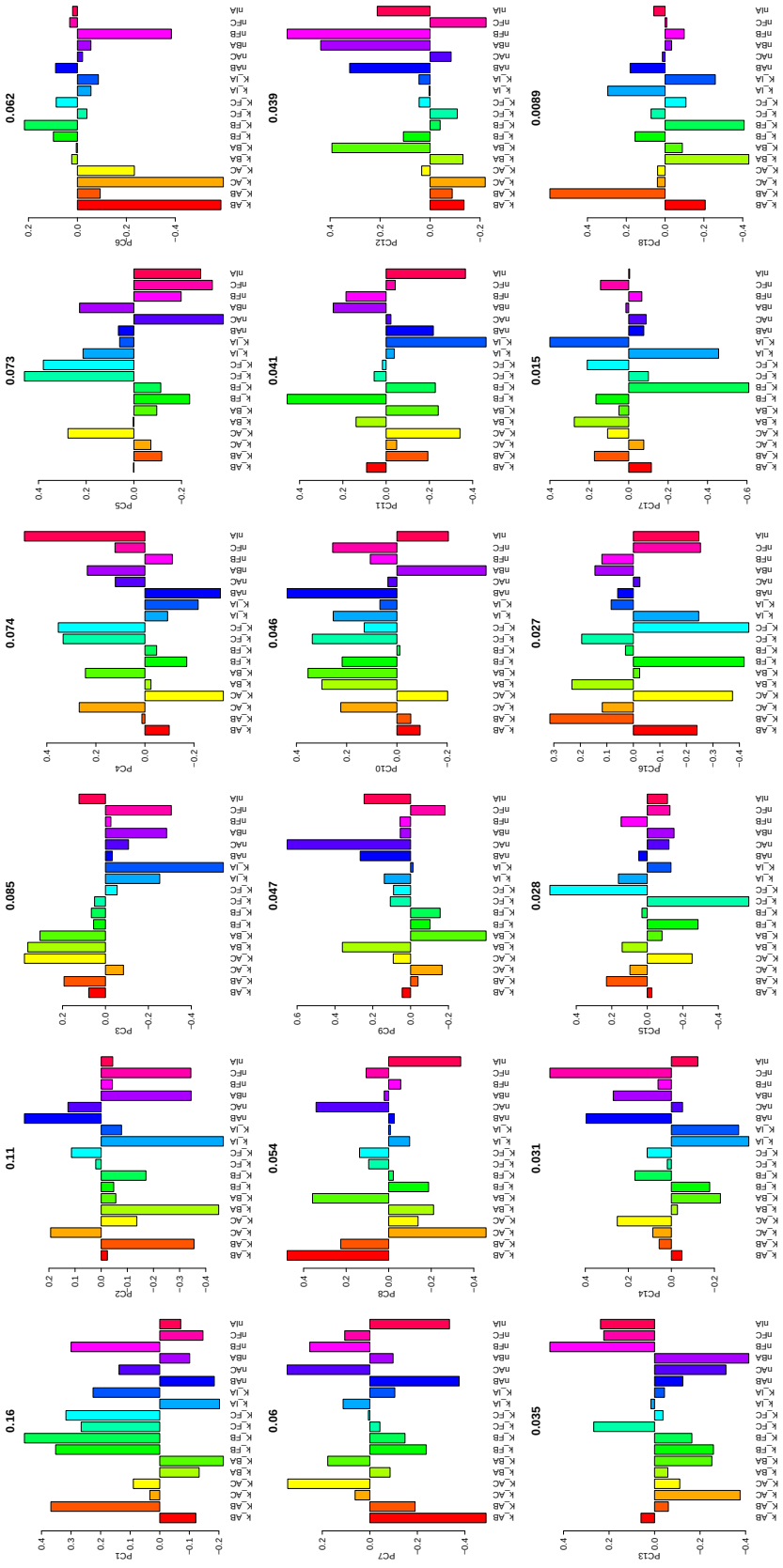

Figure S3: Biochemical adaptation: principal component analysis of the posterior distribution for model 4 in the case when cooperativity is included.

Figure S4: Robust oscillator design: posterior distribution for model 2 to achieve limits cycles via a hopf bifurcation.

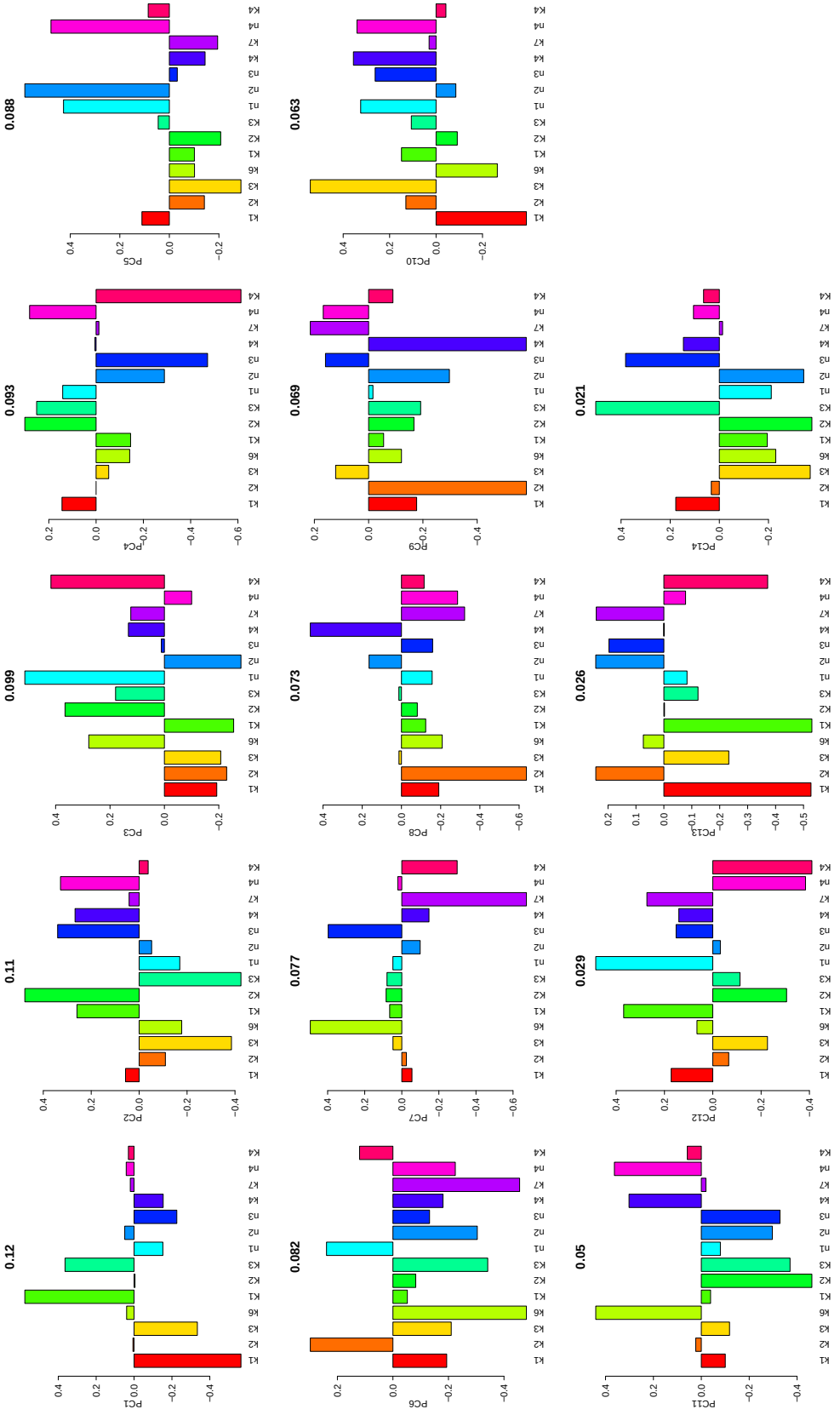

Figure S5: Robust oscillator design: principal component analysis of the posterior distribution for model 2 to achieve limits cycles via a hopf bifurcation.

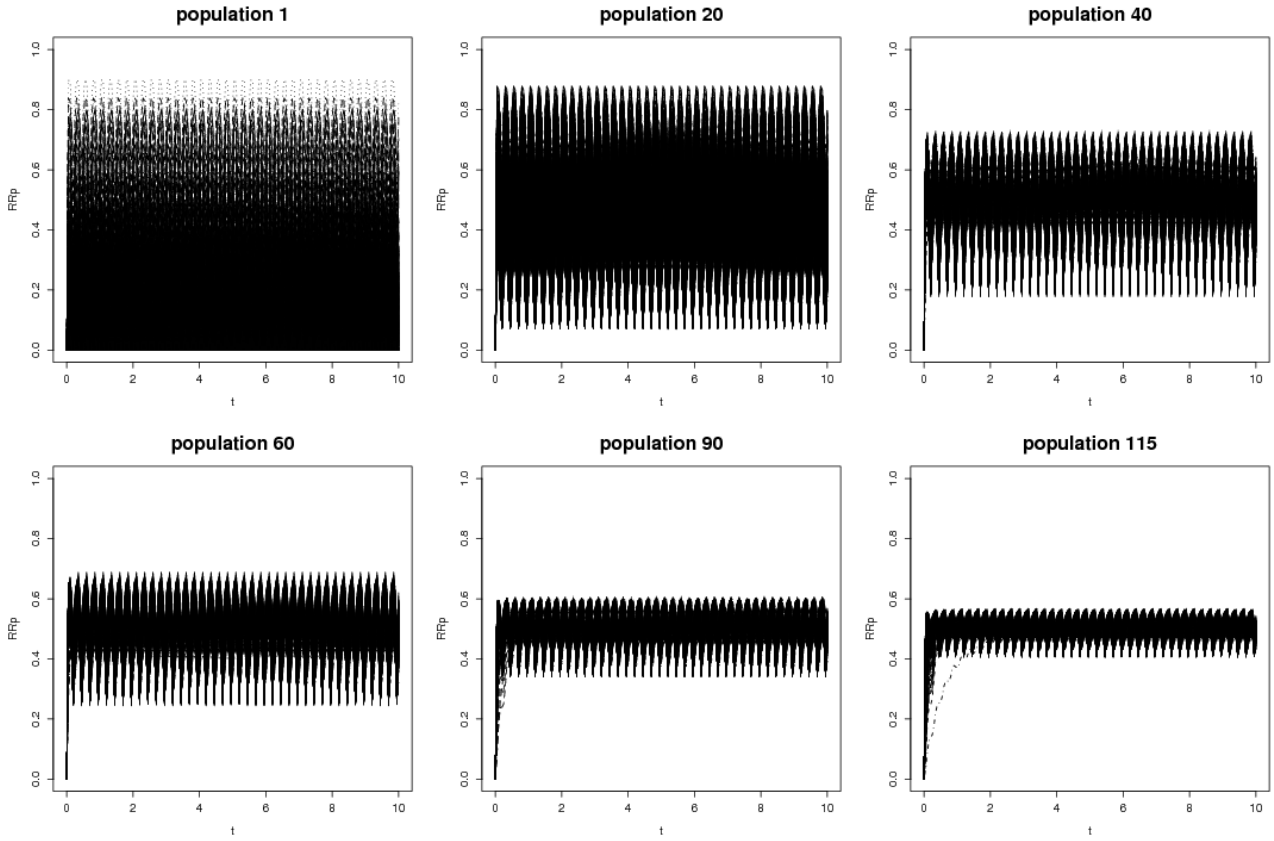

Figure S6: Two component systems: evolution to the noise reduction behavior.

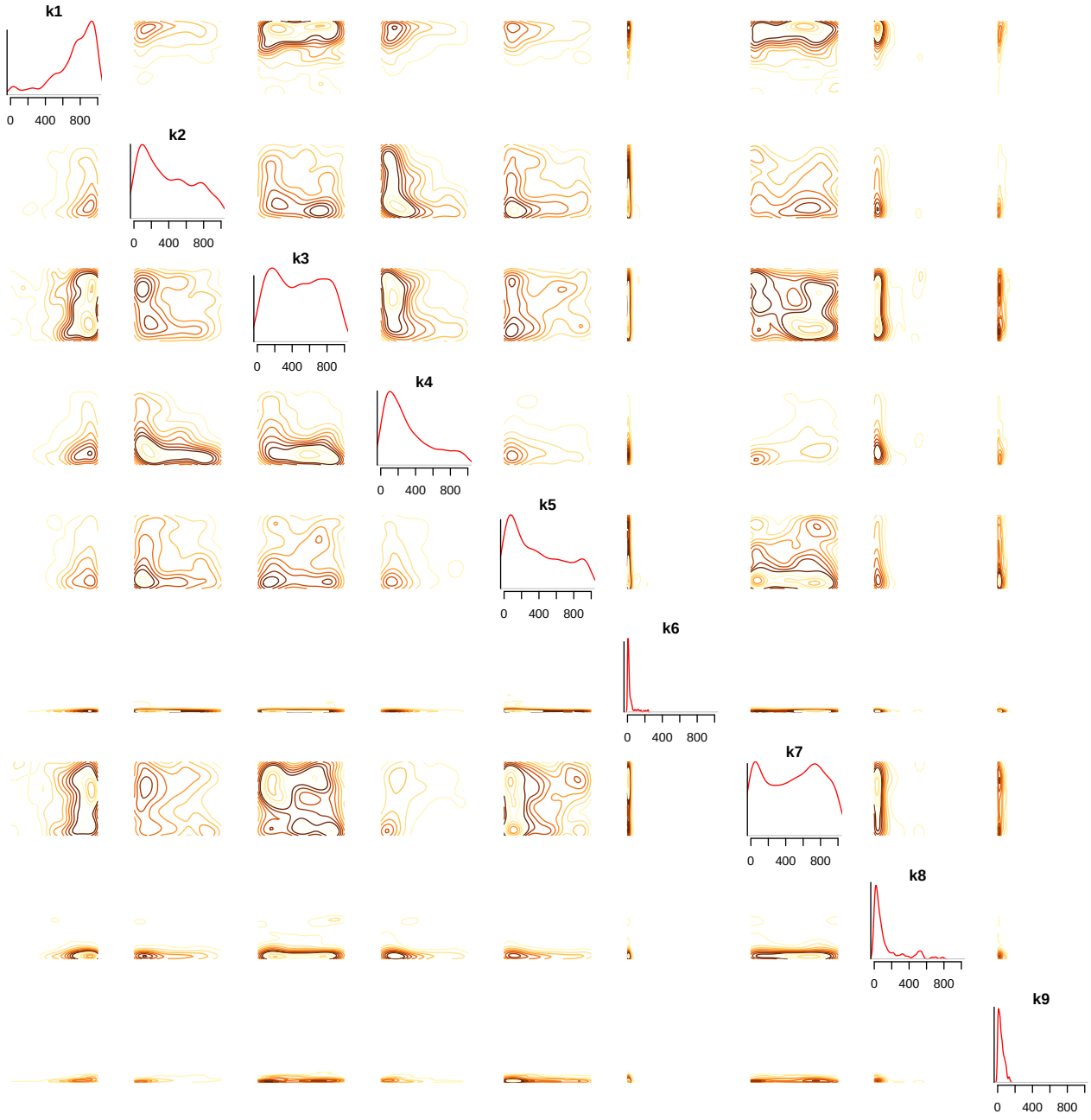

Figure S7: Two component systems: posterior distribution for the unorthodox system to achieve the noise reduction behavior.

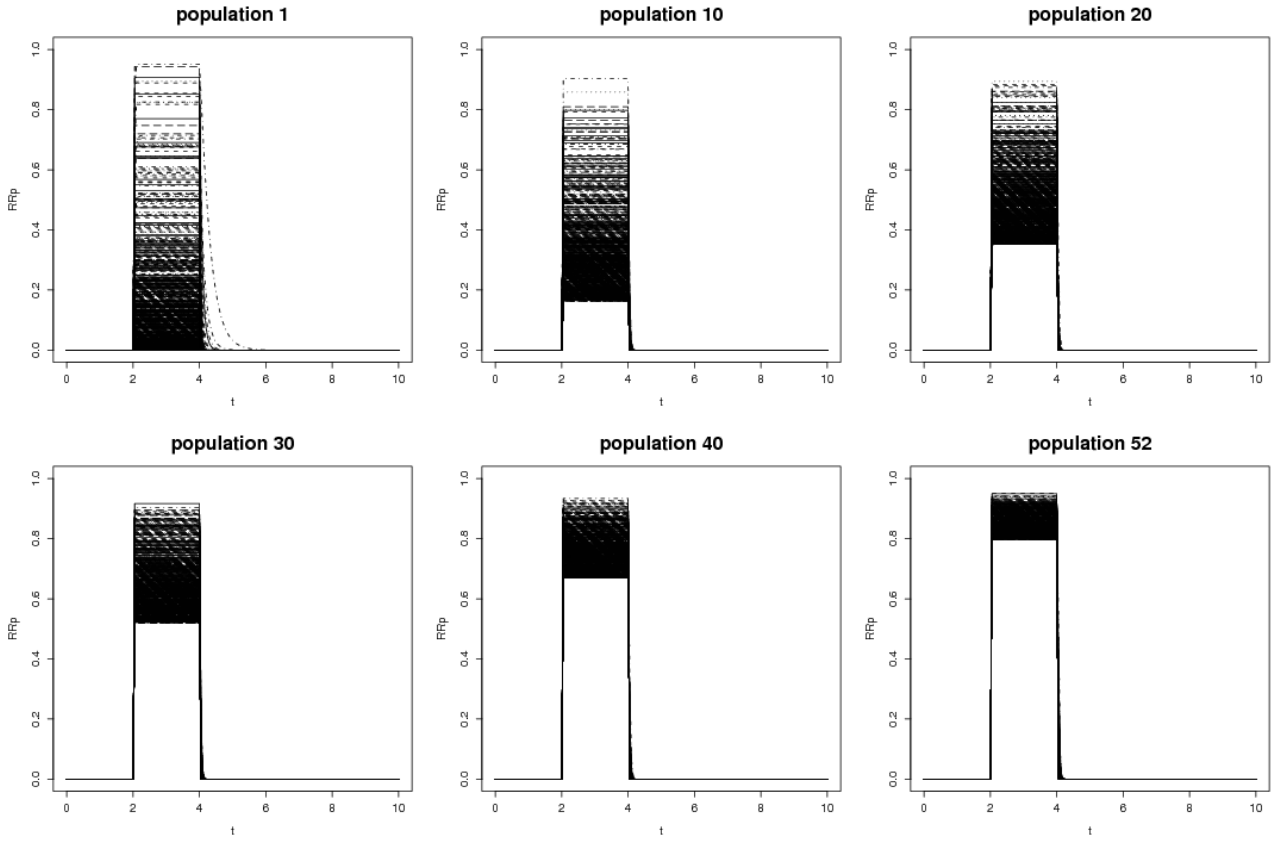

Figure S8: Two component systems: evolution to the signal reproduction behavior.

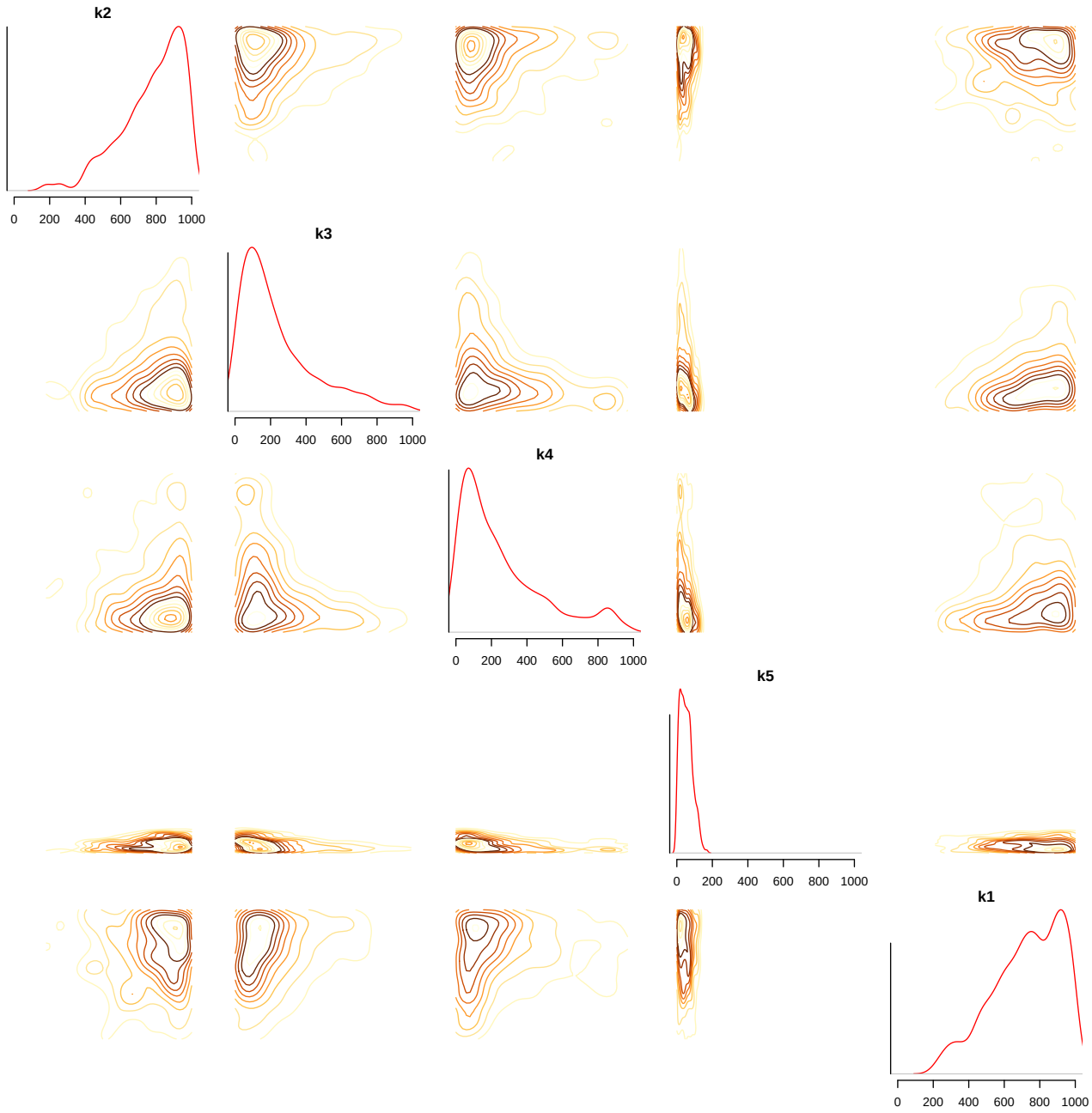

Figure S9: Two component systems: posterior distribution for the orthodox system to achieve the signal reproduction behavior.

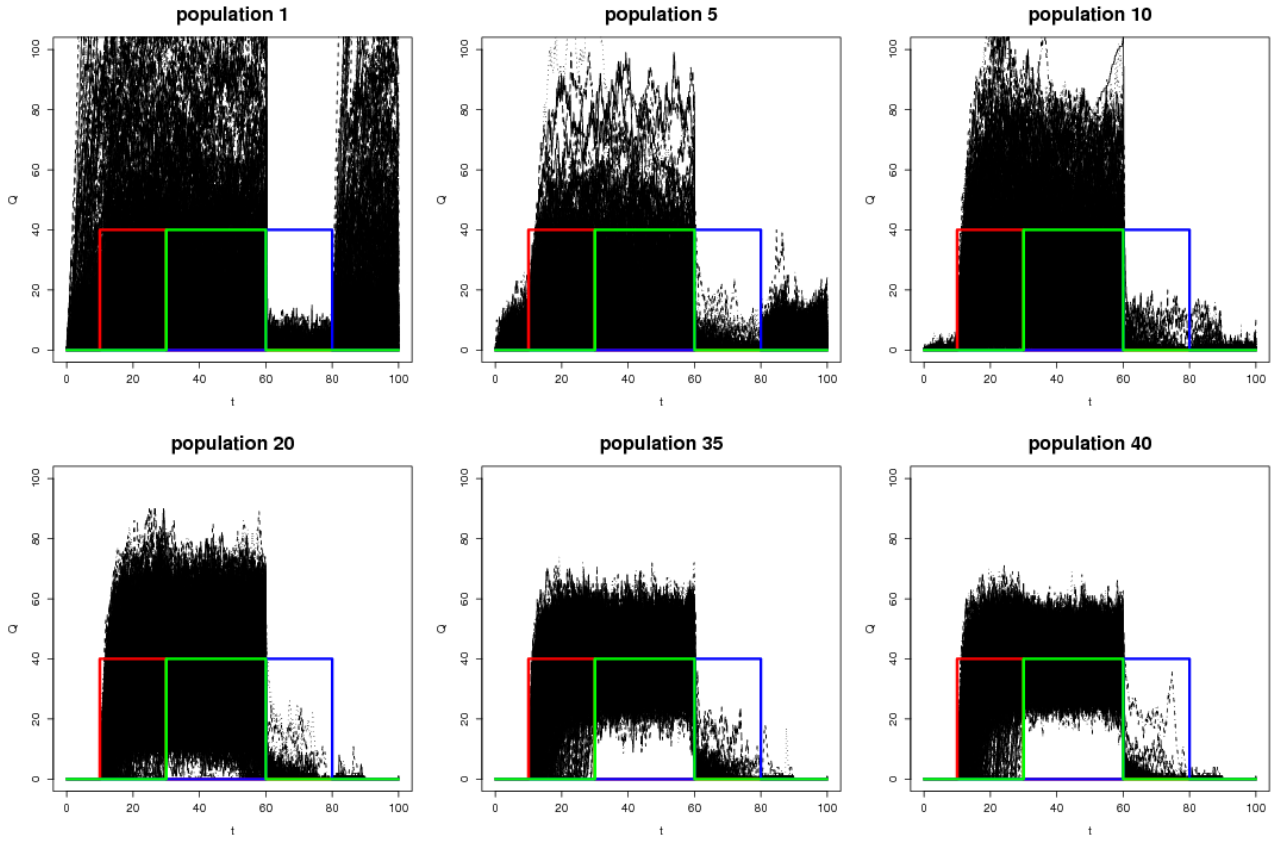

Figure S10: Stochastic genetic toggle switch: evolution to the desired behavior.

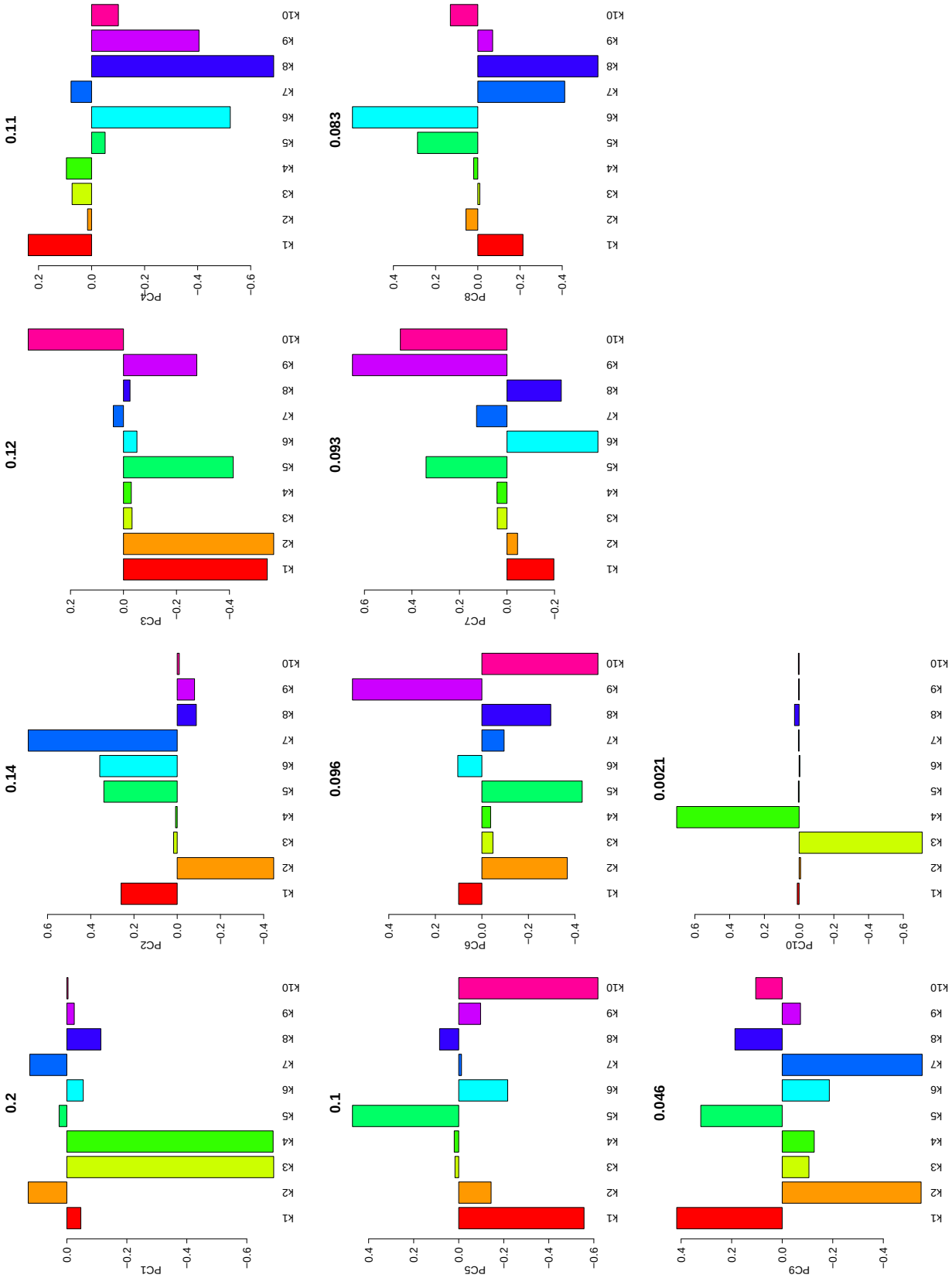

Figure S11: Stochastic genetic toggle switch: principal component analysis of the posterior distribution for model 2.