Thermodynamically consistent Bayesian analysis of closed biochemical reaction systems

Garrett Jenkinson,[1] Xiaogang Zhong,[2] and John Goutsias[1§]

[1]Whitaker Biomedical Engineering Institute, The Johns Hopkins University, Baltimore, MD 21218, USA

[2] Department of Applied Mathematics and Statistics, The Johns Hopkins University, Baltimore, MD 21218, USA

[§] Corresponding author

Email addresses:

GJ: jenkinson@jhu.edu

XZ: xzhong4@jhu.edu

JG: goutsias@jhu.edu

BMC Bioinformatics 2010, 11:547

Abstract

Background

Estimating the rate constants of a biochemical reaction system with known stoichiometry from noisy time series measurements of molecular concentrations is an important step for building predictive models of cellular function. Inference techniques currently available in the literature may produce rate constant values that defy necessary constraints imposed by the fundamental laws of thermodynamics. As a result, these techniques may lead to biochemical reaction systems whose concentration dynamics could not possibly occur in nature. Therefore, development of a thermodynamically consistent approach for estimating the rate constants of a biochemical reaction system is highly desirable.

Results

We introduce a Bayesian analysis approach for computing thermodynamically consistent estimates of the rate constants of a closed biochemical reaction system with known stoichiometry given experimental data. Our method employs an appropriately designed prior probability density function that effectively integrates fundamental biophysical and thermodynamic knowledge into the inference problem. Moreover, it takes into account experimental strategies for collecting informative observations of molecular concentrations through perturbations. The proposed method employs a maximization-expectation-maximization algorithm that provides thermodynamically feasible estimates of the rate constant values and computes appropriate measures of estimation accuracy. We demonstrate various aspects of the proposed method on synthetic data obtained by simulating a subset of a well-known model of the EGF/ERK signaling pathway, and examine its robustness under conditions that violate key assumptions. Software, coded in MATLAB®, which implements all Bayesian analysis techniques discussed in this paper, is available free of charge at http://www.cis.jhu.edu/∼goutsias/CSS%20lab/software.html.

Conclusions

Our approach provides an attractive statistical methodology for estimating thermodynamically feasible values for the rate constants of a biochemical reaction system from noisy time series observations of molecular concentrations obtained through perturbations. The proposed technique is theoretically sound and computationally feasible, but restricted to quantitative data obtained from closed biochemical reaction systems. This necessitates development of similar techniques for estimating the rate constants of open biochemical reaction systems, which are more realistic models of cellular function.

Background

Biochemical reaction systems are popular models of cellular function. These models are extensively used to represent the inter-connectivity and functional relationships among molecular species in cells and, most often, they provide accurate description of cellular behavior. Inferring a biochemical reaction system from experimental data is an important step towards building mathematical and computational tools for the analysis of cellular systems. This step requires both structure (stoichiometry) identification as well as parameter (rate constant) estimation [1–4]. Due however to the large combinatorial complexity of determining the stoichiometry of a biochemical reaction system, solving this problem requires large amounts of high quality experimental data and substantial computational resources, which are not usually available in practice.

Recently, several approaches have been proposed in the literature for addressing a simpler problem, known as *model calibration*. The objective of model calibration is to adjust the kinetic parameters of a biochemical reaction system with given stoichiometry in order to obtain a sufficiently good match between simulated and observed dynamics; e.g. see [2, 5–11].

Among known model calibration techniques, the ones based on Bayesian analysis [7, 10, 11] are perhaps the most versatile. Bayesian analysis allows us to effectively incorporate biophysical knowledge into the problem at hand and naturally draw statistical conclusions about the unknown kinetic parameters. This is done by employing a probability density function that encapsulates prior information about the rate constants of a biochemical reaction system and by deriving a posterior probability density function over the kinetic parameters after experimental data have been collected. By taking into account the experimental data and the information contained in the prior, the posterior density summarizes all knowledge available about the unknown kinetic parameters and quantifies uncertainty about their true values [12, 13]. Moreover, the posterior allows us to quantify our confidence about estimation accuracy, compute probabilities over alternative calibrations, and design additional experiments to improve inference.

Most published model calibration techniques do not take into account constraints on the reaction rate constants imposed by the fundamental laws of thermodynamics. If these constraints, known as Wegscheider conditions [14, 15], are not explicitly considered by a model calibration technique, then the method will spend most time examining impossible kinetic parameter sets and will most probably produce a biochemical reaction system that is not physically realistic [16]. This issue has been recently recognized in the literature, and new modeling formalisms have been suggested in an effort to address it [17–20]. The proposed formalisms describe a biochemical reaction system by well-defined thermodynamic parameters whose values always guarantee that the reaction rate constants satisfy the Wegscheider conditions. For example, in [19, 20], a biochemical reaction system is parameterized in terms of molecular capacities and reaction resistances, by using a thermodynamic kinetic modeling (TKM) formalism that enjoys a number of advantages over the ones suggested in [17, 18].

We believe that parameterizing a biochemical reaction system in terms of capacities and resistances is unnecessary and, in certain instances, problematic. It has been pointed out in [19] that different choices for the TKM parameters can lead to the same concentration dynamics. As a consequence, the TKM parameters cannot be determined uniquely from concentration measurements. A way to address this problem is to take the capacities to be the equilibrium concentrations (which is always possible in closed biochemical reaction systems), in which case the capacities are constrained by conservation relationships imposed by the system stoichiometry. Then, parameter estimation in the TKM formalism may be possible by arbitrarily fixing a subset of capacity values and estimating the remaining capacities and resistances. However, this approach can be very cumbersome when dealing with molecular perturbations (as we do in this paper) or when merging estimated TKM models, since, in both cases, the capacities may not refer to compatible equilibrium concentrations. It has been suggested in [19] that a way to merge two models using the TKM formalism is to first convert the capacities and resistance to the rate parameters, merge the two models, and then convert back to the TKM formalism. However, this approach seems to be overly complicated, especially in view of the model calibration methodology presented here.

In this paper, we introduce a thermodynamically consistent Bayesian analysis approach to model calibration that does not require reparametrization. Our approach relies on statistically modeling the reaction rate constants of the forward reactions as well as the equilibrium constants of individual reactions. We restrict our attention to closed systems (or systems that can be approximately considered to be closed), since thermodynamic analysis of such systems is easier to handle than open systems. The proposed approach controls thermodynamic consistency of the reaction rate constants by employing well-defined relationships between the kinetic parameters of a biochemical reaction system, imposed by the Wegscheider conditions. By embedding these relationships within an iterative algorithm that finds the mode of the posterior density, we arrive at a thermodynamically consistent Bayesian estimate for the rate constants.

Bayesian analysis can be appreciably influenced by the choice of the prior probability density functions. This is particularly true in systems biology problems in which only a small number of observations is usually available. It is therefore important to focus our effort on constructing appropriate prior densities for the unknown rate constants of the forward reactions and the equilibrium constants of individual reactions. Although a number of choices may be possible, it is imperative to use fundamental biophysical and thermodynamic principles to derive informative prior densities that effectively encapsulate such principles.

By using the classical Arrhenius formula of chemical kinetics [21], we construct an appropriate prior density for the log-rate constants of the forward reactions. To do so, we assume that the prefactor and activation energy associated with the Arrhenius formula are both random variables following log-normal and exponential distributions, respectively. This approach takes into account unpredictable changes in biochemical conditions affecting the structure of the reactant molecules and the probability of reaction after collision. On the other hand, by exploiting the thermodynamic relationship between rate constants, equilibrium concentrations, and stoichiometric coefficients, we derive an analytical expression for the joint prior density of the logarithms of the equilibrium constants. This expression depends on steady-state concentration measurements and on the stoichiometry of the biochemical reaction system under consideration.

Another important issue associated with the inference problem considered in this paper is the need to collect an informative set of measurements that can lead to sufficiently accurate parameter estimation. It has been increasingly recognized in the literature that a powerful approach to accomplish this goal in problems of systems biology is to selectively perturb key molecular components and measure the effects of these perturbations on the underlying concentrations [22–24]. We follow this strategy here and assume that we can selectively perturb, one at a time, the initial concentrations of a selected number of molecular species in a biochemical reaction system, by increasing or decreasing their values without altering the underlying stoichiometry. This can be achieved by a variety of experimental techniques, such as RNA interference (RNAi), transfection, or molecular injection. Therefore, our approach combines Bayesian analysis with current experimental practices, thus bridging the gap between statistical inference approaches and experimental design.

The Bayesian analysis technique discussed in this paper requires numerical evaluation of a number of statistical summaries of the posterior density. Although several methods are available to deal with this problem (e.g., see [25, 26]), we employ here a maximization-expectation-maximization (MEM) strategy that calculates a thermodynamically consistent estimate of the reaction rate constants as well as Monte Carlo estimates of posterior summaries used to evaluate the quality of inference. This strategy is based on sequentially combining a powerful stochastic optimization technique, known as simultaneous perturbation stochastic approximation (SPSA) [27], with Markov chain Monte Carlo (MCMC) sampling [25]. Our experience with extensive synthetic experiments, based on data obtained by simulating a subset of a well-known model of the EGF/ERK signaling pathway, indicates that the proposed algorithm is robust, producing excellent estimation results even in cases of high measurement errors and limited time measurements.

This paper is structured as follows. In the "Methods" section, we provide a brief overview of biochemical reaction systems, discuss how to model perturbations, and present a standard statistical model for the measurements. We then outline our Bayesian analysis approach to model calibration and present our choices for the prior and posterior densities. By emphasizing the fact that the prior density must assign zero probability over the thermodynamically infeasible region of the parameter space, and by employing an encompassing prior approach to Bayesian analysis, we derive an appropriate posterior density that satisfies this condition. We finally outline our proposed methodology for computing thermodynamically consistent Bayesian estimates of the kinetic parameters and for assessing estimation accuracy.

In the "Results/Discussion" section, we provide simulation results, based on a subset of a well-established model of the EGF/ERK signal transduction pathway. These results illustrate key aspects of the proposed model calibration methodology and show its potential for producing sufficiently accurate thermodynamically consistent estimates of a biochemical reaction system from noisy time-series measurements of molecular concentrations.

Finally, in the "Conclusions" section, we discuss a key statistical advantage of the proposed model calibration methodology, viewed from a bias-variance tradeoff perspective. Moreover, we provide suggestions for further research to address a number of practical issues associated with model calibration, such as estimating the initial concentrations and their perturbations, dealing with partially observed or missing data, and extending the proposed technique to the case of open biochemical reaction systems.

Extensive mathematical and computational details are required to rigorously formulate, derive, and understand various aspects of the proposed approach. We provide these details in three Additional files accompanying this paper. In Additional file 1, we present a detailed exposition of the underlying theory, whereas, in Additional file 2, we carefully discuss computational implementation. Finally, in Additional file 3, we provide all necessary details pertaining the biochemical reaction system we use in our simulations. Well-documented software, coded in MATLAB$^{®}$, which implements all Bayesian analysis techniques discussed in this paper, is available to interested readers free of charge at http://www.cis.jhu.edu/$\sim$goutsias/CSS%20lab/software.html.

Methods

Biochemical reaction systems

In this paper, we consider a biochemical reaction system comprised of N molecular species X_1, X_2, ..., X_N that interact through M coupled reactions of the form:

$$\sum_{n=1}^{N} \nu_{nm} X_n \underset{k_{2m}}{\overset{k_{2m-1}}{\rightleftarrows}} \sum_{n=1}^{N} \nu'_{nm} X_n, \quad m \in \mathcal{M} := \{1, 2, \dots, M\}. \tag{1}$$

The parameters k_{2m-1} and k_{2m} are the rate constants of the forward and reverse reactions, whereas, ν_{nm}, $\nu'_{nm} \geq 0$ are the stoichiometries of the reactants and products. Note that $k_{2m-1}, k_{2m} > 0$, for all $m \in \mathcal{M}$, since irreversible reactions are thermodynamically not possible in a closed biochemical reaction system [19]. We will assume that the system is well-mixed (homogeneous) with constant temperature and volume. We will also assume that the molecular concentrations evolve continuously as a function of time and that all reactions can be sufficiently characterized by the mass action rate law. In this case, we can describe the dynamic evolution of the molecular concentrations in the system by the following chemical kinetic equations:

$$\frac{dx_n^{(p)}(t)}{dt} = \sum_{m \in \mathcal{M}} s_{nm} \rho_m^{(p)}(t), \;\; t \in \mathcal{T}, \; n \in \mathcal{N}, \; p \in \mathcal{P}, \tag{2}$$

initialized by

$$x_n^{(p)}(0) = \begin{cases} c_p + \pi_p \,, & \text{if } \; n = p \neq 0 \\ c_n \,, & \text{if } \; p = 0 \;\text{ or }\; n \neq p \neq 0 \,, \end{cases} \tag{3}$$

where $\rho_m^{(p)}(t)$ is the net flux of the m^{th} reaction at time t, given by

$$\rho_m^{(p)}(t) = k_{2m-1} \prod_{i \in \mathcal{N}} [x_i^{(p)}(t)]^{\nu_{im}} - k_{2m} \prod_{i \in \mathcal{N}} [x_i^{(p)}(t)]^{\nu'_{im}}, \tag{4}$$

s_{nm} is the *net* stoichiometry coefficient of the n^{th} molecular species associated with the m^{th} reaction, defined by $s_{nm} := \nu'_{nm} - \nu_{nm}$, and $\mathcal{T} := [0, t_{\text{max}}]$ is an observation time window of interest.

Equations (2)–(4) are based on the assumption that we can selectively perturb, one at a time, the concentrations of molecular species in a set $\mathcal{P}$, by increasing or decreasing their values at time $t = 0$ without altering the underlying stoichiometry. For notational convenience, we include 0 in $\mathcal{P}$ and assign $p = 0$ to the original unperturbed system. In this case, $x_n^{(0)}(t)$ is the concentration of the n^{th} molecular species in the unperturbed system at time t, whereas, $x_n^{(p)}(t)$, for $p \neq 0$, is the concentration of the n^{th} molecular species at time t, obtained by perturbing the initial concentration of the p^{th} species. In (3), $\pi_p \geq -c_p$ quantifies the perturbation applied on the initial concentration c_p of the p^{th} molecular species at time $t = 0$. When $-c_p \leq \pi_p < 0$, the initial concentration of the p^{th} molecular species is reduced, a situation that can be achieved by a variety of experimental techniques, such as RNA interference (RNAi). On the other hand, when $\pi_p > 0$, the initial concentration of the p^{th} molecular species is increased, a situation that can be achieved by transfection or molecular injection.

Due to the enormous complexity of biological reaction networks, (1) is used to model a limited number of molecular interactions embedded within a larger and more complex system. Mass flow between the biochemical reaction system given by (1) and its surroundings complicates modeling. As a matter of fact, some molecular concentrations in the system may be influenced by unknown reactions, not modeled by (1), or by partially known reactions with reactants regulated by unknown biochemical mechanisms. To address this problem, we will assume that there is no appreciable mass transfer between the biochemical reaction system and its surroundings during the observation time interval $\mathcal{T} = [0, t_{\text{max}}]$. As a consequence, we can assume that (1) characterizes a *closed* biochemical reaction system within $\mathcal{T}$. Moreover, we will assume that the system reaches *quasi-equilibrium* at some time $t_* \leq t_{\text{max}}$, after which its thermodynamic properties do not appreciably change for $t_* < t \leq t_{\text{max}}$. Note however that the quasi-equilibrium assumption does not necessarily imply that the biochemical reaction system will be at thermodynamic equilibrium after time t_{max}, since mass transfer may take place at some time $t > t_{\text{max}}$. Although we may be able to satisfy these assumptions by appropriately designed *synthetic* or *in vitro* biological experiments, the assumptions are certainly not satisfied *in vivo*. For this reason, we believe that future research must be focused on extending the approaches and techniques discussed in this paper to the case of *open* biochemical reaction systems.

Measurements

We will now specify an appropriate model for the available measurements. We will assume that, by an appropriately designed experiment, we can obtain noisy measurements $\boldsymbol{y} := \{y_n^{(p)}(t_q), n \in \mathcal{N}, p \in \mathcal{P}, q \in \mathcal{Q}\}$ and $\overline{\boldsymbol{y}} := \{y_n^{(p)}(t_{Q+1}), n \in \mathcal{N}, p \in \mathcal{P}\}$ of the concentrations of all molecular species in the unperturbed and perturbed systems at a limited number of distinct time points $t_1 < t_2 < \cdots < t_Q < t_{Q+1}$ in $\mathcal{T}$, where $\mathcal{Q} := \{1, 2, \ldots, Q\}$. We will also assume that these measurements are related to the true concentrations $x_n^{(p)}(t_q)$ by

$$y_n^{(p)}(t_q) = \ln\left[\epsilon_n^{(p)}(t_q)\, x_n^{(p)}(t_q)\right] = \ln x_n^{(p)}(t_q) + \eta_n^{(p)}(t_q), \quad n \in \mathcal{N}, \;\; p \in \mathcal{P}, \tag{5}$$

for $q = 1, 2, \ldots, Q+1$, where $\epsilon_n^{(p)}(t_q)$ is a *multiplicative* random error factor and $\eta_n^{(p)}(t_q) := \ln \epsilon_n^{(p)}(t_q)$. The assumption of multiplicative errors is common in most data acquisition procedures, such as DNA microarray-based genomics and mass spectrometry-based proteomics [28–30], whereas, the logarithm is used to obtain a convenient additive error model for the measurements.

In the following, we will assume that the biochemical reaction system, and all its perturbed versions, is sufficiently close to steady-state at time point t_{Q+1}. We can justify this assumption by taking $t_* \leq t_{Q+1} \leq t_{\text{max}}$ and by recalling our previous assumption that the biochemical reaction system is at thermodynamic quasi-equilibrium at times $t_* \leq t \leq t_{\text{max}}$. Our Bayesian analysis approach is based on data $\boldsymbol{y}$, whereas, we use the steady-state measurements $\overline{\boldsymbol{y}}$ to derive a joint probability density function for the logarithms $\{\ln(k_{2m-1}/k_{2m}), m \in \mathcal{M}\}$ of the equilibrium constants of the reactions needed for specifying the posterior density.

Finally, we will assume that the error components $\eta_n^{(p)}(t_q)$ are statistically independent zero-mean Gaussian random variables. The Gaussian assumption is quite common in genomic problems and has been experimentally verified in some cases; e.g., see [31]. This assumption is usually justified by the central limit theorem and the premise that the errors are due to a large number of independent multiplicative error sources. We may attempt to justify the independence assumption between measurement errors by arguing that an error occurred in a particular measurement may only be due to the acquisition process used to obtain that measurement and, hence, it may not affect the error values of other measurements. In general, however, this is only a mathematically convenient assumption that may not be realistic. We experimentally demonstrate later that, at least for the example considered in this paper, the proposed estimation methodology is quite effective even in the case of non-Gaussian and correlated measurement errors. For simplicity, we finally assume equal error variances; i.e., we will assume that var$[\eta_n^{(p)}(t_q)] = \sigma^2$, for every n, p, and q. This assumption is not crucial to our approach and can be relaxed if necessary.

Bayesian model calibration

In this paper, we deal with the following problem: Given noisy concentration measurements $\boldsymbol{y}$ and $\overline{\boldsymbol{y}}$, we want to calculate thermodynamically consistent estimates of the log-rate constants $\boldsymbol{\kappa} := \{\kappa_{2m-1} := \ln k_{2m-1}, \kappa_{2m} := \ln k_{2m}, m \in \mathcal{M}\}$ of a closed biochemical reaction system, such that (2), initialized by (3), produce molecular concentrations $x_n^{(p)}(t)$ that "best" match (in some well-defined sense) the available measurements.

We should note here that it is convenient to estimate the logarithms of the rate constants instead of the constants themselves. By focusing on the logarithms, we can reduce the dynamic range of rate constant values and make their estimation numerically easier. To simplify our developments, we will assume that the initial concentrations $\{c_n, n \in \mathcal{N}\}$ and perturbations $\{\pi_p, p \in \mathcal{P}\}$ are known or have been estimated by an appropriate experimental procedure. When this is not true, these quantities must be treated as unknown parameters and estimated from data, together with the rate constants, provided that a sufficient amount of data is available to allow reliable estimation.

Given data $\boldsymbol{y}$, the objective of Bayesian analysis is to evaluate the *posterior* probability density function $p(\boldsymbol{\kappa} \mid \boldsymbol{y})$, which summarizes our belief about the log-rate constants $\boldsymbol{\kappa}$ *after* the data $\boldsymbol{y}$ have been collected. It can be shown [see Equations (S-1.4) and (S-1.5) in Additional file 1] that

$$p(\boldsymbol{\kappa} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{\kappa}) \int p(\boldsymbol{\kappa} \mid \boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z}, \tag{6}$$

where $p \propto q$ denotes that p is proportional to q, and

$$p(\boldsymbol{y} \mid \boldsymbol{\kappa}) = \int p(\boldsymbol{y} \mid \boldsymbol{\kappa}, \sigma^2) p(\sigma^2 \mid \boldsymbol{\kappa}) d\sigma^2, \tag{7}$$

with $\boldsymbol{z} = \{z_m, m \in \mathcal{M}\}$ being the set of log-equilibrium constants of the reactions, defined by

$$z_m := \ln \frac{k_{2m-1}}{k_{2m}} = \kappa_{2m-1} - \kappa_{2m}, \quad \text{for } m \in \mathcal{M}. \tag{8}$$

Note that the prior density of the log-rate constants $\boldsymbol{\kappa}$ depends on $\boldsymbol{z}$. For this reason, we view $\boldsymbol{z}$ as a set of random hyperparameters (in Bayesian analysis, parameters used to specify prior densities are known as hyperparameters), specified by means of the prior density $p(\boldsymbol{z})$.

The posterior density $p(\boldsymbol{\kappa} \mid \boldsymbol{y})$ takes into account our prior belief about the rate constant values and the data formation process, summarized by the prior density $p(\boldsymbol{z})$ of the log-equilibrium constants, the conditional prior density $p(\boldsymbol{\kappa} \mid \boldsymbol{z})$ of the log-rate constants given the log-equilibrium constants, the conditional probability density $p(\sigma^2 \mid \boldsymbol{\kappa})$ of the error variance given the log-rate constants, and the likelihood $p(\boldsymbol{y} \mid \boldsymbol{\kappa}, \sigma^2)$. However, the posterior density is hard to interpret, especially in high-dimensional problems that involve many parameters, such as the problem we are dealing with here. As a consequence, the main objective of Bayesian analysis is to produce numerical information that can be effectively used to summarize the posterior density and simplify the task of statistical inference to the extent possible. Typical summaries include measures of location and scale of the posterior, which are used to produce estimates for the parameter values and to evaluate the accuracy of such estimates, respectively.

It is clear from (6) that, to evaluate the posterior $p(\boldsymbol{\kappa} \mid \boldsymbol{y})$, we need to compute the "effective" prior density $\int p(\boldsymbol{\kappa} \mid \boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z}$ as well as the "effective" likelihood $\int p(\boldsymbol{y} \mid \boldsymbol{\kappa}, \sigma^2) p(\sigma^2 \mid \boldsymbol{\kappa}) d\sigma^2$. To do so, we must specify the aforementioned densities $p(\sigma^2 \mid \boldsymbol{\kappa})$, $p(\boldsymbol{z})$, $p(\boldsymbol{\kappa} \mid \boldsymbol{z})$, and $p(\boldsymbol{y} \mid \boldsymbol{\kappa}, \sigma^2)$. We discuss this problem next.

Prior density of error variance

In general, it is difficult to derive an informative prior probability density function $p(\sigma^2 \mid \boldsymbol{\kappa})$ for the error variance. To deal with this problem, we assume here that the error variance is independent of the rate constants; i.e., we assume that $p(\sigma^2 \mid \boldsymbol{\kappa}) = p(\sigma^2)$. Moreover, we assume that σ^2 follows an inverse gamma distribution, in which case

$$p(\sigma^2) = \frac{b^a}{\Gamma(a)} \, (\sigma^2)^{-(a+1)} \, e^{-b/\sigma^2}, \tag{9}$$

for two parameters $a, b > 0$.

The independence assumption between σ^2 and $\boldsymbol{\kappa}$ is reasonable, in view of the fact that the errors are mainly due to the experimental methodology used to obtain the measurements, whereas, the rate constants are due to biophysical principles underlying the biochemical reaction system. On the other hand, the choice given by (9) has been well-justified in Bayesian analysis. In fact, the inverse gamma distribution is the conjugate prior for the variance of additive Gaussian errors [13]. Conjugate priors are common in Bayesian analysis, since they often lead to attractive analytical and computational simplifications. Note that $\mathrm{E}[\sigma^2] = b/(a-1)$ and $\mathrm{var}[\sigma^2] = \{\mathrm{E}[\sigma^2]\}^2/(a-2) = b^2/[(a-1)^2(a-2)]$, for $a > 2$. Therefore, the parameters a, b control the location and scale of the inverse gamma distribution given by (9). We illustrate this prior in Figure S-1.3 of Additional file 1. In the following, we treat a and b as hyperparameters with known values. For a practical method to determine these values, the reader is referred to Additional file 1.

Prior density of log-equilibrium constants

Before we consider the problem of specifying a prior density for the log-equilibrium constants $\boldsymbol{z}$, we first investigate how much information about $\boldsymbol{z}$ can be extracted from measurements.

It is a direct consequence of thermodynamic analysis that, at steady-state, the net flux of each reaction in a closed biochemical reaction system must be zero. This implies that

$$k_{2m-1} \prod_{n \in \mathcal{N}} [\overline{x}_n^{(p)}]^{\,\nu_{nm}} = k_{2m} \prod_{n \in \mathcal{N}} [\overline{x}_n^{(p)}]^{\,\nu'_{nm}}, \quad \text{for all } \ m \in \mathcal{M}, \ p \in \mathcal{P}, \tag{10}$$

by virtue of (4), where $\{\overline{x}_n^{(p)} > 0, n \in \mathcal{N}\}$ are the stationary concentrations when the initial concentration of the p^{th} molecular species is perturbed (thermodynamic analysis dictates that these concentrations must be nonzero, provided that the initial concentrations are nonzero). As a matter of fact, (10) is equivalent to the following constraints on the reaction rate constants (see Additional file 1):

$$\prod_{m \in \mathcal{M}} \left(\frac{k_{2m-1}}{k_{2m}} \right)^{r_m} = 1, \qquad \text{for all } \ \boldsymbol{r} \in \text{null}(\mathbb{S}), \tag{11}$$

known as Wegscheider conditions [14, 15], where r_m is the m^{th} element of the $M \times 1$ vector $\boldsymbol{r}$, $\mathbb{S}$ is the $N \times M$ stoichiometry matrix of the biochemical reaction system with elements s_{nm}, and null($\mathbb{S}$) is the null space of $\mathbb{S}$. As a consequence, for a biochemical reaction system to be physically realizable, it is required that the reaction rates satisfy the thermodynamically imposed Wegscheider conditions.

From (8) and (10), note that

$$z_m = \frac{1}{P+1} \sum_{p \in \mathcal{P}} \sum_{n \in \mathcal{N}} s_{nm} \ln \overline{x}_n^{(p)}, \quad \text{for all } \ m \in \mathcal{M}. \tag{12}$$

By employing (5) and (12), we can show that $z_m = \widetilde{y}_m - \widetilde{\eta}_m$, where

$$\widetilde{y}_m := \frac{1}{P+1} \sum_{p \in \mathcal{P}} \sum_{n \in \mathcal{N}} s_{nm} y_n^{(p)}(t_{Q+1}) \quad \text{ and } \quad \widetilde{\eta}_m := \frac{1}{P+1} \sum_{p \in \mathcal{P}} \sum_{n \in \mathcal{N}} s_{nm} \eta_n^{(p)}(t_{Q+1}). \tag{13}$$

Using this result and some straightforward algebra (see Additional file 1), we can show that, given $\widetilde{\boldsymbol{y}} := \{\widetilde{y}_m, m \in \mathcal{M}\}$, which can be calculated from the measurements $\overline{\boldsymbol{y}} = \{y_n^{(p)}(t_{Q+1}), n \in \mathcal{N}, p \in \mathcal{P}\}$ of the steady-state molecular concentrations and (13), we can construct the posterior density $p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}})$ of $\boldsymbol{z}$ by

$$p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}}) \propto \left[\frac{2b}{P+1} + (\boldsymbol{z} - \widetilde{\boldsymbol{y}})^T \mathbb{H}^{-1} (\boldsymbol{z} - \widetilde{\boldsymbol{y}}) \right]^{-(M/2+a)}, \tag{14}$$

where $\mathbb{H}$ is an $M \times M$ matrix with elements $h_{mm'} = \sum_{n \in \mathcal{N}} s_{nm} s_{nm'}$, and a, b are the two hyperparameters associated with the prior density of the measurement variance, given by (9).

The previous result suggests that we may be able to use $p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}})$ as an informative prior for the log-equilibrium constants $\boldsymbol{z}$; i.e., we may be able to replace $p(\boldsymbol{z})$ by $p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}})$ in (6). At a first glance, this idea may not seem appropriate. However, it perfectly agrees with the fact that, in Bayesian analysis, hyperparameters are often estimated directly from data [13]. Since we have shown here that steady-state measurements can be effectively used to construct the entire posterior probability density function of $\boldsymbol{z}$, it seems reasonable to use this posterior as a prior density for $\boldsymbol{z}$. Note however that, by replacing $p(\boldsymbol{z})$ with $p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}})$ in (6), we must make sure that $\widetilde{\boldsymbol{y}}$ is independent of $\boldsymbol{y}$ (see Additional file 1). Otherwise, our choice for $p(\boldsymbol{z})$ may not lead to a proper posterior density $p(\boldsymbol{\kappa} \mid \boldsymbol{y})$ (i.e., it may not lead to a density that is finite for all $\boldsymbol{y}$). Note that the independence of $\widetilde{\boldsymbol{y}}$ and $\boldsymbol{y}$ is assured by the independence between the measurement errors $\{\eta_n^{(p)}(t_{Q+1}), n \in \mathcal{N}, p \in \mathcal{P}\}$ and $\{\eta_n^{(p)}(t_q), n \in \mathcal{N}, p \in \mathcal{P}, q \in \mathcal{Q}\}$.

An important observation here is that evaluation of $p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}})$, given by (14), may not be possible, since the matrix $\mathbb{H}$ may not be invertible. We can address this problem by decorrelating $\boldsymbol{z}$ using the singular value decomposition (SVD) of matrix $\mathbb{H}$. As a consequence, we obtain $\mathbb{H} = \mathbb{U}_0 \mathbb{D}_0 \mathbb{U}_0^T$, where $\mathbb{D}_0$ is an invertible diagonal matrix containing the nonzero singular values of $\mathbb{H}$, and $\mathbb{U}_0$ is an appropriately constructed matrix (see Additional file 1 for details). In this case, instead of using (14) for $p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}})$, we must use

$$p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}}) \propto \left[\frac{2b}{P+1} + (\boldsymbol{z} - \widetilde{\boldsymbol{y}})^T \mathbb{U}_0 \mathbb{D}_0^{-1} \mathbb{U}_0^T (\boldsymbol{z} - \widetilde{\boldsymbol{y}}) \right]^{-a}, \tag{15}$$

which we can always evaluate, since matrix $\mathbb{D}_0$ is invertible.

Prior density of log-rate constants

To specify the (conditional) prior density $p(\boldsymbol{\kappa} \mid \boldsymbol{z})$ of the log-rate constants of a biochemical reaction system, we will first derive a prior probability density function $p(\kappa_{2m-1})$ for the log-rate constant of the m^{th} forward reaction. To do so, we use the well-known Arrhenius formula of chemical kinetics [21] and set $k_{2m-1} = \alpha_m \exp\{-E_m / k_B T\}$, where α_m is the prefactor, E_m is the activation energy of the reaction, k_B is the Boltzmann constant ($k_B = 1.3806504 \times 10^{-23}$J/K), and T is the temperature. Unfortunately, we cannot predict the values of the prefactor and activation energy precisely. To deal with this problem, we set $\alpha_m = \alpha_m^0 \exp\{g_m\}$ and $E_m = E_m^0 + U_m$, where α_m^0, E_m^0 are the predictable portions of the prefactor and activation energy, respectively, and g_m, U_m are two random variables that model the unpredictable portions of these quantities. In the Additional file 1, we argue that it is reasonable to model g_m as a zero-mean Gaussian random variable with standard deviation λ_m, and U_m is an exponential random variable with mean and standard deviation $k_B T_m^*$, where T_m^* is a temperature larger than T. As a consequence, we obtain the following prior density for the log-rate constant κ_{2m-1} of the m^{th} forward reaction [see Equation (S-1.31) in Additional file 1]:

$$p(\kappa_{2m-1}) = \frac{e^{\lambda_m^2/2\tau_m^2}}{2\tau_m} \operatorname{erfc}\left[\frac{1}{\sqrt{2}}\left(\frac{\lambda_m}{\tau_m} + \frac{\kappa_{2m-1} - \kappa_m^0}{\lambda_m}\right)\right] e^{(\kappa_{2m-1}-\kappa_m^0)/\tau_m}, \tag{16}$$

where $\tau_m := T_m^*/T > 1$, $\kappa_m^0 := \ln \alpha_m^0 - E_m^0/k_B T$, and erfc$[\cdot]$ is the complementary error function. We illustrate this prior in Figure S-1.1 of Additional file 1.

Basic thermodynamic arguments (see Additional file 1) imply that z_m, defined by (8), is a constant characteristic to the m^{th} reaction. Since $\kappa_{2m} = \kappa_{2m-1} - z_m$, this implies that the rate constants κ_{2m} and κ_{2m-1} are two dependent random variables, given z_m, with joint probability density $p(\kappa_{2m}, \kappa_{2m-1} \mid z_m) = \delta(\kappa_{2m} - \kappa_{2m-1} + z_m)p(\kappa_{2m-1})$, where $\delta(\cdot)$ is the Dirac delta function [see Equation (S-1.37) in Additional file 1]. By assuming that the reaction rate constants of different reactions are mutually independent given the z's (which is reasonable if we assume that all common factors affecting these rates, such as temperature and pressure, are kept fixed), we obtain

$$p(\boldsymbol{\kappa} \mid \boldsymbol{z}) = \prod_{m \in \mathcal{M}} \delta(\kappa_{2m} - \kappa_{2m-1} + z_m)p(\kappa_{2m-1}). \tag{17}$$

Equations (16) and (17) provide an analytical form for the prior density of the log-rate constants. To use this expression, we must determine appropriate values for $\boldsymbol{\phi} := \{\kappa_m^0, \tau_m, \lambda_m, m \in \mathcal{M}\}$, which can be treated as hyperparameters. Although we could treat $\boldsymbol{\phi}$ as random, we will choose here known values for these parameters. This is motivated by the fact that $\boldsymbol{\phi}$ determines the location and scale of the prior densities of the forward rate constants; see Figure S-1.1 in Additional file 1. In certain problems of interest, there might be enough information to determine possible ranges for the forward rate constant values. As a consequence, we can use this information, together with an appropriate procedure, to effectively determine values for $\boldsymbol{\phi}$. The reader is referred to Additional file 1 for details on how to do so.

Effective likelihood

Calculating the effective likelihood $p(\boldsymbol{y} \mid \boldsymbol{\kappa})$, given by (7), is straightforward. From (5), (7), and (9), we can show that

$$p(\boldsymbol{y} \mid \boldsymbol{\kappa}) \propto \int \frac{1}{\sigma^{N(P+1)Q+2(a+1)}} \exp\left\{ -\frac{\varphi(\boldsymbol{\kappa}, \boldsymbol{y})}{2\sigma^2} \right\} d\sigma^2, \tag{18}$$

where

$$\varphi(\boldsymbol{\kappa}, \boldsymbol{y}) := 2b + \sum_{n \in \mathcal{N}} \sum_{p \in \mathcal{P}} \sum_{q \in \mathcal{Q}} [y_n^{(p)}(t_q) - \ln x_n^{(p)}(t_q)]^2. \tag{19}$$

By setting $\xi = \varphi(\boldsymbol{\kappa}, \boldsymbol{y})/2\sigma^2$ in (18), we obtain

$$\begin{aligned} p(\boldsymbol{y} \mid \boldsymbol{\kappa}) &\propto [\varphi(\boldsymbol{\kappa}, \boldsymbol{y})]^{-a-N(P+1)Q/2} \int \xi^{a+N(P+1)Q/2-1} e^{-\xi} d\xi \\ &\propto [\varphi(\boldsymbol{\kappa}, \boldsymbol{y})]^{-a-N(P+1)Q/2}. \end{aligned} \tag{20}$$

Note that evaluating $\varphi(\boldsymbol{\kappa}, \boldsymbol{y})$ at given values of $\boldsymbol{\kappa}$ and $\boldsymbol{y}$ requires integration of the system of ordinary differential equations (2).

Posterior density

Our previous developments lead finally to an analytical formula for the posterior density $p(\boldsymbol{\kappa} \mid \boldsymbol{y})$ of the log-rate constants. Indeed, (6), (15), (17), (19), and (20), lead to

$$p(\boldsymbol{\kappa} \mid \boldsymbol{y}) \propto \frac{\omega(\boldsymbol{\kappa})}{[\psi(\boldsymbol{\kappa}, \boldsymbol{y})]^a [\varphi(\boldsymbol{\kappa}, \boldsymbol{y})]^\beta}\,, \tag{21}$$

with

$$\begin{aligned} \omega(\boldsymbol{\kappa}) &= \prod_{m \in \mathcal{M}} \operatorname{erfc}\left[\frac{1}{\sqrt{2}} \left(\frac{\lambda_m}{\tau_m} + \frac{\kappa_{2m-1} - \kappa_m^0}{\lambda_m} \right) \right] e^{\kappa_{2m-1}/\tau_m} \\ \psi(\boldsymbol{\kappa}, \boldsymbol{y}) &= \frac{2b}{P+1} + \sum_{m \in \mathcal{M}} \sum_{m' \in \mathcal{M}} \theta_{mm'} (\kappa_{2m-1} - \kappa_{2m} - \widetilde{y}_m)(\kappa_{2m'-1} - \kappa_{2m'} - \widetilde{y}_{m'}) \\ \varphi(\boldsymbol{\kappa}, \boldsymbol{y}) &= 2b + \sum_{n \in \mathcal{N}} \sum_{p \in \mathcal{P}} \sum_{q \in \mathcal{Q}} [y_n^{(p)}(t_q) - \ln x_n^{(p)}(t_q)]^2 \\ \beta &= a + N(P+1)Q/2\,, \end{aligned} \tag{22}$$

where $\theta_{mm'}$ are the elements of matrix $\mathbb{U}_0 \mathbb{D}_0^{-1} \mathbb{U}_0^T$ obtained from the SVD decomposition of $\mathbb{S}^T \mathbb{S}$, and $\widetilde{y}_m$ is given by (13).

Note that the posterior density of the log-rate constants is a compromise between the prior and the likelihood. The prior terms $\omega(\boldsymbol{\kappa})$ and $\psi(\boldsymbol{\kappa}, \boldsymbol{y})$ penalize log-rate values that do not fit well with available a-priori information, whereas, the likelihood term $\varphi(\boldsymbol{\kappa}, \boldsymbol{y})$ penalizes log-rate values that produce concentration dynamics which deviate appreciably from measurements. As the number $N(P+1)Q$ of available measurements increases, this compromise is controlled to a greater extent by the data through the factor $\varphi(\boldsymbol{\kappa}, \boldsymbol{y})$.

A problem arises with the posterior density $p(\boldsymbol{\kappa} \mid \boldsymbol{y})$, given by (21) and (22), since nonzero probabilities may be assigned to thermodynamically infeasible log-rate constants. A Bayesian analyst might argue that we have correctly done our job by formulating the problem as we did and that it is the data which will rule out the possibility that our biochemical reaction system can be characterized by thermodynamically infeasible parameters. However, we choose to trust thermodynamics far more than we would trust noisy data and appropriately modify the posterior density based on our knowledge that the kinetic parameters must satisfy the Wegscheider conditions given by (11).

By using the Wegscheider conditions, we can decompose the $2M$ log-rate constants $\boldsymbol{\kappa}$ into two mutually exclusive sets: $M + M_1$ "free" log-rate constants $\boldsymbol{\kappa}_f$ and $M - M_1$ "dependent" log-rate constants $\boldsymbol{\kappa}_d$, where $M_1 = \text{rank}(\mathbb{S})$ (see Additional file 1). Although parameters $\boldsymbol{\kappa}_f$ can take any value, parameters $\boldsymbol{\kappa}_d$ must be equal to $\mathbb{W}\boldsymbol{\kappa}_f$ for the Wegscheider conditions to be satisfied, where $\mathbb{W}$ is an appropriately defined matrix. One way to incorporate the constraint $\boldsymbol{\kappa}_d = \mathbb{W}\boldsymbol{\kappa}_f$ into our Bayesian analysis problem is to treat it as prior information and apply it on the prior density of the unconstrained problem. This principle forms the basis of an attractive strategy for incorporating constraints into Bayesian analysis, known as encompassing prior approach (EPA) [32]. By following EPA, we can replace the previously discussed encompassing "effective" prior density $\int p(\boldsymbol{\kappa} \mid \boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z}$ by the following probability density function:

$$p_W(\boldsymbol{\kappa}_f, \boldsymbol{\kappa}_d) := \frac{\delta(\boldsymbol{\kappa}_d - \mathbb{W}\boldsymbol{\kappa}_f) \displaystyle\int p(\boldsymbol{\kappa}_f, \boldsymbol{\kappa}_d \mid \boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z}}{\displaystyle\int\int\int \delta(\boldsymbol{\kappa}_d - \mathbb{W}\boldsymbol{\kappa}_f) p(\boldsymbol{\kappa}_f, \boldsymbol{\kappa}_d \mid \boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z} d\boldsymbol{\kappa}_f d\boldsymbol{\kappa}_d}, \qquad (23)$$

where δ is the Dirac delta function. Clearly, this density assigns zero probability to kinetic parameters that do not satisfy the Wegscheider conditions, since $\delta(\boldsymbol{\kappa}_d - \mathbb{W}\boldsymbol{\kappa}_f) = 0$, if $\boldsymbol{\kappa}_d \neq \mathbb{W}\boldsymbol{\kappa}_f$. Note now that the log-rate constants $\boldsymbol{\kappa}_d$ are of no immediate interest, since their values can be determined as soon as the values of the log-rate constants $\boldsymbol{\kappa}_f$ have been estimated. As a consequence, we can treat $\boldsymbol{\kappa}_d$ as "nuisance" parameters and integrate them out of the problem [13]. This integration, together with the updated prior density we presented above, leads to the following *marginal* posterior density of the log-rate constants $\boldsymbol{\kappa}_f$:

$$p_W(\boldsymbol{\kappa}_f \mid \boldsymbol{y}) \propto \int \delta(\boldsymbol{\kappa}_d - \mathbb{W}\boldsymbol{\kappa}_f) p(\boldsymbol{\kappa}_f, \boldsymbol{\kappa}_d \mid \boldsymbol{y}) d\boldsymbol{\kappa}_d = p(\boldsymbol{\kappa}_f, \mathbb{W}\boldsymbol{\kappa}_f \mid \boldsymbol{y}). \tag{24}$$

Clearly, the values of the marginal posterior $p_W(\boldsymbol{\kappa}_f \mid \boldsymbol{y})$ are proportional to the corresponding values of the original posterior density $p(\boldsymbol{\kappa}_f, \boldsymbol{\kappa}_d \mid \boldsymbol{y})$ over the thermodynamically feasible region of the parameter space, given by the hyperplane $\boldsymbol{\kappa}_d = \mathbb{W}\boldsymbol{\kappa}_f$. In the following, we will base our Bayesian analysis approach on $p_W(\boldsymbol{\kappa}_f \mid \boldsymbol{y})$.

Computing the posterior mode

In a Bayesian setting, we use the location of the posterior density over the parameter space to provide an estimate of the unknown parameter values. Typically, two measures of location are employed, namely the mode and the mean of the posterior. The posterior mean minimizes the mean-square error between the estimated and true parameters, whereas, the posterior mode is more likely to produce dynamics that closely resemble the true dynamics (see Additional file 1 for why this is true). We note here that the main objective of parameter estimation in biochemical reaction systems is not necessarily to determine parameter values that are “close” to the true values (e.g., in the mean square sense) but to obtain appropriate values for the rate constants so that the resulting molecular concentration dynamics closely reproduce the dynamics observed in the true system [33]. As a consequence, we choose the posterior mode as our parameter estimator.

The posterior log-density $\ln p_W(\boldsymbol{\kappa}_f \mid \boldsymbol{y})$ is usually not concave, especially when a limited amount of highly noisy data $\boldsymbol{y}$ is available. As a consequence, there is no optimization algorithm that can find the posterior mode in a finite number of steps. A method to address this problem would be to randomly sample the parameter space at a predefined (and usually large) number of points and use these points to initialize an optimization algorithm, such as the simultaneous perturbation stochastic approximation (SPSA) algorithm discussed in the Additional file 2. We can then calculate the parameters and the associated values of $\ln p_W(\boldsymbol{\kappa}_f \mid \boldsymbol{y})$ obtained by each initialization after a set number of optimization steps, and declare the parameters associated with the highest log-posterior value as being the desired mode estimates.

Unfortunately, SPSA (and as a matter of fact any other appropriate optimization algorithm) is computationally costly, especially in the case of large biochemical reaction systems. Therefore, using SPSA in the previous multi-seed strategy may result in a computationally prohibitive approach for finding the posterior mode. In order to reduce computations, we may choose only a small number of initial points that we believe are sufficiently proximal to the posterior mode. Two such points might be the prior and posterior means. As a matter of fact, as the data sample size tends to infinity, we expect that the posterior mean will coincide with the posterior mode, since, under suitable regularity conditions, the posterior density converges asymptotically to a Gaussian distribution [12, 34]. This simple idea leads to the sequential maximization-expectation-maximization (MEM) algorithm we discuss in the Additional file 2. According to this algorithm, we perform a relatively small number of SPSA iterations, initialized by the prior mode, to obtain a posterior mode estimate $\widehat{\boldsymbol{\kappa}}_{f,1}^{\text{mode}}$. We then use an MCMC algorithm, initialized by $\widehat{\boldsymbol{\kappa}}_{f,1}^{\text{mode}}$, to obtain an estimate of the posterior mean $\widehat{\boldsymbol{\kappa}}_{f}^{\text{mean}}$. Subsequently, we perform another set of SPSA iterations, initialized by $\widehat{\boldsymbol{\kappa}}_{f}^{\text{mean}}$, to obtain the posterior mode estimate $\widehat{\boldsymbol{\kappa}}_{f,2}^{\text{mode}}$. We finally set $\widehat{\boldsymbol{\kappa}}_{f}^{\text{mode}}$ to be the log-rate constants that produce the maximum posterior value during all SPSA and MCMC iterations, and set the optimal estimate $\widehat{\boldsymbol{\kappa}}$ of the log-rate constants $\boldsymbol{\kappa}$ equal to $\{\widehat{\boldsymbol{\kappa}}_{f}^{\text{mode}}, \mathbb{W}\widehat{\boldsymbol{\kappa}}_{f}^{\text{mode}}\}$.

Estimation accuracy

One way to quantify the accuracy of the posterior mode estimate $\widehat{\kappa}_{f}^{\text{mode}}$ of a "free" log-rate constant κ_f is to calculate and report the root mean square error (RMSE), given by

$$\epsilon_{\text{RMSE}}(\widehat{\kappa}_{f}^{\text{mode}}) = \sqrt{\text{E}[(\kappa_f - \widehat{\kappa}_{f}^{\text{mode}})^2 \mid \boldsymbol{y}]} = \left\{ \int (\kappa_f - \widehat{\kappa}_{f}^{\text{mode}})^2 p_W(\boldsymbol{\kappa}_f \mid \boldsymbol{y}) d\boldsymbol{\kappa}_f \right\}^{1/2}. \tag{25}$$

A small value of ϵ_{RMSE} provides us with confidence that the estimated value of that constant is accurate. On the other hand, the estimate may be perceived as inaccurate if ϵ_{RMSE} is exceedingly large.

Another useful metric for evaluating estimation accuracy is $D := \ln \det[\mathbb{V}]/(M + M_1)$, where $\det[\mathbb{V}]$ is the determinant of the posterior covariance matrix $\mathbb{V} = \text{E}[(\boldsymbol{\kappa}_f - \widehat{\boldsymbol{\kappa}}_{f}^{\text{mode}})(\boldsymbol{\kappa}_f - \widehat{\boldsymbol{\kappa}}_{f}^{\text{mode}})^T \mid \boldsymbol{y}]$. Note that D is the average of the log-eigenvalues of $\mathbb{V}$ and is related to the well-known D-optimal criterion used in experimental design [27]. We can use D to quantify the overall accuracy of a model calibration result, with smaller values of D indicating better overall accuracy.

Note that the RMSE's ϵ_{RMSE} can be computed from the diagonal elements of $\mathbb{V}$. It turns out the we can approximate ϵ_{RMSE} and D from an estimate $\widehat{\mathbb{V}}$ of the posterior covariance matrix $\mathbb{V}$ obtained during the second (MCMC) phase of the proposed MEM algorithm (see Additional file 2 for details).

When the true values $\boldsymbol{\kappa}^{\text{true}}$ of the log-rate constants are known (which is the case when we use simulated data to evaluate the performance of the proposed Bayesian analysis approach, as we do in this paper), we can provide a more direct evaluation of estimation performance. As we have mentioned previously, calculating a measure of "closeness" (such as the square error) between the estimated and true parameter values may not be quite appropriate here. Since, in reality, our objective is to estimate the rate constant values so that the biochemical reaction system produces dynamics that closely match the true molecular dynamics, it may be more appropriate to use, as measures of estimation performance, the following *median* and *maximum* absolute error criteria:

$$\begin{aligned}\epsilon_{\text{MED-AE}} &= \underset{n\in\mathcal{N},\, p\in\mathcal{P}}{\text{med}} \left\{ \frac{\displaystyle\int_{\mathcal{T}} \left| \widehat{x}_n^{(p)}(t) - x_n^{(p)}(t) \right| dt}{\displaystyle\int_{\mathcal{T}} x_n^{(p)}(t)dt} \right\} \\ \epsilon_{\text{MAX-AE}} &= \max_{n\in\mathcal{N},\, p\in\mathcal{P}} \left\{ \frac{\displaystyle\int_{\mathcal{T}} \left| \widehat{x}_n^{(p)}(t) - x_n^{(p)}(t) \right| dt}{\displaystyle\int_{\mathcal{T}} x_n^{(p)}(t)dt} \right\},\end{aligned} \tag{26}$$

where $\{x_n^{(p)}(t), t \in \mathcal{T}\}$ and $\{\widehat{x}_n^{(p)}(t), t \in \mathcal{T}\}$ are the true and estimated dynamics of the n^{th} molecular species under the p^{th} perturbation, produced by the biochemical reaction system with log-rate constants $\boldsymbol{\kappa}^{\text{true}}$ and $\widehat{\boldsymbol{\kappa}} = \{\widehat{\boldsymbol{\kappa}}_f^{\text{mode}}, \mathbb{W}\widehat{\boldsymbol{\kappa}}_f^{\text{mode}}\}$, respectively. Clearly, $\epsilon_{\text{MED-AE}}$ and $\epsilon_{\text{MAX-AE}}$ provide measures of closeness between the estimated molecular responses $\{\widehat{x}_n^{(p)}(t), t \in \mathcal{T}, n \in \mathcal{N}, p \in \mathcal{P}\}$ and the true molecular responses $\{x_n^{(p)}(t), t \in \mathcal{T}, n \in \mathcal{N}, p \in \mathcal{P}\}$, normalized by the corresponding true integrated responses $\{\int_{\mathcal{T}} x_n^{(p)}(t)dt, n \in \mathcal{N}, p \in \mathcal{P}\}$. Normalization is required in order to make sure that no one species dominates the error values more than any other. Finally, note that half of the normalized absolute errors will be between 0 and $\epsilon_{\text{MED-AE}}$, whereas, the remaining half will be between $\epsilon_{\text{MED-AE}}$ and $\epsilon_{\text{MAX-AE}}$.

Results/Discussion

To illustrate key aspects of the previous Bayesian analysis methodology, we now consider a numerical example based on a subset of a well-established model of the EGF/ERK signal transduction pathway proposed by Schoeberl *et al.* [35]. This model corresponds to an open biochemical reaction system, since it contains irreversible reactions as well as reactions governed by Michaelis-Menten kinetics that involve molecular species not included in the model. We extract a closed subset of the Schoeberl model by choosing the largest connected section that contains only reversible reactions governed by mass action kinetics. The resulting biochemical reaction system is depicted in Figure 1 and is comprised of $N = 13$

molecular species that interact through $M = 9$ reversible reactions. Of course, we could attempt to generate a closed biochemical reaction system for the entire EGF/ERK signaling pathway, by including all relevant molecular species not considered by the Schoeberl model (e.g., ADP, ATP, intermediate forms in catalyzed reactions, etc.). However, since we are only interested in demonstrating the potential and key properties of our Bayesian analysis methodology, we found this to be unnecessary. We feel that the biochemical reaction system depicted in Figure 1 leads to a sufficiently rich numerical example that serves the main purpose of this section well.

In specifying the model depicted in Figure 1, we must provide three sets of physically reasonable values: true rate constant values, initial concentrations, and experimentally feasible perturbations to the initial concentrations. Published values for the reaction rate constants associated with our example are given in Equation (S-3.1) of Additional file 3. However, these values do not correspond to a thermodynamically feasible biochemical reaction system, since they do not satisfy the Wegscheider conditions, given by (11). We should point out here that this is a common problem in systems biology. Reaction rate values are usually amalgamated from various independent sources in the literature, so it is highly unlikely that these values will correspond to a thermodynamically feasible biochemical reaction system. As a consequence, it is desirable to develop a method that uses published values for the reaction rate constants and calculates an appropriate set of thermodynamically feasible values that can be considered as the "true" parameter values. In Additional file 3, we calculate "true" values for the log-rate constants by using a linear least squares approach to project the published values onto the thermodynamically feasible hyperplane. The resulting "true" values are given in Equation (S-3.3) of Additional file 3.

Regarding the initial concentrations, we use the values specified in [35, 36], with two minor modifications. First, molecular species with zero initial concentrations are modified to have a small number of molecules present. We do this to accommodate the fact that, in a real cellular system, these molecular species are constitutively expressed. The second modification comes from the fact that we are no longer modeling the entire EGF/ERK signaling cascade and, therefore, we must account for the upstream EGF stimulus. To take this into account, we increase the initial concentration of the most upstream molecular species in our model, namely (EGF-EGFR*)2-GAP. The initial concentrations used are given by Equation (S-3.4) in Additional file 3.

To specify appropriate perturbations to the initial molecular concentrations, note that molecular complexes, such as dimers, trimers, etc., are far more difficult to perturb than simple monomeric molecular species. For this reason, we focus our perturbation efforts on Shc*, Grb2, and Sos. Since Shc* is commercially available in a purified and quantified form, we will assume that we can increase its initial concentration by a factor of 100 using molecular injection. We will also assume that we can perturb Grb2 and Sos by RNAi, resulting in a decrease in their initial concentrations by a factor of 100. Thus, we set $\pi_1 = 99c_1$, $\pi_2 = -.99c_2$, and $\pi_4 = -.99c_4$.

To avoid specifying different hyperparameter values for the prior densities of the forward log-rate constants, we assume here that all densities share the same known values $\{\kappa^0, \tau, \lambda\}$, where $\kappa^0 = -5.1010$, $\tau = 1.8990$, and $\lambda = 0.7409$, whereas, we set $a = 3$ and $b = 1$ for the hyperparameters of the prior density of the variance σ^2 of the measurement errors. These choices correspond to the prior densities depicted in Figure S-1.2(a) and Figure S-1.3(a) in Additional file 1. We implement our Bayesian analysis approach using the MEM algorithm described in Additional file 2, with $I = 5{,}000$ SPSA iterations in each maximization step and a total of $L = 50{,}000$ MCMC iterations in the expectation step. Finally, we observe the biochemical reaction system within a time period of 1 min.

In Figure 2, we depict a typical result obtained by the proposed Bayesian analysis algorithm. In this figure, we compare the estimated log-rate values (blue) with the thermodynamically consistent true log-rate values (red) as well as the corresponding concentration dynamics of selected molecular species in the unperturbed biochemical reaction system. We have obtained these results by measuring the concentration dynamics in the unperturbed and perturbed systems at $Q = 6$ logarithmically-spaced time points (green circles), with the measurements being corrupted by independent and identically distributed (i.i.d.) zero-mean Gaussian noise with standard deviation $\sigma = 0.3$. Moreover, we summarize the estimated posterior RMSE values, given by (25), in Table 1. Finally, the calculated median and maximum absolute error values, given by (26), are 3.03×10^{-2} and 1.68×10^{-1}, respectively.

The concentration dynamics produced by the estimated rate constant values match well the dynamics produced by the true values. As a matter of fact, the calculated median and maximum absolute error values imply that half of the relative integrated absolute error values between the estimated and true concentration dynamics (across all molecular species and all applied perturbations) are smaller than 3.03%, whereas, the remaining values are between 3.03% and 16.8%. On the other hand, the estimated posterior RMSE values summarized in Table 1 indicate a high probability that, given the concentration measurements, the log-rate values will lie within a relatively small region around the corresponding posterior mode values.

We expect that, in general, by selecting appropriate perturbations and by increasing the number of concentration data collected during an experiment, we can improve estimation accuracy. However, how can one know if the right perturbations have been applied on the biochemical reaction system and if enough data has been collected in a practical situation? Inspection of RMSE values can provide an answer to these important questions. If the estimated RMSE values of the log-rate constants of many reactions are large, it may be worth collecting additional data by increasing P and Q. Additional data can improve estimation accuracy by shrinking the RMSE values to a size that indicates an acceptable degree of uncertainty. However, if the biochemical reaction system is insensitive to a given kinetic parameter, then the RMSE associated with that reaction may remain large even as the quality of data improves. Therefore, additional data should only be collected when the RMSE values are large and sensitivity analysis indicates that the values of the rate constants associated with these RMSE values appreciably affect the system dynamics.

The RMSE values do not provide a global measure of estimation accuracy, since some parameters may have small RMSE values and some may have large values. To address this problem, we may instead employ the D-optimal criterion as a measure of estimation accuracy. As a matter of fact, we can effectively use the D-optimal criterion as a guide for selecting appropriate perturbations and for determining the data sampling scheme we must use in order to increase estimation accuracy. In Table 2, for example, we summarize estimated values of D, for the case of uniform and logarithmic sampling, calculated for different values of Q. Clearly, the sampling scheme used may appreciably affect estimation performance. For each value of Q, uniform sampling results in higher values of D than logarithmic sampling. As a consequence, we must use logarithmic sampling over uniform sampling, since the former may produce better estimation accuracy than the latter. This is expected, since uniform sampling may result in measuring steady-state concentrations much more often than (short-lived) transient concentrations. On the other hand, logarithmic sampling may be used to gather valuable information about the transient behavior of a biochemical reaction system while placing less emphasis on its steady-state dynamics (which only provide information about the equilibrium constants of the underlying reactions). The results depicted in Table 2 also suggest an appropriate value for Q. If our goal is to find the smallest value Q^* of Q (an objective dictated by the high cost of experimentally measuring molecular concentrations) which results in a value of D that is no less than, say 5%, of the value obtained when $Q = Q^* - 1$, then we must set $Q^* = 6$.

In Table 3, we summarize the estimated values of D obtained from seven different perturbation experiments (logarithmic sampling is used with $Q = 6$). Moreover, we report the D values obtained by repeating an experiment that does not use molecular perturbations. Experimental replication may be an effective approach to obtain additional data, especially when molecular perturbations are costly or difficult to apply. Our formulation allows us to consider this scenario by setting $\pi_p = 0$, for every $p \in \mathcal{P}$. The data collected this way correspond to repeating the same experiment $P + 1$ times, where P is the number of elements in $\mathcal{P}$. The maximum experimental replication considered in Table 3 uses $P = 3$, which corresponds to repeating the same experiment four times. This produces the same amount of data as the data obtained by perturbing the initial concentrations of Shc*, Grb2, and Sos, one at a time. The values depicted in Table 3 suggest that perturbing the initial concentrations of Shc*, Grb2, and Sos may be the right thing to do, since this produces the lowest value of D and, thus, it may result in better estimation performance as compared to perturbing the initial concentrations of one or two of these molecular species. In this case, however, it may also be acceptable to replicate an experiment that does not use molecular perturbations, since the minimum value of D is only 9.31% lower than the D value obtained by repeating the experiment four times.

One of the underlying assumptions associated with the proposed Bayesian analysis algorithm is that the measurement errors are statistically independent, following a zero-mean Gaussian distribution with standard deviation σ. To assess the adequacy of this assumption and evaluate its implication on estimation performance, we depict in Table 4 calculated median and maximum absolute error values obtained when the measurement errors $\eta_n^{(p)}$ in (5) are i.i.d. zero-mean Gaussian with standard deviation σ, i.i.d. zero-mean uniform within the interval $[-\sqrt{3}\sigma, \sqrt{3}\sigma]$, with standard deviation σ, and correlated zero-mean stationary Gaussian with autocorrelation $\mathrm{E}[\eta_n^{(p)}(t_1)\eta_n^{(p)}(t_2)] = \sigma^2 \exp\{-|t_1 - t_2|\}$. We consider different values for the standard deviation, namely $\sigma = 0.1, 0.2, 0.3, 0.4, 0.5$, and measure the concentration dynamics in the unperturbed and perturbed systems at $Q = 6$ logarithmically spaced time points. Table 4 shows clearly that violation of the i.i.d. Gaussian assumption may lead to reduction in estimation accuracy, especially when the measurement errors are correlated, due to an increase in the maximum absolute error values. However, the calculated median absolute error values indicate that the proposed algorithm is relatively robust to the statistical behavior of the measurement errors, producing reasonable estimates for at least half of the concentration dynamics. In Figure 3, we depict results obtained by the proposed Bayesian analysis algorithm when measuring the concentration dynamics in the unperturbed and perturbed systems at $Q = 6$ logarithmically-spaced time points (green circles), with the measurements being corrupted by correlated zero-mean stationary Gaussian errors with standard deviation $\sigma = 0.3$. These results compare favorably to the ones depicted in Figure 2. In this case, the calculated median absolute error value is 1.48×10^{-2}, which is 62.8% smaller that the value obtained when the errors are i.i.d. zero-mean Gaussian, whereas, the calculated maximum absolute error value is 7.32×10^{-2}, which is 31.7% larger that the value obtained when the errors are i.i.d. zero-mean Gaussian.

Conclusions

In this paper, we have introduced a novel Bayesian analysis technique for estimating the kinetic parameters (rate constants) of a closed biochemical reaction system from time measurements of noisy concentration dynamics. The proposed procedure enjoys a clear advantage over other published estimation techniques: the estimated kinetic parameters satisfy the Wegscheider conditions imposed by the fundamental laws of thermodynamics. As a consequence, it always leads to physically plausible biochemical reaction systems.

From a statistical perspective, there are additional advantages for thermodynamically restricting the kinetic parameters of a biochemical reaction system to satisfy the Wegscheider conditions. This may be seen through the well-known bias-variance tradeoff in estimation [27]. The mean squared error of a given estimator can be decomposed into a bias term and a variance term. In general, imposing constraints on the estimator may increase its bias but decrease its variance (hence the tradeoff). However, if the true parameter values satisfy the constraints, then the variance may decrease without increasing the bias term [27]. Since the true values of the kinetic parameters must lie on the thermodynamically feasible manifold in the parameter space, confining the Bayesian estimator to this manifold (which is of lower dimension than the parameter space itself) may lead to lower mean squared error due to a smaller variance. Since the thermodynamically feasible manifold is of lower dimension than the parameter space, gains in variance (and hence improvements in the mean squared error) are expected to be large. This may be seen through the "curse of dimensionality," which refers to the exponential increase in the volume of the parameter space as its dimension grows, making estimation exponentially harder in higher dimensional spaces (in our example, the unconstrained parameter space has 12.5% more dimensions than the thermodynamically feasible subspace). The Wegscheider conditions reduce the dimensionality of the parameter space to a feasible region in which estimation may be easier. Thus, the proposed Bayesian analysis procedure improves on other estimation techniques by producing a statistically superior, physically meaningful and plausible estimate for the kinetic parameters of a closed biochemical reaction system.

The Bayesian analysis methodology discussed in this paper has been formulated by assuming that all initial concentrations and perturbations are precisely known and that concentration measurements can be obtained by directly sampling all system dynamics. However, current experimental practices in quantitative systems biology restrict the amount and type of data that can be collected from living cells. As a consequence, further research is needed to develop approaches that can accommodate this important issue and make a Bayesian analysis approach to parameter estimation better applicable to systems biology problems.

If the initial concentrations and the perturbations applied on these concentrations are not known, then we may try to estimate them together with the unknown kinetic parameters. Although formulation of this problem is similar to the one considered in this paper, the additional computational burden will be substantial. Moreover, while quantitative biochemical techniques are improving, the vast majority of data available in problems of systems biology are obtained by measuring ratios of molecular concentrations (e.g., by using techniques such as SILAC [37]). Estimation of the rate constants of a biochemical reaction system from concentration measurements available as ratios relative to a reference system requires special consideration and extensive modification of the proposed Bayesian analysis procedure. Finally, it is very important to address the problem of missing observations. This is a common problem in systems biology, since it is not possible to monitor and measure the concentrations of all molecular species present in the system. Although appropriate modifications to the proposed algorithm can lead to a Bayesian analysis approach that can handle missing data, we think that development of a practically effective way to address this problem is challenging. Our future plan is to expand and improve the Bayesian analysis procedure discussed in this paper in order to provide practical solutions to the previous problems.

It is worth noting here that the estimation procedure suggested in this paper applies only to closed biochemical reaction systems (or to approximations of closed systems embedded in a larger open system). However, a cell is an open system, since it effectively interacts with its environment. If we include the cell's environment into our system and monitor the combined system until steady-state (i.e., until cell death), then we would have the necessary closed system. Unfortunately, this is clearly an unrealistic scenario. As a consequence, there is also a need to develop a theoretical and computational approach for dealing with thermodynamically consistent parameter estimation in open biochemical reaction systems.

To conclude, it has been argued in a recent paper [33] that most models of computational systems biology are "sloppy," in the sense that many parameters of such models do not appreciably alter system behavior.

A key conclusion of this paper is that collective fitting procedures (such as the Bayesian analysis technique presented in the present paper) are far more desirable than piecewise construction of a biochemical reaction system model from individual parameter estimates (which is how most models are constructed when investigators scour the literature for individual rate constant values). Moreover, it has been pointed out in [33] that using a method to obtain precise parameter values may be difficult, even with an unlimited amount of data, since the behavior of a sloppy model is insensitive to the values of most parameters. As a consequence, the authors suggest that, instead of focusing on the quality of parameter estimation, it will be more wise to focus on the quality of prediction achieved by an estimated model (as we have also argued in this paper).

To a certain extent, our Bayesian analysis approach addresses some of the issues raised in [33]. By imposing the Wegscheider conditions on the kinetic parameters of a biochemical reaction system, we can effectively constrain these parameters to a thermodynamically feasible manifold in the parameter space, thus reducing sloppiness. Moreover, we can effectively use the RMSE values and the D-optimal criterion to determine an appropriate experimental design and distinguish those estimated values that can be trusted from those that cannot. For example, if the RMSE value associated with a kinetic parameter is small, then we may trust these values. On the other hand, a large RMSE value may indicate high uncertainty in the estimated parameter values, which may be untrustworthy. As we mentioned before, if a sensitivity analysis approach, such as the one proposed in [38], indicates that the kinetic parameters associated with large RMSE values are influential parameters, then we must reduce these RMSE values to an acceptable level of uncertainty by adopting a new and more effective experimental design approach. On the other hand, if these parameters correspond to a non-influential reaction, then we can accept the estimated values with no further consideration, since high uncertainty in the exact values of these parameters will not affect the predicted concentration dynamics.

Authors' contributions

JG developed the basic Bayesian analysis framework, derived most theoretical results, coded a substantial portion of the software, and wrote the final version of the paper. GJ derived a number of theoretical results, generated much of the material in Additional files 2 & 3, coded much of the final version of the software, and interpreted the obtained computational results. XZ and JG advised on several theoretical aspects of the paper, algorithm design, and data analysis, and wrote early versions of the software. GJ and JG advised on various theoretical aspects of the paper, on algorithm design, and data analysis. All authors read and approved the final version of the paper.

Acknowledgements

GJ has been supported by The National Defense Science and Engineering Graduate (NDSEG) Fellowship through the Department of Defense. JG acknowledges the National Science Foundation (NSF) for support of this research. The authors thank the anonymous reviewers for their insightful comments, which resulted in a great improvement of the paper. Special thanks to Ali Afshar for identifying several problems with the original version of this paper and for suggesting appropriate fixes.

References

1. Crampin EJ, Schnell S, McSharry PE: **Mathematical and computational techniques to deduce complex biochemical reaction mechanisms**. *Prog. Biophys. Mol. Bio.* 2004, **86**:77–112.
2. Feng XJ, Rabitz H: **Optimal identification of biochemical reaction networks**. *Biophys. J.* 2004, **86**:1270–1281.
3. Maria G: **A review of algorithms and trends in kinetic model identification for chemical and biochemical systems**. *Chem. Biochem. Eng. Q.* 2004, **18**(3):195–222.
4. Papin JA, Hunter T, Palsson BO, Subramaniam S: **Reconstruction of cellular signalling networks and analysis of their properties**. *Nat. Rev. Mol. Cell Bio.* 2005, **6**:99–111.
5. Ronen M, Rosenberg R, Shraiman BI, Alon U: **Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics**. *P. Natl. Acad. Sci. USA* 2002, **99**(16):10555–10560.
6. Rodriguez-Fernandez M, Mendes P, Banga JR: **A hybrid approach for efficient and robust parameter estimation in biochemical pathways**. *BioSystems* 2006, **83**:248–265.
7. Liebermeister W, Klipp E: **Bringing metabolic networks to life: integration of kinetic, metabolic, and proteomic data**. *Theor. Biol. Med. Model.* 2006, **3**(42).
8. Quach M, Brunel N, d'Alché Buc F: **Estimating parameters and hidden variables in non-linear state-space models based on ODEs for biological networks inference**. *Bioinformatics* 2007, **23**(23):3209–3216.
9. Balsa-Canto E, Peifer M, Banga JR, Timmer J, Fleck C: **Hybrid optimization method with general switching strategy for parameter estimation**. *BMC Syst. Biol.* 2008, **26**.
10. Klinke DJ: **An empirical Bayesian approach for model-based inference of cellular signaling networks**. *BMC Bioinformatics* 2009, **10**(371).
11. Mazur J, Ritter D, Reinelt G, Kaderali L: **Reconstructing nonlinear dynamic models of gene regulation using stochastic sampling**. *BMC Bioinformatics* 2009, **10**(448).
12. Berger JO: *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag, 2nd edition 1985.
13. Gelman A, Carlin JB, Stern HS, Rubin DB: *Bayesian Data Analysis*. Boca Raton, Florida: Chapman and Hall/CRC, 2nd edition 2004.
14. Heinrich R, Schuster S: *The Regulation of Cellular Systems*. New York: Chapman & Hall 1996.
15. Vlad MO, Ross J: **Thermodynamically based constraints for rate coefficients of large biochemical networks**. *WIREs Syst. Biol. Med.* 2009, **1**:348–358.

16. Alberty RA: **Princple of detailed balance in kinetics**. *J. Chem. Educ.* 2004, **81**(8):1206–1209.

17. Liebermeister W, Klipp E: **Bringing metabolic networks to life: convenience rate law and thermodynamic constraints**. *Theor. Biol. Med. Model.* 2006, **3**(41).

18. Yang J, Bruno WJ, Hlavacek WS, Pearson JE: **On imposing detailed balance in complex reaction mechanisms**. *Biophys. J.* 2006, **91**:1136–1141.

19. Ederer M, Gilles ED: **Thermodynamically feasible kinetic models of reaction networks**. *Biophys. J.* 2007, **92**:1846–1857.

20. Ederer M, Gilles ED: **Thermodynamic constraints in kinetic modeling: Thermodynamic-kinetic modeling in comparison to other approaches**. *Eng. Life Sci.* 2008, **8**(5):467–476.

21. Berry RS, Rice SA, Ross J: *Physical Chemistry*. New York: Oxford University Press, 2nd edition 2000.

22. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L: **Integrated genomic and proteomic analyses of a systematically perturbed metabolic network**. *Science* 2001, **292**:929–934.

23. Tegnér J, Yeung MKS, Hasty J, Collins JJ: **Reverse engineering gene networks: Integrating genetic perturbations with dynamical modeling**. *Proc. Nat. Acad. Sci. USA* 2003, **100**:5944–5949.

24. Zak DE, Gonye GE, Schwaber JS, Doyle FJ: **Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks: Insights from an identifiability analysis of an in silico network**. *Genome Res.* 2003, **13**:2396–2405.

25. Liu JS: *Monte Carlo Strategies in Scientific Computing*. New York: Springer-Verlag 2001.

26. Moles CG, Mendes P, Banga JR: **Parameter estimation in biochemical pathways: A comparison of global optimization methods**. *Genome Res.* 2007, **13**:2467–2474.

27. Spall JC: *Introduction to Stochastic Search and Optimization: Estimation, Simulation and Control*. New York: Wiley-Interscience 2003.

28. Šašik R, Calvo E, Corbeil J: **Statistical analysis of high-density oligonucleotide arrays: a multiplicative noise model**. *Bioinf.* 2002, **18**:1633–1640.

29. Anderle M, Roy S, Lin H, Becker C, Joho K: **Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum**. *Bioinf.* 2004, **20**:3575–3582.

30. Listgarten J, Emili A: **Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry**. *Mol. Cell. Proteomics* 2005, **4**:419–434.

31. Molina H, Parmigiani G, Pandey A: **Assessing reproducibility of a protein dynamics study using in vivo labeling and liquid chromatography tandem mass spectrometry**. *Anal. Chem.* 2005, **77**:2739–2744.

32. Klugkist I, Hoijtink H: **The Bayes factor for inequality and about equality constrained models**. *Comput. Stat. Data An.* 2007, **51**:6367–6379.

33. Gutenkunst RN, Waterfall JJ, Casey FP, Brown KS, Myers CR, Sethna JP: **Universally sloppy parameter sensitivities in systems biology models**. *PLOS Comp. Biol.* 2007, **3**(10):1871–1878.

34. Walker AM: **On the asymptotic behaviour of posterior distributions**. *J. Roy. Stat. Soc. B Met.* 1969, **31**:80–88.

35. Schoeberl B, Eichler-Jonsson C, Gilles ED, Müller G: **Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors**. *Nat. Biotechnol.* 2002, **20**:370–375.

36. Li C, Donizelli M, Rodriguez N, Dharuri H, Endler L, Chelliah V, Li L, He E, Henry A, Stefan MI, Snoep JL, Hucka M, Novère NL, Laibe C: **BioModels database: An enhanced, curated and annotated resource for published quantitative kinetic models**. *BMC Syst. Biol.* 2010, **4**:92.

37. Ong SE, Mann M: **A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC)**. *Nat. Protoc* 2006, **1**:2650–2660.

38. Zhang HX, Dempsey WP, Goutsias J: **Probabilistic sensitivity analysis of biochemical reaction systems**. *J. Chem. Phys.* 2009, **131**(094101):1–20.

Figure legends

Figure 1: A subset of the EGF/ERK signal transduction pathway model proposed in [35]. The biochemical reaction system is comprised of $N = 13$ molecular species that interact through $M = 9$ reactions. Bayesian analysis is focused on estimating the values of the 18 rate constants associated with the reactions.

Figure 2: True (red) vs. estimated (blue) log-rate values and selected molecular dynamics in the unperturbed biochemical reaction system depicted in Figure 1. The results are based on measuring the dynamics in the unperturbed and perturbed systems at $Q = 6$ logarithmically-spaced time points (green circles). Perturbations are applied on the initial concentrations of Shc*, Grb2, and Sos, one at a time. The measurement errors are i.i.d. zero-mean Gaussian with standard deviation $\sigma = 0.3$.

Figure 3: True (red) vs. estimated (blue) log-rate values and selected molecular dynamics in the unperturbed biochemical reaction system depicted in Figure 1. The results are based on measuring the dynamics in the unperturbed and perturbed systems at $Q = 6$ logarithmically-spaced time points (green circles). Perturbations are applied on the initial concentrations of Shc*, Grb2, and Sos, one at a time. The measurement errors are correlated zero-mean Gaussian with standard deviation $\sigma = 0.3$.

Tables

Table 1: Estimated posterior RMSE values for the case of i.i.d. zero-mean Gaussian errors with standard deviation $\sigma = 0.3$. Logarithmic sampling is used with $Q = 6$. The log-rate constants κ_8 and κ_{14} are "dependent" variables. Therefore, no RMSE values are reported for these variables.

κ_1	κ_3	κ_5	κ_7	κ_9	κ_{11}	κ_{13}	κ_{15}	κ_{17}
0.2414	0.1578	0.1838	0.2950	0.1426	0.1683	0.0968	0.4474	0.1484
κ_2	κ_4	κ_6	κ_8	κ_{10}	κ_{12}	κ_{14}	κ_{16}	κ_{18}
0.2594	0.2095	0.1704	–	0.2124	0.2136	–	0.5093	0.0494

Table 2: Estimated values of the D-optimal criterion for uniform and logarithmic sampling schemes. The measurement errors are i.i.d. zero-mean Gaussian with standard deviation $\sigma = 0.3$.

Q	uniform	logarithmic	% change
2	-1.7697	-2.3500	–
3	-2.0030	-3.4287	45.90%
4	-2.3752	-3.7432	9.17%
5	-2.6115	-4.1173	9.99%
6	-2.3492	-4.1039	-0.33%

Table 3: Estimated values of the D-optimal criterion for different replications and perturbations. The measurement errors are i.i.d. zero-mean Gaussian with standard deviation $\sigma = 0.3$. Logarithmic sampling is used with $Q = 6$.

Perturbation	D
NO: 1 replication	-3.0123
NO: 2 replications	-3.4950
NO: 3 replications	-3.7544
YES: Shc*	-3.1398
YES: Grb2	-3.0747
YES: Sos	-3.4531
YES: Shc*, Grb2	-3.9279
YES: Shc*, Sos	-3.7716
YES: Grb2, Sos	-3.6363
YES: Shc*, Grb2, Sos	-4.1039

Table 4: Median and maximum absolute error values under a variety of measurement error conditions. Logarithmic sampling is used with $Q = 6$.

mean $= 0$	i.i.d. Gaussian	i.i.d. Uniform	correlated Gaussian
$\sigma = 0.1$	3.98×10^{-3}	8.64×10^{-3}	1.48×10^{-2}
	5.56×10^{-2}	4.81×10^{-2}	7.32×10^{-2}
$\sigma = 0.2$	1.01×10^{-2}	1.78×10^{-2}	3.09×10^{-2}
	8.29×10^{-2}	1.30×10^{-1}	1.89×10^{-1}
$\sigma = 0.3$	3.03×10^{-2}	1.78×10^{-2}	3.05×10^{-2}
	1.68×10^{-1}	1.30×10^{-1}	2.46×10^{-1}
$\sigma = 0.4$	2.19×10^{-2}	2.56×10^{-2}	1.04×10^{-1}
	2.27×10^{-1}	1.41×10^{-1}	3.67×10^{-1}
$\sigma = 0.5$	2.67×10^{-2}	3.86×10^{-2}	6.43×10^{-2}
	2.48×10^{-1}	3.32×10^{-1}	3.10×10^{-1}

Additional files provided with this submission

Additional file 1: addfile-1.pdf, 224K

In this document, we provide theoretical details necessary to understand the Bayesian analysis approach introduced in the Main text.

Additional file 2: addfile-2.pdf, 82K

This document contains a detailed description of the computational algorithms used for implementing various steps of the proposed Bayesian analysis approach.

Additional file 3: addfile-3.pdf, 80K

In this document, we list the biochemical reactions associated with our numerical example and provide thermodynamically consistent values for the rate constants as well as appropriate values for the initial molecular concentrations.

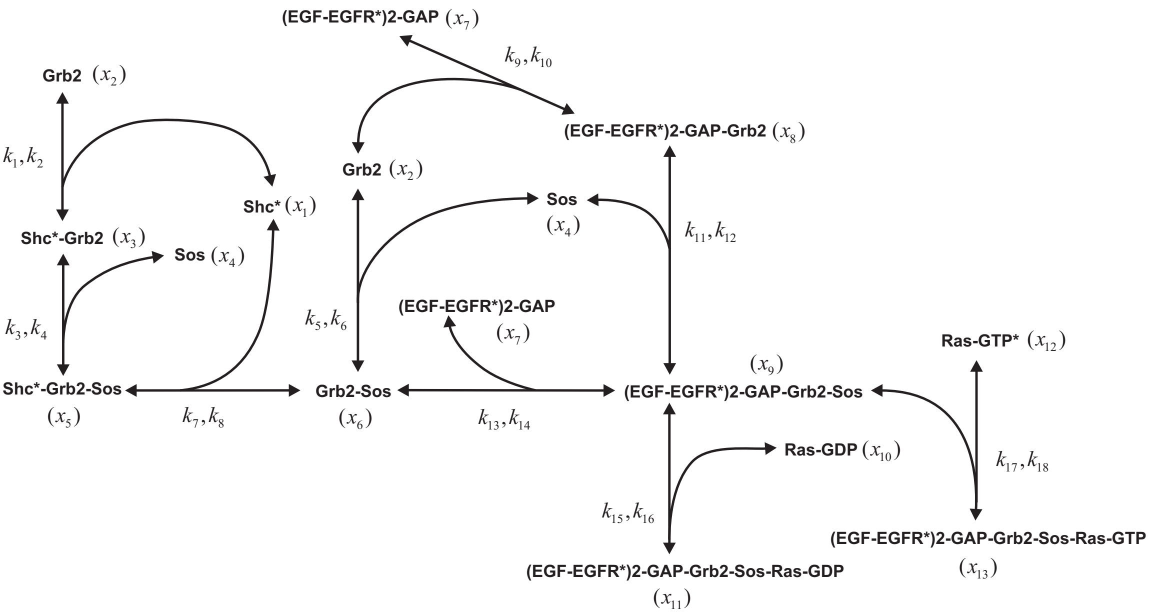
(EGF-EGFR*)2-GAP (x_7)
k_9, k_{10}
Grb2 (x_2)
k_1, k_2
(EGF-EGFR*)2-GAP-Grb2 (x_8)
Grb2 (x_2)
Shc* (x_1)
Sos (x_4)
Shc*-Grb2 (x_3)
Sos (x_4)
k_{11}, k_{12}
k_3, k_4
k_5, k_6
(EGF-EGFR*)2-GAP (x_7)
Ras-GTP* (x_{12})
(x_9)
Shc*-Grb2-Sos (x_5)
Grb2-Sos (x_6)
(EGF-EGFR*)2-GAP-Grb2-Sos
k_7, k_8
k_{13}, k_{14}
Ras-GDP (x_{10})
k_{17}, k_{18}
k_{15}, k_{16}
(EGF-EGFR*)2-GAP-Grb2-Sos-Ras-GTP (x_{13})
(EGF-EGFR*)2-GAP-Grb2-Sos-Ras-GDP (x_{11})

FIGURE 1

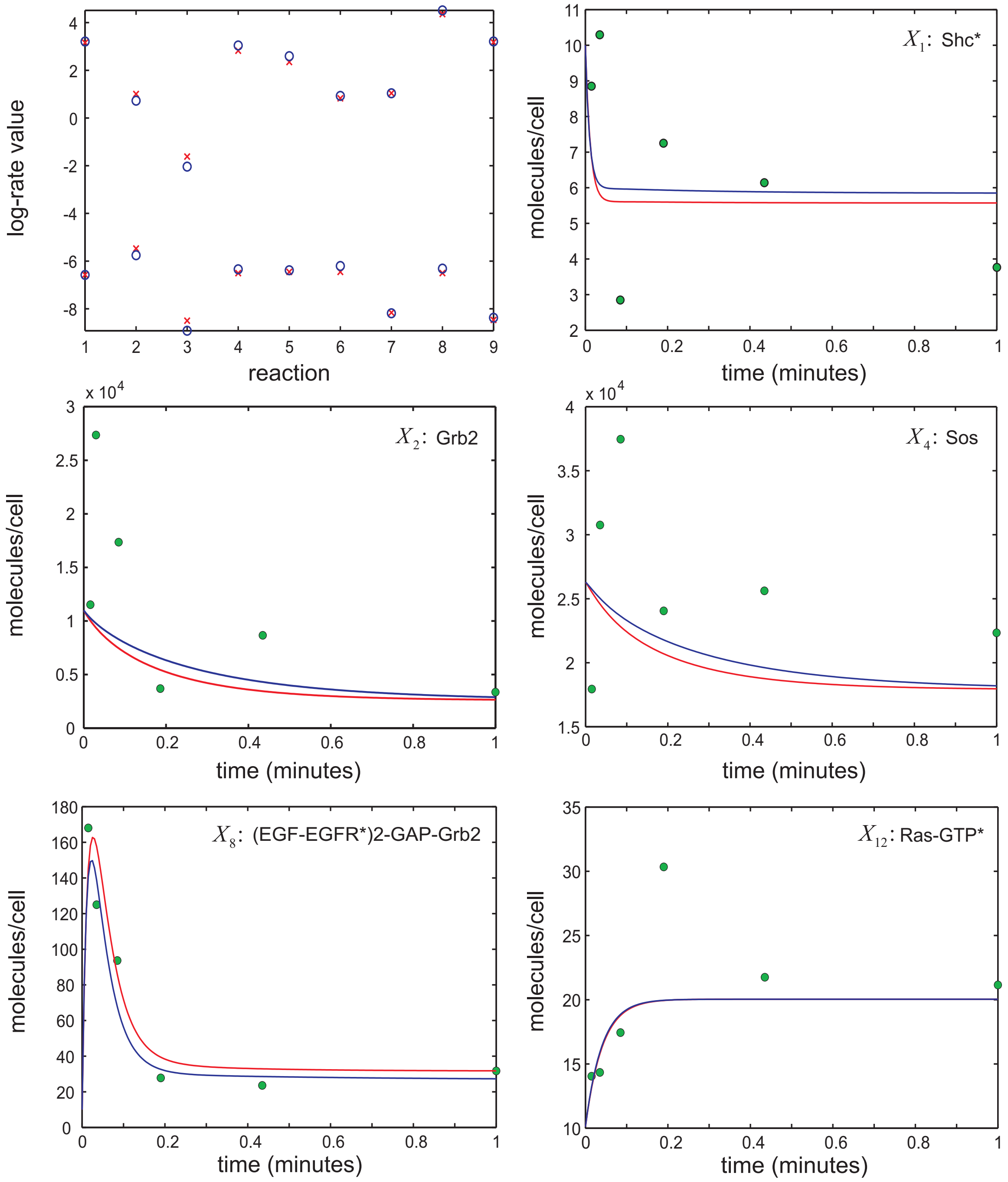
log-rate value
reaction
4
2
0
-2
-4
-6
-8
1
2
3
4
5
6
7
8
9
molecules/cell
time (minutes)
X_1: Shc*
11
10
9
8
7
6
5
4
3
2
0
0.2
0.4
0.6
0.8
1
x 10^4
X_2: Grb2
3
2.5
2
1.5
1
0.5
0
X_4: Sos
4
3.5
3
2.5
2
1.5
X_8: (EGF-EGFR*)2-GAP-Grb2
180
160
140
120
100
80
60
40
20
0
X_{12}: Ras-GTP*
35
30
25
20
15
10

FIGURE 2

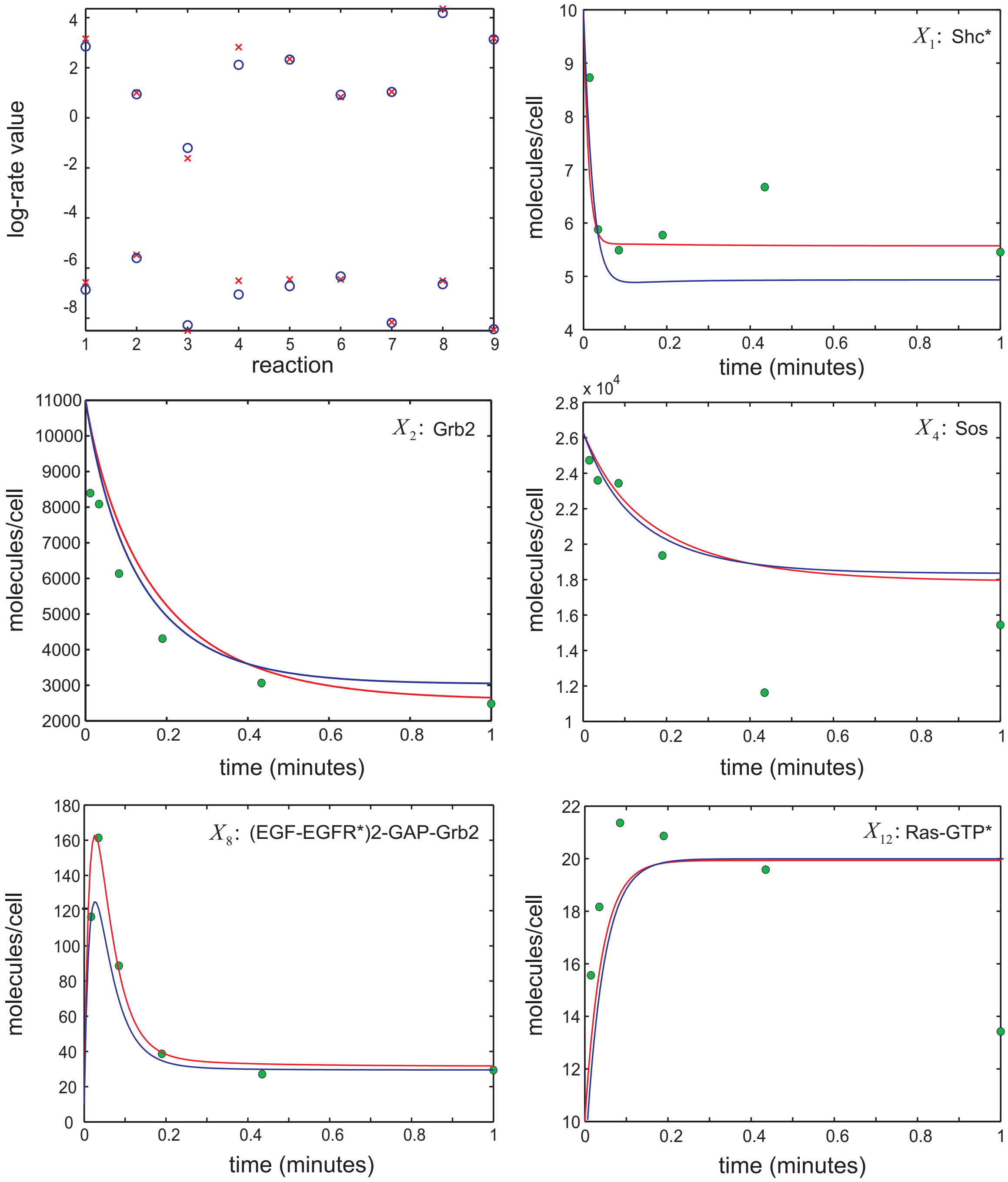
log-rate value
reaction
4
2
0
-2
-4
-6
-8
1
2
3
4
5
6
7
8
9
molecules/cell
time (minutes)
X_1: Shc*
X_2: Grb2
X_4: Sos
x 10^4
X_8: (EGF-EGFR*)2-GAP-Grb2
X_{12}: Ras-GTP*
0
0.2
0.4
0.6
0.8
1

FIGURE 3

ADDITIONAL FILE 1

Thermodynamically consistent Bayesian analysis of closed biochemical reaction systems

THEORY

Garrett Jenkinson,[1] Xiaogang Zhong,[2] and John Goutsias*[1]

[1]Whitaker Biomedical Engineering Institute, The Johns Hopkins University, Baltimore, MD 21218, USA
[2]Department of Applied Mathematics and Statistics, The Johns Hopkins University, Baltimore, MD 21218, USA

Email: Garrett Jenkinson - jenkinson@jhu.edu; Xiaogang Zhong - xzhong4@jhu.edu; John Goutsias*- goutsias@jhu.edu;

*Corresponding author

In this document, we provide theoretical details necessary to understand the Bayesian analysis approach introduced in the Main text. We start by deriving a general formula for the posterior density associated with our Bayesian analysis approach. We then discuss the Wegscheider conditions and their implications on the reaction rate constants of a closed biochemical reaction system. Subsequently, we derive appropriate prior probability density functions for the log-equilibrium constants as well as for the log-rate constants, and discuss a practical method for determining the hyperparameters associated with these priors. Finally, we present mathematical arguments to support our belief that the posterior mode should be a more preferable estimator of the kinetic parameters of a biochemical reaction system than the posterior mean.

Bayesian analysis

To develop a Bayesian analysis approach for estimating the kinetic parameters (rate constants) of a biochemical reaction system, we take the log-rate constants $\boldsymbol{\kappa}$ and the error variance σ^2 to be random variables. Note however that the probability density function of $\boldsymbol{\kappa}$ we consider in this paper depends on the log-equilibrium constants $\boldsymbol{z} = \{z_m, m \in \mathcal{M}\}$, where $z_m := \ln(k_{2m-1}/k_{2m})$, which we treat as *hyperparameters* (details to follow). Since we do not know the exact values of these hyperparameters, we take them to be random variables as well. By following this approach, we can write the *posterior* joint density of $\boldsymbol{\kappa}$, $\boldsymbol{z}$, and σ^2, given data $\boldsymbol{y}$, as[1]

$$p(\boldsymbol{\kappa}, \boldsymbol{z}, \sigma^2 \mid \boldsymbol{y}) = p(\sigma^2 \mid \boldsymbol{\kappa}, \boldsymbol{y}) p(\boldsymbol{\kappa} \mid \boldsymbol{z}, \boldsymbol{y}) p(\boldsymbol{z} \mid \boldsymbol{y}) \,. \tag{S-1.1}$$

Since our main objective is to estimate $\boldsymbol{\kappa}$, we are interested in evaluating the posterior density $p(\boldsymbol{\kappa} \mid \boldsymbol{y})$. To do so, we must integrate the log-equilibrium constants $\boldsymbol{z}$ and the error variance σ^2 out of the posterior joint density $p(\boldsymbol{\kappa}, \boldsymbol{z}, \sigma^2 \mid \boldsymbol{y})$. In this case,

$$p(\boldsymbol{\kappa} \mid \boldsymbol{y}) = \int \int p(\boldsymbol{\kappa}, \boldsymbol{z}, \sigma^2 \mid \boldsymbol{y}) d\sigma^2 d\boldsymbol{z} = \int p(\boldsymbol{\kappa} \mid \boldsymbol{z}, \boldsymbol{y}) p(\boldsymbol{z} \mid \boldsymbol{y}) d\boldsymbol{z}, \tag{S-1.2}$$

by virtue of (S-1.1). On the other hand,

$$p(\boldsymbol{\kappa} \mid \boldsymbol{z}, \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\kappa}) p(\boldsymbol{\kappa} \mid \boldsymbol{z}) p(\boldsymbol{z})}{p(\boldsymbol{z} \mid \boldsymbol{y}) p(\boldsymbol{y})} \,. \tag{S-1.3}$$

As a consequence of (S-1.2) and (S-1.3), we have that

$$\begin{aligned} p(\boldsymbol{\kappa} \mid \boldsymbol{y}) &= \frac{p(\boldsymbol{y} \mid \boldsymbol{\kappa})}{p(\boldsymbol{y})} \int p(\boldsymbol{\kappa} \mid \boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z} \\ &\propto p(\boldsymbol{y} \mid \boldsymbol{\kappa}) \int p(\boldsymbol{\kappa} \mid \boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z}, \end{aligned} \tag{S-1.4}$$

where $p \propto q$ denotes that p is proportional to q, whereas,

$$p(\boldsymbol{y} \mid \boldsymbol{\kappa}) = \int p(\boldsymbol{y} \mid \boldsymbol{\kappa}, \sigma^2) p(\sigma^2 \mid \boldsymbol{\kappa}) d\sigma^2. \tag{S-1.5}$$

The term $p(\boldsymbol{y} \mid \boldsymbol{\kappa})$ is simply the average of the likelihood $p(\boldsymbol{y} \mid \boldsymbol{\kappa}, \sigma^2)$ over the conditional prior density $p(\sigma^2 \mid \boldsymbol{\kappa})$ of σ^2 given the values of the log-rate constants $\boldsymbol{\kappa}$. We refer to $p(\boldsymbol{y} \mid \boldsymbol{\kappa})$ as the "effective" likelihood. Moreover, we refer to $\int p(\boldsymbol{\kappa} \mid \boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z}$ as the "effective" prior, since this term contains the prior information about the log-rate constants $\boldsymbol{\kappa}$ after the prior information about the log-equilibrium constants is integrated out of the problem.

[1] It is clear that $\boldsymbol{z}$ can be directly calculated from $\boldsymbol{\kappa}$. Consequently, conditioning on both $\boldsymbol{\kappa}$ and $\boldsymbol{z}$ is equivalent to conditioning only on $\boldsymbol{\kappa}$. Therefore, $p(\sigma^2 \mid \boldsymbol{\kappa}, \boldsymbol{z}, \boldsymbol{y}) = p(\sigma^2 \mid \boldsymbol{\kappa}, \boldsymbol{y})$ and $p(\boldsymbol{y} \mid \boldsymbol{\kappa}, \boldsymbol{z}) = p(\boldsymbol{y} \mid \boldsymbol{\kappa})$.

Wegscheider conditions

As we mentioned in the Main text, the thermodynamic constraints given by Equation (10) are equivalent to the Wegscheider conditions given by Equation (11). To show this fact, let us first assume that Equation (10) is satisfied. Then, for every $\boldsymbol{r} = \{r_m, m \in \mathcal{M}\} \in \text{null}(\mathbb{S})$,

$$\begin{aligned}
\ln \prod_{m \in \mathcal{M}} \left(\frac{k_{2m-1}}{k_{2m}} \right)^{r_m} &= \sum_{m \in \mathcal{M}} r_m \ln \frac{k_{2m-1}}{k_{2m}} \\
&= \sum_{m \in \mathcal{M}} r_m \ln \frac{\prod_{n \in \mathcal{N}} [\overline{x}_n^{(p)}]^{\nu'_{nm}}}{\prod_{n \in \mathcal{N}} [\overline{x}_n^{(p)}]^{\nu_{nm}}} \\
&= \sum_{m \in \mathcal{M}} r_m \ln \prod_{n \in \mathcal{N}} [\overline{x}_n^{(p)}]^{s_{nm}} \\
&= \sum_{n \in \mathcal{N}} \left(\sum_{m \in \mathcal{M}} s_{nm} r_m \right) \ln \overline{x}_n^{(p)} \\
&= 0,
\end{aligned}$$

where $\{\overline{x}_n^{(p)}, n \in \mathcal{N}\}$ are the stationary concentrations in the system when the initial concentration of the p^{th} molecular species is perturbed. This shows Equation (11).

On the other hand, if Equation (11) is satisfied, then

$$\begin{aligned}
\sum_{m \in \mathcal{M}} r_m \ln \frac{k_{2m-1}}{k_{2m}} \frac{\prod_{n \in \mathcal{N}} [\overline{x}_n^{(p)}]^{\nu_{nm}}}{\prod_{n \in \mathcal{N}} [\overline{x}_n^{(p)}]^{\nu'_{nm}}} &= \sum_{m \in \mathcal{M}} \ln \left(\frac{k_{2m-1}}{k_{2m}} \right)^{r_m} - \sum_{m \in \mathcal{M}} r_m \ln \prod_{n \in \mathcal{N}} [\overline{x}_n^{(p)}]^{s_{nm}} \\
&= \sum_{m \in \mathcal{M}} \ln \left(\frac{k_{2m-1}}{k_{2m}} \right)^{r_m} - \sum_{n \in \mathcal{N}} \left(\sum_{m \in \mathcal{M}} s_{nm} r_m \right) \ln \overline{x}_n^{(p)} \\
&= \sum_{m \in \mathcal{M}} \ln \left(\frac{k_{2m-1}}{k_{2m}} \right)^{r_m} = 0, \qquad \text{(S-1.6)}
\end{aligned}$$

for every $\boldsymbol{r} \in \text{null}(\mathbb{S})$. We can take the m^{th} element r_m of $\boldsymbol{r}$ to be

$$r_m = k_{2m-1} \prod_{n \in \mathcal{N}} [\overline{x}_n^{(p)}]^{\nu_{nm}} - k_{2m} \prod_{n \in \mathcal{N}} [\overline{x}_n^{(p)}]^{\nu'_{nm}},$$

since, from Equations (2) and (4) in the Main text, we have $\mathbb{S}\boldsymbol{r} = 0$, which implies that $\boldsymbol{r} \in \text{null}(\mathbb{S})$. Note now that

$$\sum_{m \in \mathcal{M}} \left(k_{2m-1} \prod_{n \in \mathcal{N}} [\overline{x}_n^{(p)}]^{\nu_{nm}} - k_{2m} \prod_{n \in \mathcal{N}} [\overline{x}_n^{(p)}]^{\nu'_{nm}} \right) \ln \frac{k_{2m-1}}{k_{2m}} \frac{\prod_{n \in \mathcal{N}} [\overline{x}_n^{(p)}]^{\nu'_{nm}}}{\prod_{n \in \mathcal{N}} [\overline{x}_n^{(p)}]^{\nu_{nm}}} = 0,$$

by virtue of (S-1.6). However, each term in the previous sum is nonnegative. Hence, for the sum to be equal to zero, each term must be zero, which implies the steady-state thermodynamic constrains given by Equation (10) in the Main text.

We will now show that the Wegscheider conditions are satisfied for all $\boldsymbol{r} \in \text{null}(\mathbb{S})$ so long as they are satisfied for $M_2 = M - M_1$ basis vectors $\{\boldsymbol{r}(j), j = 1, 2, \ldots, M_1\}$ of the null space of $\mathbb{S}$, where $M_1 = \text{rank}(\mathbb{S})$. Note that, for any $\boldsymbol{r} \in \text{null}(\mathbb{S})$,

$$\boldsymbol{r} = \sum_{j=1}^{M_2} a_j \boldsymbol{r}(j),$$

for some scalar coefficients $\alpha_j, j = 1, 2, \ldots, M_2$. As a consequence, and from Equations (8) and (11) in the Main text, we have that

$$\begin{aligned}
\ln \prod_{m \in \mathcal{M}} \left(\frac{k_{2m-1}}{k_{2m}} \right)^{r_m} &= \sum_{m \in \mathcal{M}} r_m \ln \frac{k_{2m-1}}{k_{2m}} \\
&= \sum_{m \in \mathcal{M}} r_m z_m \\
&= \sum_{m \in \mathcal{M}} \left[\sum_{j=1}^{M_2} a_j r_m(j) \right] z_m \\
&= \sum_{j=1}^{M_2} a_j \sum_{m \in \mathcal{M}} z_m r_m(j) \\
&= \sum_{j=1}^{M_2} a_j \sum_{m \in \mathcal{M}} r_m(j) \ln \frac{k_{2m-1}}{k_{2m}} \\
&= \sum_{j=1}^{M_2} a_j \ln \prod_{m \in \mathcal{M}} \left(\frac{k_{2m-1}}{k_{2m}} \right)^{r_m(j)} \\
&= 0,
\end{aligned}$$

for every $\boldsymbol{r} = \{r_m, m \in \mathcal{M}\} \in \text{null}(\mathbb{S})$, since the Wegscheider conditions are assumed to be satisfied by the basis vectors of the null space of $\mathbb{S}$. This shows that the Wegscheider conditions are also satisfied for every $\boldsymbol{r} \in \text{null}(\mathbb{S})$.

Let us now rearrange the columns and rows of the stoichiometry matrix $\mathbb{S}$ (by appropriately relabeling the molecular species and reactions) so that the first M_1 columns are linearly independent, whereas, the remaining M_2 columns are linearly dependent on the first columns. In this case, we can write the stoichiometry matrix $\mathbb{S}$ in the following block matrix form:

$$\mathbb{S} = \begin{bmatrix} \mathbb{S}_{11} & \mathbb{S}_{12} \\ \mathbb{S}_{21} & \mathbb{S}_{22} \end{bmatrix},$$

where $\mathbb{S}_{11}$ is an *invertible* $M_1 \times M_1$ matrix, whereas, $\mathbb{S}_{12}$, $\mathbb{S}_{21}$, and $\mathbb{S}_{22}$ are $M_1 \times M_2$, $(N - M_1) \times M_1$, and $(N - M_1) \times M_2$ matrices, respectively. It is a well-known fact (e.g., see [1]) that the general solution of $\mathbb{S}\boldsymbol{r} = \boldsymbol{0}$ is given by $\boldsymbol{r}' = -\mathbb{S}_{11}^{-1}\mathbb{S}_{12}\boldsymbol{r}''$, for an arbitrary $\boldsymbol{r}''$, where $\boldsymbol{r}'$, $\boldsymbol{r}''$ are $M_1 \times 1$ and $M_2 \times 1$ vectors, respectively, such that

$$\boldsymbol{r} = \begin{bmatrix} \boldsymbol{r}' \\ \boldsymbol{r}'' \end{bmatrix}.$$

This implies that the columns of matrix

$$\mathbb{B} = \begin{bmatrix} -\mathbb{S}_{11}^{-1}\mathbb{S}_{12} \\ \\ \mathbb{I}_{M_2} \end{bmatrix},$$

where $\mathbb{I}_{M_2}$ is the $M_2 \times M_2$ identity matrix, form a basis for the null space of $\mathbb{S}$. As a consequence of this result and the fact that the Wegscheider conditions are satisfied so long they are satisfied by a basis vector of null($\mathbb{S}$), we can conclude that the Wegscheider conditions, given by Equation (11) in the Main text, are equivalent to the following conditions [2]:

$$\kappa_{2m'} \;=\; \kappa_{2m'-1} + \sum_{m \in \mathcal{M}_1} [\mathbb{S}_{11}^{-1}\mathbb{S}_{12}]_{m,m'} \left(\kappa_{2m} - \kappa_{2m-1}\right), \quad \text{for every } \; m' \in \mathcal{M}_2, \tag{S-1.7}$$

where $\mathcal{M}_1 = \{1, 2, \ldots, M_1\}$, $\mathcal{M}_2 = \{M_1 + 1, M_1 + 2, \ldots, M\}$, and $[\mathbb{S}_{11}^{-1}\mathbb{S}_{12}]_{m,m'}$ is the element of the m^{th} row and the m'^{th} column of matrix $\mathbb{S}_{11}^{-1}\mathbb{S}_{12}$.

Equation (S-1.7) allows us to specify arbitrary values for the forward and reverse log-rate constants $\{\kappa_{2m-1}, m \in \mathcal{M}\}$, $\{\kappa_{2m}, m \in \mathcal{M}_1\}$, and calculate the reverse log-rate constants $\{\kappa_{2m}, m \in \mathcal{M}_2\}$ so that the Wegscheider conditions are satisfied. If we denote by $\boldsymbol{\kappa}_f$ the $(M + M_1)$ "free" log-rate constants $\{\kappa_{2m-1}, m \in \mathcal{M}, \kappa_{2m}, m \in \mathcal{M}_1\}$ and by $\boldsymbol{\kappa}_d$ the M_2 "dependent" log-rate constants $\{\kappa_{2m}, m \in \mathcal{M}_2\}$, then we can write (S-1.7) in the following compact form:

$$\boldsymbol{\kappa}_d = \mathbb{W}\boldsymbol{\kappa}_f, \tag{S-1.8}$$

where $\mathbb{W}$ is the $(M - M_1) \times (M + M_1)$ matrix implementing the right hand side of (S-1.7), given by

$$\mathbb{W} = \left[-(\mathbb{S}_{11}^{-1}\mathbb{S}_{12})^T \mid \mathbb{I}_{M_2} \mid (\mathbb{S}_{11}^{-1}\mathbb{S}_{12})^T\right]. \tag{S-1.9}$$

Thus, for any arbitrary $\boldsymbol{\kappa}_f$, we can construct a thermodynamically feasible set of kinetic parameters by determining the dependent parameters $\boldsymbol{\kappa}_d$ according to (S-1.8), and by setting $\boldsymbol{\kappa} = \{\boldsymbol{\kappa}_f, \boldsymbol{\kappa}_d\}$.

Prior density of log-equilibrium constants

From Equations (8) and (10) in the Main text, we have that

$$z_m = \sum_{n \in \mathcal{N}} s_{nm} \ln \overline{x}_n^{(p)}, \quad \text{for all } \; p \in \mathcal{P},$$

where $s_{nm} = \nu'_{nm} - \nu_{nm}$ is the net stoichiometry coefficient and $\{\overline{x}_n^{(p)}, n \in \mathcal{N}\}$ are the stationary concentrations when the initial concentration of the p^{th} molecular species is perturbed. Therefore,

$$z_m = \frac{1}{P+1} \sum_{p \in \mathcal{P}} \sum_{n \in \mathcal{N}} s_{nm} \ln \overline{x}_n^{(p)}, \quad \text{for all } \; m \in \mathcal{M}. \tag{S-1.10}$$

On the other hand, Equation (5) in the Main text gives

$$\sum_{p \in \mathcal{P}} \sum_{n \in \mathcal{N}} s_{nm} \ln x_n^{(p)}(t_q) = \sum_{p \in \mathcal{P}} \sum_{n \in \mathcal{N}} s_{nm} y_n^{(p)}(t_q) - \sum_{p \in \mathcal{P}} \sum_{n \in \mathcal{N}} s_{nm} \eta_n^{(p)}(t_q). \tag{S-1.11}$$

If we assume that the biochemical reaction system and all its perturbed versions are sufficiently close to steady-state at some time point t_{Q+1}, then (S-1.10) and (S-1.11) approximately imply that

$$z_m = \widetilde{y}_m - \widetilde{\eta}_m, \tag{S-1.12}$$

where

$$\widetilde{y}_m := \frac{1}{P+1} \sum_{p \in \mathcal{P}} \sum_{n \in \mathcal{N}} s_{nm} y_n^{(p)}(t_{Q+1}), \tag{S-1.13}$$

and

$$\widetilde{\eta}_m := \frac{1}{P+1} \sum_{p \in \mathcal{P}} \sum_{n \in \mathcal{N}} s_{nm} \eta_n^{(p)}(t_{Q+1}).$$

To proceed, note that

$$\begin{aligned} p(\boldsymbol{z} \mid \overline{\boldsymbol{y}}) &= p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}}, \overline{\boldsymbol{y}}) \\ &= \int p(\boldsymbol{z}, \sigma^2 \mid \widetilde{\boldsymbol{y}}, \overline{\boldsymbol{y}}) d\sigma^2 \\ &= \int p(\boldsymbol{z} \mid \sigma^2, \widetilde{\boldsymbol{y}}, \overline{\boldsymbol{y}}) p(\sigma^2 \mid \widetilde{\boldsymbol{y}}, \overline{\boldsymbol{y}}) d\sigma^2 \\ &= \int p(\boldsymbol{z} \mid \sigma^2, \widetilde{\boldsymbol{y}}) p(\sigma^2 \mid \overline{\boldsymbol{y}}) d\sigma^2. \end{aligned} \tag{S-1.14}$$

This is due to the fact that $\widetilde{\boldsymbol{y}} := \{\widetilde{y}_m, m \in \mathcal{M}\}$ can be calculated from $\overline{\boldsymbol{y}} := \{y_n^{(p)}(t_{Q+1}), n \in \mathcal{N}, p \in \mathcal{P}\}$ by means of (S-1.13), in which case $p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}}, \overline{\boldsymbol{y}}) = p(\boldsymbol{z} \mid \overline{\boldsymbol{y}})$ and $p(\sigma^2 \mid \widetilde{\boldsymbol{y}}, \overline{\boldsymbol{y}}) = p(\sigma^2 \mid \overline{\boldsymbol{y}})$. Moreover, since

$z_m = \widetilde{y}_m - \widetilde{\eta}_m$, $\overline{\boldsymbol{y}}$ does not provide further information about $\boldsymbol{z}$, given σ^2 and $\widetilde{\boldsymbol{y}}$, in which case $p(\boldsymbol{z} \mid \sigma^2, \widetilde{\boldsymbol{y}}, \overline{\boldsymbol{y}}) = p(\boldsymbol{z} \mid \sigma^2, \widetilde{\boldsymbol{y}})$.

Equation (S-1.14) implies that, in order to calculate the probability density function $p(\boldsymbol{z} \mid \overline{\boldsymbol{y}})$, we must determine the probability density functions $p(\boldsymbol{z} \mid \sigma^2, \widetilde{\boldsymbol{y}})$ and $p(\sigma^2 \mid \overline{\boldsymbol{y}})$. Note that, given σ^2 and $\widetilde{\boldsymbol{y}}$, the log-equilibrium constants $\boldsymbol{z}$ follow a multivariate Gaussian distribution with means and covariances

$$\mathrm{E}[z_m \mid \sigma^2, \widetilde{\boldsymbol{y}}] = \widetilde{y}_m \qquad \text{and} \qquad \mathrm{cov}[z_m, z_{m'} \mid \sigma^2, \widetilde{\boldsymbol{y}}] = \frac{\sigma^2}{P+1} \sum_{n \in \mathcal{N}} s_{nm} s_{nm'} \,.$$

This implies that

$$p(\boldsymbol{z} \mid \sigma^2, \widetilde{\boldsymbol{y}}) = \frac{(P+1)^{M/2}}{(2\pi)^{M/2} \sigma^M |\mathbb{H}|^{1/2}} \exp\left\{ -\frac{P+1}{2\sigma^2} (\boldsymbol{z} - \widetilde{\boldsymbol{y}})^T \mathbb{H}^{-1} (\boldsymbol{z} - \widetilde{\boldsymbol{y}}) \right\}, \tag{S-1.15}$$

where $\mathbb{H}$ is an $M \times M$ matrix with elements $h_{mm'} = \sum_{n \in \mathcal{N}} s_{nm} s_{nm'}$. Note that $\mathbb{H} = \mathbb{S}^T \mathbb{S}$, where $\mathbb{S}$ is the $N \times M$ stoichiometry matrix of the biochemical reaction system with elements s_{nm}.

To determine the probability density function $p(\sigma^2 \mid \overline{\boldsymbol{y}})$ of the error variance, note that

$$p(\overline{\boldsymbol{x}}, \sigma^2 \mid \overline{\boldsymbol{y}}) = \frac{p(\overline{\boldsymbol{y}} \mid \overline{\boldsymbol{x}}, \sigma^2) p(\overline{\boldsymbol{x}}, \sigma^2)}{p(\overline{\boldsymbol{y}})}, \tag{S-1.16}$$

where $\overline{\boldsymbol{x}} := \{\ln \overline{x}_n^{(p)}, n \in \mathcal{N}, p \in \mathcal{P}\}$ are the stationary concentrations when the initial concentration of the p^{th} molecular species is perturbed. Moreover,

$$p(\overline{\boldsymbol{x}}, \sigma^2 \mid \overline{\boldsymbol{y}}) = p(\overline{\boldsymbol{x}} \mid \sigma^2, \overline{\boldsymbol{y}}) p(\sigma^2 \mid \overline{\boldsymbol{y}}). \tag{S-1.17}$$

As a consequence of (S-1.16) and (S-1.17), we have

$$\begin{aligned}
p(\sigma^2 \mid \overline{\boldsymbol{y}}) &= \frac{p(\overline{\boldsymbol{x}}, \sigma^2 \mid \overline{\boldsymbol{y}})}{p(\overline{\boldsymbol{x}} \mid \sigma^2, \overline{\boldsymbol{y}})} \\
&= \frac{p(\overline{\boldsymbol{y}} \mid \overline{\boldsymbol{x}}, \sigma^2) p(\overline{\boldsymbol{x}}, \sigma^2)}{p(\overline{\boldsymbol{x}} \mid \sigma^2, \overline{\boldsymbol{y}}) p(\overline{\boldsymbol{y}})} \\
&= \frac{p(\overline{\boldsymbol{y}} \mid \overline{\boldsymbol{x}}, \sigma^2) p(\overline{\boldsymbol{x}}) p(\sigma^2)}{p(\overline{\boldsymbol{x}} \mid \sigma^2, \overline{\boldsymbol{y}}) p(\overline{\boldsymbol{y}})} \\
&\propto \frac{p(\overline{\boldsymbol{y}} \mid \overline{\boldsymbol{x}}, \sigma^2)}{p(\overline{\boldsymbol{x}} \mid \sigma^2, \overline{\boldsymbol{y}})} \, p(\sigma^2) \,,
\end{aligned}$$

where we use the fact that $\overline{\boldsymbol{x}}$ and σ^2 are statistically independent (since $\overline{\boldsymbol{x}}$ is determined from $\boldsymbol{\kappa}$ and we have assumed in the Main text that $\boldsymbol{\kappa}$ and σ^2 are statistically independent). It is now not difficult to see that

$p(\overline{\boldsymbol{y}} \mid \overline{\boldsymbol{x}}, \sigma^2) = p(\overline{\boldsymbol{x}} \mid \sigma^2, \overline{\boldsymbol{y}})$, due to the statistical independence and Gaussianity of the error terms $\eta_n^{(p)}$ in Equation (5) of the Main text. Therefore,

$$p(\sigma^2 \mid \overline{\boldsymbol{y}}) = p(\sigma^2) = \frac{b^a}{\Gamma(a)} \, (\sigma^2)^{-(a+1)} \, e^{-b/\sigma^2}, \tag{S-1.18}$$

which does not depend on $\overline{\boldsymbol{y}}$, by virtue of Equation (9) in the Main text.

As a consequence of (S-1.14) and (S-1.18), we now have that

$$p(\boldsymbol{z} \mid \overline{\boldsymbol{y}}) = p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}}, \overline{\boldsymbol{y}}) = p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}}). \tag{S-1.19}$$

Finally, from (S-1.14), (S-1.15), (S-1.18), and (S-1.19), and after some straightforward, albeit tedious, algebraic manipulations, we can show that

$$\begin{aligned} p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}}) &\propto \int \frac{1}{\sigma^{2(a+1)+M}} \exp\left\{ -\frac{P+1}{2\sigma^2} (\boldsymbol{z} - \widetilde{\boldsymbol{y}})^T \mathbb{H}^{-1} (\boldsymbol{z} - \widetilde{\boldsymbol{y}}) - \frac{b}{\sigma^2} \right\} d\sigma^2 \\ &\propto \left[\frac{2b}{P+1} + (\boldsymbol{z} - \widetilde{\boldsymbol{y}})^T \mathbb{H}^{-1} (\boldsymbol{z} - \widetilde{\boldsymbol{y}}) \right]^{-(M/2+a)}. \end{aligned} \tag{S-1.20}$$

A problem associated with the previous formulation is that it may not be possible to evaluate the prior density $p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}})$ of the log-equilibrium constants, given by (S-1.20), since the matrix $\mathbb{H}$ may not be invertible. Indeed, if $\boldsymbol{r}$ is a (nonzero) vector in the null space of the $N \times M$ stoichiometry matrix $\mathbb{S}$ (i.e., if $\mathbb{S}\,\boldsymbol{r} = 0$), then $\boldsymbol{r}^T \mathbb{H}\,\boldsymbol{r} = \boldsymbol{r}^T \mathbb{S}^T \mathbb{S}\,\boldsymbol{r} = 0$, which shows that matrix $\mathbb{H}$ is positive semi-definite and, hence, not necessarily invertible.

To address the previous problem, we will follow a well-known technique known as *decorrelation* or *whitening* that allows us to transform the dependent random variables $\boldsymbol{z}$ into the statistically independent zero-mean random variables $\boldsymbol{z}_0$ and obtain a form of matrix $\mathbb{H}$ that is always invertible. Subsequently, this will allow us to derive a form for the probability density function $p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}})$ that we can always evaluate.

Let us consider an $(M - M_0)$-dimensional zero-mean Gaussian random vector $\boldsymbol{z}_0$ with identity covariance matrix, where M_0 in the number of zero eigenvalues of $\mathbb{H}$. Our first objective is to determine an $M \times (M - M_0)$ matrix $\mathbb{H}_0$, such that

$$\boldsymbol{z} = \mathbb{H}_0 \boldsymbol{z}_0 + \widetilde{\boldsymbol{y}}\,. \tag{S-1.21}$$

Note that $\mathrm{E}[\boldsymbol{z} \mid \sigma^2, \widetilde{\boldsymbol{y}}] = \widetilde{\boldsymbol{y}}$, as expected from (S-1.12), whereas, $\mathbb{H} = [(P+1)/\sigma^2]\, \mathbb{H}_0\, \mathbb{H}_0^T$. We can use singular value decomposition (SVD) [3] to decompose matrix $\mathbb{H} = \mathbb{S}^T\mathbb{S}$ into $\mathbb{H} = \mathbb{U}\mathbb{D}\mathbb{U}^T$, form the

$(M - M_0) \times (M - M_0)$ diagonal matrix $\mathbb{D}_0$ by removing the last M_0 zero singular values from $\mathbb{D}$ and the $M \times (M - M_0)$ submatrix $\mathbb{U}_0$ of $\mathbb{U}$ by removing the last M_0 columns of $\mathbb{U}$. Then, $\mathbb{H} = \mathbb{U}_0\mathbb{D}_0\mathbb{U}_0^T$, which implies $\mathbb{H}_0 = \sigma\, \mathbb{U}_0\mathbb{D}_0^{1/2}/\sqrt{P+1}$. Note now that $\mathbb{U}_0^T\mathbb{U}_0 = \mathbb{I}$, where $\mathbb{I}$ denotes the identity matrix. As a consequence, given vectors $\boldsymbol{z}$ and $\widetilde{\boldsymbol{y}}$, (S-1.21) has a unique solution, given by

$$\boldsymbol{z}_0 = \frac{\sqrt{P+1}}{\sigma}\, \mathbb{D}_0^{-1/2}\mathbb{U}_0^T(\boldsymbol{z} - \widetilde{\boldsymbol{y}})\,. \tag{S-1.22}$$

This formula transforms the dependent random variables $\boldsymbol{z}$ into the statistically independent zero-mean random variables $\boldsymbol{z}_0$.

Note now that

$$\begin{aligned}
p(\boldsymbol{z} \mid \sigma^2, \widetilde{\boldsymbol{y}}) &= \int p(\boldsymbol{z}, \boldsymbol{z}_0 \mid \sigma^2, \widetilde{\boldsymbol{y}})\, d\boldsymbol{z}_0 \\
&= \int p(\boldsymbol{z} \mid \boldsymbol{z}_0, \sigma^2, \widetilde{\boldsymbol{y}}) p(\boldsymbol{z}_0)\, d\boldsymbol{z}_0 \\
&= \frac{1}{(2\pi)^{(M-M_0)/2}} \int \delta(\boldsymbol{z} - \mathbb{H}_0\boldsymbol{z}_0 - \widetilde{\boldsymbol{y}}) \exp\left\{-\frac{1}{2}\, \boldsymbol{z}_0^T\boldsymbol{z}_0\right\} d\boldsymbol{z}_0 \\
&= \frac{\int \delta(\boldsymbol{z} - \mathbb{H}_0\boldsymbol{z}_0 - \widetilde{\boldsymbol{y}}) d\boldsymbol{z}_0}{(2\pi)^{(M-M_0)/2}} \exp\left\{-\frac{P+1}{2\sigma^2}(\boldsymbol{z} - \widetilde{\boldsymbol{y}})^T\mathbb{U}_0\mathbb{D}_0^{-1}\mathbb{U}_0^T(\boldsymbol{z} - \widetilde{\boldsymbol{y}})\right\} \\
&= \frac{\int \delta(\mathbb{H}_0\boldsymbol{z}_0)\, d\boldsymbol{z}_0}{(2\pi)^{(M-M_0)/2}} \exp\left\{-\frac{P+1}{2\sigma^2}(\boldsymbol{z} - \widetilde{\boldsymbol{y}})^T\mathbb{U}_0\mathbb{D}_0^{-1}\mathbb{U}_0^T(\boldsymbol{z} - \widetilde{\boldsymbol{y}})\right\},
\end{aligned}$$

by virtue of (S-1.21) and (S-1.22), where $\delta(\cdot)$ is the Dirac delta function. This result shows that

$$p(\boldsymbol{z} \mid \sigma^2, \widetilde{\boldsymbol{y}}) \propto \exp\left\{-\frac{P+1}{2\sigma^2}(\boldsymbol{z} - \widetilde{\boldsymbol{y}})^T\mathbb{U}_0\mathbb{D}_0^{-1}\mathbb{U}_0^T(\boldsymbol{z} - \widetilde{\boldsymbol{y}})\right\},$$

which leads to [compare with (S-1.20)]:

$$\begin{aligned}
p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}}) &\propto \int \frac{1}{\sigma^{2(a+1)}} \exp\left\{-\frac{P+1}{2\sigma^2}(\boldsymbol{z} - \widetilde{\boldsymbol{y}})^T\mathbb{U}_0\mathbb{D}_0^{-1}\mathbb{U}_0^T(\boldsymbol{z} - \widetilde{\boldsymbol{y}}) - \frac{b}{\sigma^2}\right\} d\sigma^2 \\
&\propto \left[\frac{2b}{P+1} + (\boldsymbol{z} - \widetilde{\boldsymbol{y}})^T\mathbb{U}_0\mathbb{D}_0^{-1}\mathbb{U}_0^T(\boldsymbol{z} - \widetilde{\boldsymbol{y}})\right]^{-a}.
\end{aligned}$$

Note that we can always evaluate this form of $p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}})$, since the diagonal matrix $\mathbb{D}_0$ is invertible.

If we now replace $p(\boldsymbol{z})$ by $p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}})$ in Equation (6) of the Main text, we have

$$\begin{aligned}
\int p(\boldsymbol{\kappa} \mid \boldsymbol{y}) d\boldsymbol{\kappa} &= \frac{1}{p(\boldsymbol{y})} \int \int p(\boldsymbol{y} \mid \boldsymbol{\kappa}) p(\boldsymbol{\kappa} \mid \boldsymbol{z}) p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}}) d\boldsymbol{\kappa} d\boldsymbol{z} \\
&= \frac{1}{p(\boldsymbol{y})} \int \int p(\boldsymbol{y} \mid \boldsymbol{\kappa}, \boldsymbol{z}) p(\boldsymbol{\kappa} \mid \boldsymbol{z}) p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}}) d\boldsymbol{\kappa} d\boldsymbol{z} \\
&= \frac{1}{p(\boldsymbol{y})} \int p(\boldsymbol{y} \mid \boldsymbol{z}) p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}}) d\boldsymbol{z} \\
&= \frac{1}{p(\boldsymbol{y})} \int p(\boldsymbol{y} \mid \boldsymbol{z}, \widetilde{\boldsymbol{y}}) p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}}) d\boldsymbol{z} \\
&= 1, \text{ for all } \boldsymbol{y},
\end{aligned}$$

since $p(\boldsymbol{y} \mid \boldsymbol{\kappa}, \boldsymbol{z}) = p(\boldsymbol{y} \mid \boldsymbol{\kappa})$ (see footnote 1), provided that $\widetilde{\boldsymbol{y}}$ is statistically independent of $\boldsymbol{y}$, in which case $p(\boldsymbol{y} \mid \boldsymbol{z}, \widetilde{\boldsymbol{y}}) = p(\boldsymbol{y} \mid \boldsymbol{z})$ and $p(\boldsymbol{y} \mid \widetilde{\boldsymbol{y}}) = p(\boldsymbol{y})$. Clearly, the independence between $\widetilde{\boldsymbol{y}}$ and $\boldsymbol{y}$ is a sufficient condition for the posterior density $p(\boldsymbol{\kappa} \mid \boldsymbol{y})$ to be proper [i.e., for $p(\boldsymbol{\kappa} \mid \boldsymbol{y})$ to be finite for every $\boldsymbol{y}$].

Prior density of log-rate constants

Let us consider a bimolecular reaction $X_1 + X_2 \to X_3$. For this reaction to occur, two events must take place: one molecule of X_1 must collide with one molecule of X_2 and, given that the two molecules have collided, the reaction must take place. Using basic probabilistic arguments and the well-known hard-sphere collision theory [4], it has been shown in [5] that the probability of a randomly selected pair of molecules X_1 and X_2 at time t to react during an infinitesimal time interval $[t, t + dt)$ is given by cdt. Here, c is known as the specific probability rate constant and is given by

$$c = \frac{\pi(r_1 + r_2)^2}{V} \sqrt{\frac{8k_B T}{\pi \mu^*}} \, \gamma, \tag{S-1.23}$$

where V is the volume of the biochemical reaction system, T is the temperature, k_B is the Boltzmann constant ($k_B = 1.3806504 \times 10^{-23}$J/K), and γ is the probability that a randomly selected pair of colliding molecules X_1 and X_2 will react. This formula is based on the assumption that each molecule X_n is a hard sphere of radius r_n and mass μ_n. In (S-1.23), $\mu^* = \mu_1 \mu_2 / (\mu_1 + \mu_2)$.

The rate constant k of the previous bimolecular reaction is associated to the specific probability rate constant c by means of $k = AVc$ (for a reaction with different reactants; see [6]), where A is the Avogadro constant ($A = 6.0221415 \times 10^{23}$mol^{-1}). If we assume that a pair X_1 and X_2 of molecules react only after collision with impact energy greater than E, then $\gamma = \exp(-E/k_B T)$ [5]. In this case,

$$k = \alpha e^{-E/k_B T}, \tag{S-1.24}$$

where

$$\alpha := A\pi(r_1 + r_2)^2 \sqrt{\frac{8k_B T}{\pi \mu^*}}. \tag{S-1.25}$$

Equation (S-1.24) is the well-known Arrhenius formula for the rate constant of a bimolecular reaction, and holds for other types of reactions (e.g., monomolecular and trimolecular) as well. The coefficient α is usually referred to as the pre-exponential factor, or simply the prefactor, whereas, E is referred to as the activation energy of the reaction. In the following, we use (S-1.24) to derive a probabilistic model for the log-rate constants of a biochemical reaction system that leads to an appropriate prior density $p(\boldsymbol{\kappa} \mid \boldsymbol{z})$ for the parameters $\boldsymbol{\kappa}$.

According to (S-1.24), the rate constant of the m^{th} forward reaction is given by

$$k_{2m-1} = \alpha_m e^{-E_m/k_B T}. \tag{S-1.26}$$

Equation (S-1.25) provides a theoretical expression for the prefactor α_m, assuming the ideal situation of both reactant molecules being perfect hard spheres. In reality, however, the situation is much more complex, and we can use (S-1.25) to predict only a portion of the true prefactor value (provided that the masses and radiuses of the reactant molecules are known). As a consequence, we can decompose the prefactor α_m into a predictable part α_m^0 and an unpredictable part ω_m, so that $\alpha_m = \alpha_m^0 \omega_m$. This implies that $\ln \alpha_m = \ln \alpha_m^0 + g_m$, where $g_m := \ln \omega_m$ is a random additive component. The multiplicative factor ω_m can also be used to model unpredictable changes in biochemical conditions or changes in the structure of the reactant molecules, which may also affect the probability of particle collision and thus α_m. Therefore, we will be using the following expression for the prefactor of the m^{th} forward reaction:

$$\alpha_m = \alpha_m^0 e^{g_m}. \tag{S-1.27}$$

We will assume that g_m is a zero-mean Gaussian noise component with standard deviation λ_m. In this case, α_m is a random variable characterized by the log-normal distribution

$$p(\alpha_m) = \frac{1}{\alpha_m \lambda_m \sqrt{2\pi}} \exp\left\{ -\left(\frac{1}{\sqrt{2}\lambda_m} \ln \frac{\alpha_m}{\alpha_m^0} \right)^2 \right\},$$

with parameters $\ln \alpha_m^0$ and λ_m. It has been pointed out in [7] that log-normal distributions are very natural for modeling biochemical processes and are a direct consequence of the thermodynamic behavior of biochemical reaction systems.

Unpredictable changes in biochemical conditions can also affect the probability of reaction after collision, or equivalently, the activation energy E_m. We may consider E_m to be a random variable that is decomposed into two terms: a deterministic term E_m^0 and a random additive term U_m, so that

$$E_m = E_m^0 + U_m, \tag{S-1.28}$$

where $E_m^0, U_m \geq 0$. This is known as the (static) random activation energy (RAE) model [8]. A commonly used probability law for the random energy component U_m is the Maxwell-Boltzmann (exponential) distribution [8]:

$$p(U_m) = \frac{1}{k_B T_m^*} \exp\left\{ -\frac{U_m}{k_B T_m^*} \right\}, \quad U_m \geq 0, \tag{S-1.29}$$

at some temperature $T_m^* > T$.[2]

As a consequence of (S-1.26), (S-1.27), and (S-1.28), we have that

$$\kappa_{2m-1} = \kappa_m^0 + g_m - w_m, \quad w_m \geq 0, \tag{S-1.30}$$

where

$$\kappa_m^0 := \ln \alpha_m^0 - \frac{E_m^0}{k_B T} \qquad \text{and} \qquad w_m := \frac{U_m}{k_B T}\,.$$

Moreover, (S-1.29) implies that the probability density function of w_m is given by the following exponential distribution:

$$p(w_m) = \frac{1}{\tau_m}\, e^{-w_m/\tau_m}, \quad w_m \geq 0, \quad \tau_m > 0,$$

where $\tau_m := T_m^*/T > 1$. Note that the expected value of w_m equals its standard deviation, with $\mathrm{E}[w_m] = \mathrm{sd}[w_m] = \tau_m$. In the following, we will assume that, for all $m \in \mathcal{M}$, w_m is statistically independent of g_m. This is a reasonable assumption, considering the fact that these two random variables result from two different biophysical mechanisms, namely molecular collision and molecular reaction, which we may consider to be statistically independent.

[2] It has been suggested in the literature (e.g., in [8]) that T_m^* must be larger than the system temperature T.

The previous modeling steps lead to the following prior probability density function $p(\kappa_{2m-1})$ for the log-rate constant of the m^{th} forward reaction:

$$p(\kappa_{2m-1}) = \frac{e^{\lambda_m^2/2\tau_m^2}}{2\tau_m} \, \text{erfc}\left[\frac{1}{\sqrt{2}}\left(\frac{\lambda_m}{\tau_m} + \frac{\kappa_{2m-1} - \kappa_m^0}{\lambda_m}\right)\right] e^{(\kappa_{2m-1}-\kappa_m^0)/\tau_m}, \tag{S-1.31}$$

where erfc$[\cdot]$ is the complementary error function [9]. To derive this result, consider a random variable

$$y = c + g - w, \tag{S-1.32}$$

where c is a constant and w, g are two statistically independent random variables so that

$$p_w(w) = \frac{1}{\tau}\, e^{-w/\tau} \ \text{(exponential)} \qquad \text{and} \qquad p_g(g) = \frac{1}{\sqrt{2\pi}\lambda}\, e^{-g^2/2\lambda^2} \ \text{(Gaussian)}.$$

If we set $u = w - g$, then $y = c - u$ and

$$p_y(y) = p_u(c - y). \tag{S-1.33}$$

Moreover, since w and g are statistically independent, we have [9]

$$\begin{aligned}
p_u(u) &= \int_{-\infty}^{\infty} p_w(x) p_g(x-u) dx \\
&= \int_0^{\infty} \frac{1}{\tau}\, e^{-x/\tau} \frac{1}{\sqrt{2\pi}\lambda} e^{-(x-u)^2/2\lambda^2} dx \\
&= \frac{1}{\sqrt{2\pi}\lambda\tau} \int_0^{\infty} e^{-(\tau x^2 - 2\tau u x + \tau u^2 + 2\lambda^2 x)/2\lambda^2\tau} dx \\
&= \frac{1}{\sqrt{2\pi}\lambda\tau} \int_0^{\infty} e^{-(x+\lambda^2/\tau - u)^2/2\lambda^2} e^{[(\lambda^2/\tau-u)^2 - u^2]/2\lambda^2} dx \\
&= \frac{1}{\tau}\left[\frac{1}{\sqrt{2\pi}\lambda}\int_{\lambda^2/\tau-u}^{\infty} e^{-\xi^2/2\lambda^2} d\xi\right] e^{[(\lambda^2/\tau-u)^2-u^2]/2\lambda^2} \ \ \text{(by setting } \xi = x + \lambda^2/\tau - u\text{)} \\
&= \frac{e^{\lambda^2/2\tau^2}}{2\tau}\, \text{erfc}\left[\frac{1}{\sqrt{2}}\left(\frac{\lambda}{\tau} - \frac{u}{\lambda}\right)\right] e^{-u/\tau}.
\end{aligned} \tag{S-1.34}$$

Finally, by combining (S-1.33) and (S-1.34), we obtain

$$p_y(y) = \frac{e^{\lambda^2/2\tau^2}}{2\tau}\, \text{erfc}\left[\frac{1}{\sqrt{2}}\left(\frac{\lambda}{\tau} + \frac{y-c}{\lambda}\right)\right] e^{(y-c)/\tau}, \tag{S-1.35}$$

which provides an analytical expression for the probability density function of y. Equation (S-1.31) is now a direct consequence of (S-1.30), (S-1.32), and (S-1.35).

Let us now focus our attention on the log-rate constant κ_{2m} of the m^{th} reverse reaction. Basic thermodynamic arguments imply that

$$k_{2m-1} = k_{2m} \prod_{n \in \mathcal{N}} \phi_n^{s_{nm}}, \tag{S-1.36}$$

where k_{2m-1} and k_{2m} are the rate constants of the m^{th} forward and reverse reactions, respectively, ϕ_n is the capacity of the n^{th} molecular species, and s_{nm} is the stoichiometry coefficients of the n^{th} molecular species associated with the m^{th} reaction. The capacity is a thermodynamic quantity characteristic to a molecular species and depends on the standard chemical potential of that species (see [10] for details). As a consequence, the log-equilibrium constant z_m of the m^{th} reaction depends only on the stoichiometry of the biochemical reaction system and the capacities of the underlying molecular species, since Equation (8) in the Main text and (S-1.36) imply that

$$z_m = \sum_{n \in \mathcal{N}} s_{nm} \ln \phi_n, \quad \text{for all } \; m \in \mathcal{M}.$$

Therefore, z_m is a constant characteristic to the m^{th} reaction.

From Equation (8) in the Main text, note that

$$\kappa_{2m} = \kappa_{2m-1} - z_m, \quad \text{for all } \; m \in \mathcal{M},$$

which implies that κ_{2m} and κ_{2m-1} are two dependent random variables, given z_m. Their joint probability density function satisfies

$$\begin{aligned} p(\kappa_{2m}, \kappa_{2m-1} \mid z_m) \; &= \; p(\kappa_{2m} \mid z_m, \kappa_{2m-1}) p(\kappa_{2m-1} \mid z_m) \\ &= \; \delta(\kappa_{2m} - \kappa_{2m-1} + z_m) p(\kappa_{2m-1}), \end{aligned} \tag{S-1.37}$$

where $\delta(\cdot)$ is the Dirac delta function. Clearly, z_m is a hyperparameter, since it characterizes the prior joint density $p(\kappa_{2m}, \kappa_{2m-1})$ of the log-rate constants of the m^{th} reaction. We will assume here that, given $\boldsymbol{z} = \{z_1, z_2, \ldots, z_M\}$, the reaction rate constants of different reactions are mutually independent. Then, the prior density of the log-rate constants will be given by

$$p(\boldsymbol{\kappa} \mid \boldsymbol{z}) = \prod_{m \in \mathcal{M}} p(\kappa_{2m}, \kappa_{2m-1} \mid z_m) = \prod_{m \in \mathcal{M}} \delta(\kappa_{2m} - \kappa_{2m-1} + z_m) p(\kappa_{2m-1}),$$

by virtue of (S-1.37).

A practical method for determining the hyperparameters $\boldsymbol{\phi}, a, b$

In practice, prior information about the rate constants of a biochemical reaction system may be available from which we may be able to estimate their minimum, maximum, and average values. Moreover, some prior information might be available about the error processes $\eta_n^{(p)}$ in Equation (5) of the Main text, which may allow us to estimate the average value and spread of their variance. In this subsection, we show how to use these values to determine the hyperparameters $\boldsymbol{\phi} = \{\kappa_m^0, \tau_m, \lambda_m, m \in \mathcal{M}\}$ associated with the prior densities $p(\kappa_{2m-1})$ of the forward log-rate constants and the hyperparameters a, b associated with the prior density $p(\sigma^2)$ of the measurement errors.

From (S-1.30), we have that

$$\mathrm{E}[\kappa_{2m-1}] = \kappa_m^0 - \tau_m \qquad \text{and} \qquad \mathrm{sd}[\kappa_{2m-1}] = \sqrt{\lambda_m^2 + \tau_m^2}\,, \tag{S-1.38}$$

by virtue of our assumption that g_m and w_m are statistically independent. Clearly, the parameter κ_m^0 controls the location of $p(\kappa_{2m-1})$, whereas, τ_m controls both location and scale. Moreover, the parameter λ_m controls the scale of $p(\kappa_{2m-1})$, without affecting its location. We illustrate this behavior in Figure S-1.1.

Let κ^{min}, κ^{max} and κ^{avg} be the minimum, maximum, and average values of a forward log-rate constant κ. Our objective is to determine the hyperparameters of the prior density $p(\kappa)$, given by (S-1.31), so that $\mathrm{E}[\kappa] = \kappa^{\text{avg}}$ and $p(\kappa) \simeq 0$, for $\kappa \leq \kappa^{\text{min}}$, $\kappa \geq \kappa^{\text{max}}$.

Since $\mathrm{E}[\kappa] = \kappa^{\text{avg}}$, we must have

$$\kappa^0 - \tau = \kappa^{\text{avg}}, \tag{S-1.39}$$

by virtue of (S-1.38). Moreover, since we want $p(\kappa) \simeq 0$, for $\kappa \leq \kappa^{\text{min}}$, we can impose the condition

$$e^{(\kappa-\kappa^0)/\tau} \leq \tau\epsilon\, e^{-\lambda^2/2}, \quad \text{for} \ \kappa \leq \kappa^{\text{min}},$$

for a sufficiently small $\epsilon > 0$, which implies that $p(\kappa) < \epsilon$, for $\kappa \leq \kappa^{\text{min}}$, by virtue of (S-1.31) and the facts that $\tau > 1$ and $\mathrm{erfc}(x) \leq 2$. Since $e^{(\kappa-\kappa^0)/\tau}$ is monotonically increasing in κ, it suffices to set $e^{(\kappa^{\text{min}}-\kappa^0)/\tau} = \tau\epsilon^*$, where

$$\epsilon^* := \epsilon\, e^{-\lambda^2/2}, \tag{S-1.40}$$

which, together with (S-1.39), implies that

$$\tau \ln \tau + (1 + \ln \epsilon^*)\tau = \kappa^{\text{min}} - \kappa^{\text{avg}}. \tag{S-1.41}$$

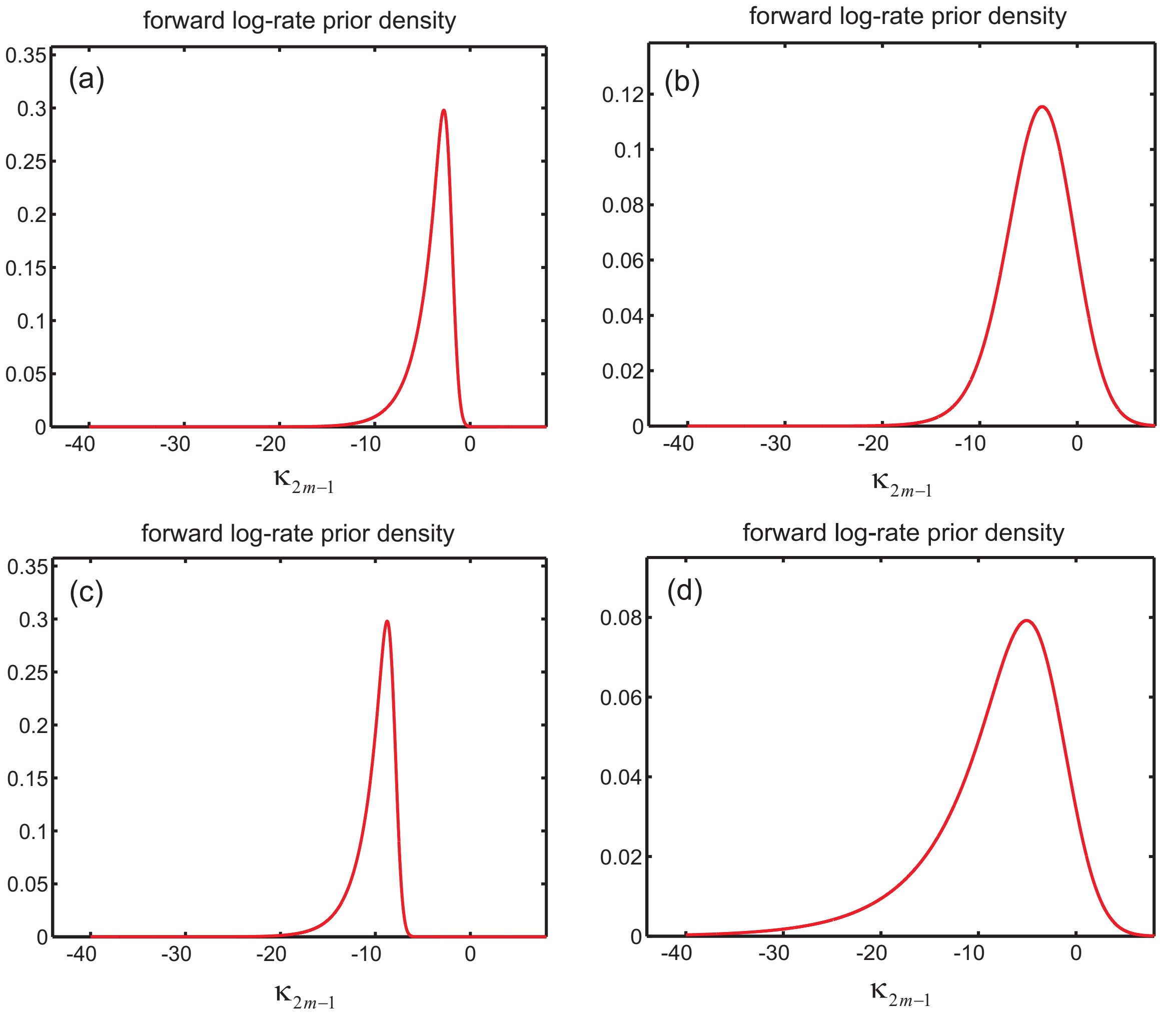

Figure S-1.1: *The prior density $p(\kappa_{2m-1})$ of the forward log-rate constant, given by (S-1.31), when: (a) $\kappa_m^0 = -2$, $\tau_m = 2$, $\lambda_m = 0.6$, (b) $\kappa_m^0 = -2$, $\tau_m = 2$, $\lambda_m = 3$, (c) $\kappa_m^0 = -8$, $\tau_m = 2$, $\lambda_m = 0.6$, and (d) $\kappa_m^0 = -2$, $\tau_m = 6$, $\lambda_m = 3$. Note that κ_m^0 controls the location of $p(\kappa_{2m-1})$, whereas, τ_m controls both location and scale. Moreover, the parameter λ_m controls the scale of $p(\kappa_{2m-1})$, without affecting its location.*

Finally, since we want $p(\kappa) \simeq 0$, for $\kappa \geq \kappa^{\max}$, we can impose the condition

$$\operatorname{erfc}\left[\frac{1}{\sqrt{2}}\left(\frac{\lambda}{\tau} + \frac{\kappa - \kappa^0}{\lambda}\right)\right] \leq 2\tau\epsilon\, e^{-\lambda^2/2}\, e^{-(\kappa-\kappa^0)/\tau}, \quad \text{for } \kappa \geq \kappa^{\max},$$

which implies that $p(\kappa) \leq \epsilon$, for $\kappa \geq \kappa^{\max}$, by virtue of (S-1.31) and the fact that $\tau > 1$. Since $\operatorname{erfc}[(\frac{\lambda}{\tau} + \frac{\kappa-\kappa^0}{\lambda})/\sqrt{2}]$ is monotonically decreasing in κ, it suffices to set

$$\operatorname{erfc}\left[\frac{1}{\sqrt{2}}\left(\frac{\lambda}{\tau} + \frac{\kappa^{\max} - \kappa^0}{\lambda}\right)\right] = 2\tau\epsilon^* e^{-(\kappa^{\max}-\kappa^0)/\tau},$$

which, together with (S-1.39), implies

$$\tau + \lambda \left(r\sqrt{2} - \frac{\lambda}{\tau} \right) = \kappa^{\max} - \kappa^{\text{avg}}, \tag{S-1.42}$$

where

$$r := \text{erfc}^{-1} \left[2\tau\epsilon^* e^{-(\kappa^{\max} - \kappa^{\text{avg}} - \tau)/\tau} \right].$$

Given $\kappa^{\min}$, $\kappa^{\max}$, and κ^{avg}, we may be able to determine the values of the hyperparameters κ^0, τ, and λ by simultaneously solving (S-1.39), (S-1.41), and (S-1.42). Unfortunately, (S-1.41) and (S-1.42) are nonlinear and they both depend on τ and λ. Hence, finding a solution to these equations is a rather difficult problem [3]. In the following, we discuss a simple approach for determining the values of the hyperparameters, which works quite well.

Equation (S-1.41) depends on λ only through ϵ^* [see (S-1.40)]. We can remove this dependence by setting ϵ^* to a sufficient small fixed value, such as 0.001. Then, we are left with a nonlinear equation for τ, which we can solve by employing an appropriate numerical method, such as Newton-Raphson [3]. Note that we must choose ϵ^* so that the resulting value of τ is greater than one.

Using the value of τ calculated in the previous step, we can calculate the value of λ by solving (S-1.42), in which case

$$\lambda = \frac{\sqrt{2}\, r\tau}{2} \pm \frac{1}{2}\sqrt{2(r^2+2)\tau^2 - 4\tau(\kappa^{\max} - \kappa^{\text{avg}})}\,. \tag{S-1.43}$$

If $\tau < 2(\kappa^{\max} - \kappa^{\text{avg}})/(r^2+2)$, we have no real-valued solution for λ. This indicates that we cannot find an appropriate prior density that satisfies the required specifications (i.e., $\text{E}[\kappa] = \kappa^{\text{avg}}$ and $p(\kappa) \simeq 0$, for $\kappa \leq \kappa^{\min}$, $\kappa \geq \kappa^{\max}$). On the other hand, (S-1.43) may produce two different nonnegative values for λ. In this case, we can use both values of λ to evaluate the corresponding prior densities $p(\kappa)$. Then, we can pick the value that leads to a prior density that best satisfies the condition $p(\kappa^{\min}) = p(\kappa^{\max}) = 0$.

After calculating λ by (S-1.43), we must use ϵ^* and (S-1.40) to determine the value of ϵ. If the resulting value is sufficiently close to zero, then we can accept the hyperparameter values. Otherwise, we must decrease ϵ^* [note from (S-1.40) that $\epsilon \geq \epsilon^*$] and repeat the previous procedure. Finally, having computed the values of τ and λ, we can calculate the value of κ^0 by setting

$$\kappa^0 = \kappa^{\text{avg}} + \tau. \tag{S-1.44}$$

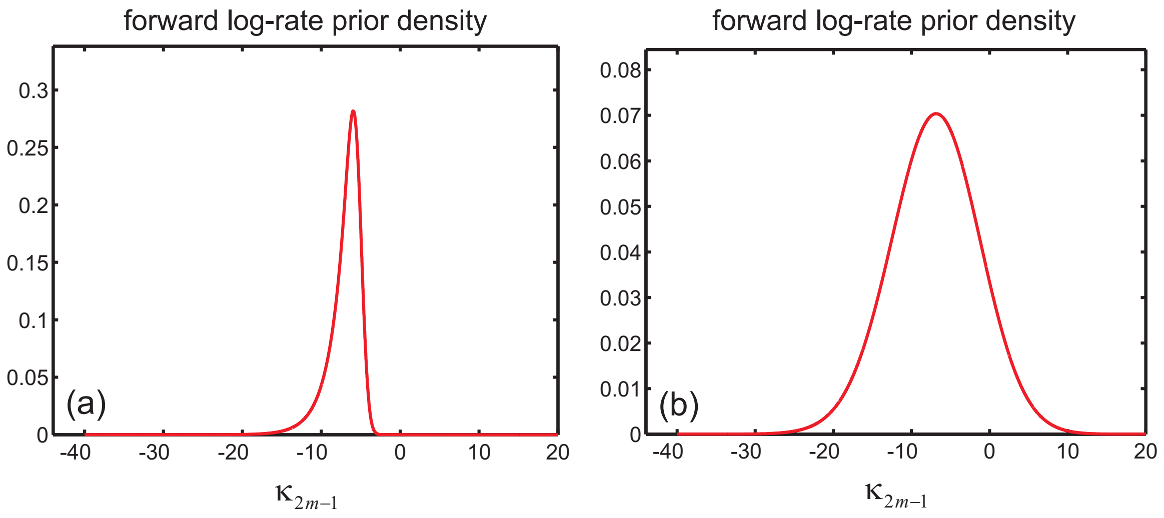

Figure S-1.2: *The prior density* $p(\kappa_{2m-1})$ *of the forward log-rate constant, given by (S-1.31), when* $\kappa_m^0 = -5.1010$*,* $\tau_m = 1.8990$*, and (a)* $\lambda_m = 0.7409$*, (b)* $\lambda_m = 5.3849$*. Note that, contrary to our expectations,* $p(-17) \not\simeq 0$ *and* $p(-3) \not\simeq 0$ *in (b).*

As an example, let us take $\kappa^{\min} = -17$, $\kappa^{\max} = -3$, and $\kappa^{\text{avg}} = -7$ (these are values we consider in the numerical example), and set $\epsilon^* = 0.001$. Then, (S-1.41) becomes $\tau \ln \tau - 5.9\tau = -10$, which is satisfied with $\tau = 1.8990$. Subsequently, (S-1.43) results in $\lambda = 0.7409$ or $\lambda \simeq 5.3849$, whereas, (S-1.44) gives $\kappa^0 = -5.1010$. The resulting prior density with $\lambda = 0.7409$ is depicted in Figure S-1.2(a). In this case, $\epsilon = 0.0013$, which is sufficiently close to zero. The resulting prior density with $\lambda = 5.3849$ is depicted in Figure S-1.2(b). Clearly, this prior is not acceptable, since it turns out that $p(\kappa^{\min}) = p(-17) \not\simeq 0$ and $p(\kappa^{\max}) = p(-3) \not\simeq 0$.

When the average value, avg, and the spread, sd, of the variance σ^2 of the measurement errors are known, we can uniquely determine the hyperparameters a, b associated with the prior error density $p(\sigma^2)$, given by Equation (9) in the Main text. This is due to the fact that $\text{E}[\sigma^2] = b/(a-1)$ and $\text{var}[\sigma^2] = \{\text{E}[\sigma^2]\}^2/(a-2) = b^2/[(a-1)^2(a-2)]$, for $a > 2$, which imply that

$$\begin{aligned} a &= 2 + \left(\frac{\text{avg}}{\text{sd}}\right)^2 \\ b &= \text{avg} \cdot (a-1). \end{aligned}$$

For example, if avg $=$ sd $= 0.5$, then $a = 3$ and $b = 1$, whereas, if avg $= 0.2$ and sd $= 0.1$, then $a = 6$ and $b = 1$. The resulting prior densities are depicted in Figure S-1.3.

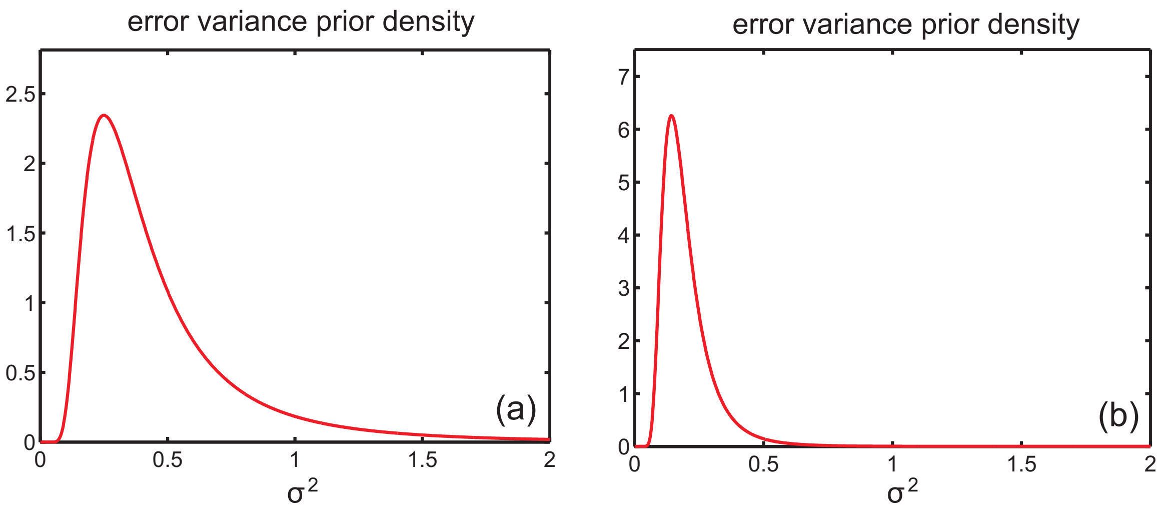

Figure S-1.3: *The prior error variance density $p(\sigma^2)$, given by Equation (9) in the Main text, with: (a) $a = 3$, $b = 1$, and (b) $a = 6$, $b = 1$. In the first case,* $\mathrm{E}[\sigma^2] = 0.5$ *and* $\mathrm{var}[\sigma^2] = 0.25$*, whereas, in the second case,* $\mathrm{E}[\sigma^2] = 0.2$ *and* $\mathrm{var}[\sigma^2] = 0.01$.

Posterior mode vs. posterior mean

We have mentioned in the Main text that the posterior mode should be a more preferable estimator of the kinetic parameters of a biochemical reaction system than the posterior mean. To see why this is true, suppose that, so long as $|\kappa_{f,m} - \kappa_{f,m}^{\text{true}}| < \epsilon_m$, for $m = 1, 2, ..., M + M_1$, the parameters $\{\boldsymbol{\kappa}_f, \mathbb{W}\boldsymbol{\kappa}_f\}$ reproduce the concentration dynamics of the biochemical reaction system faithfully. Note that a small ϵ_m corresponds to a rate constant that appreciably affects the concentration dynamics, whereas, a large ϵ_m corresponds to a rate constant whose value has little or no effect on the dynamics. If $c(\boldsymbol{\kappa}_f, \boldsymbol{\kappa}_f^{\text{true}})$ is the cost of estimating the true log-rate constants $\boldsymbol{\kappa}_f^{\text{true}}$ by $\boldsymbol{\kappa}_f$, then we can set

$$c(\boldsymbol{\kappa}_f, \boldsymbol{\kappa}_f^{\text{true}}) = \begin{cases} 0, & \text{if } |\kappa_{f,m} - \kappa_{f,m}^{\text{true}}| < \epsilon_m, \ \text{ for } m = 1, 2, ..., M + M_1 \\ 1, & \text{otherwise .} \end{cases} \tag{S-1.45}$$

As a consequence, we would like to find the optimal estimator $\widehat{\boldsymbol{\kappa}}_f$ of $\boldsymbol{\kappa}_f^{\text{true}}$ by minimizing the mean posterior cost $\mathrm{E}[c(\boldsymbol{\kappa}_f, \boldsymbol{\kappa}_f^{\text{true}}) \,|\, \boldsymbol{y}]$ with respect to $\boldsymbol{\kappa}_f$. Note that (S-1.45) implies that

$$\mathrm{E}[c(\boldsymbol{\kappa}_f, \boldsymbol{\kappa}_f^{\text{true}}) \,|\, \boldsymbol{y}] = 1 - \Pr(\{|\kappa_{f,m} - \kappa_{f,m}^{\text{true}}| < \epsilon_m, \ \text{ for } m = 1, 2, ..., M + M_1\} \,|\, \boldsymbol{y}). \tag{S-1.46}$$

Hence, we can find the optimal estimator $\widehat{\boldsymbol{\kappa}}_f$ by maximizing the probability

$$\Pr(\{|\kappa_{f,m} - \kappa_{f,m}^{\text{true}}| < \epsilon_m, \ \text{ for } m = 1, 2, ..., M + M_1\} \mid \boldsymbol{y}) = \int_{\kappa_{f,1} - \epsilon_1}^{\kappa_{f,1} + \epsilon_1} \cdots \int_{\kappa_{f,M+M_1} - \epsilon_{M+M_1}}^{\kappa_{M+M_1} + \epsilon_{M+M_1}} p_W(\boldsymbol{\kappa}' | \boldsymbol{y}) d\boldsymbol{\kappa}',$$

with respect to $\boldsymbol{\kappa}_f$.

Clearly, the optimal solution is the center of a hypercube with edge lengths $\{2\epsilon_1, 2\epsilon_2, ..., 2\epsilon_{M+M_1}\}$ with the highest probability given by (S-1.46). When all ϵ_m's are small (i.e., when all parameters appreciably affect the system dynamics), the hypercube will be small as well. In this case, $\widehat{\boldsymbol{\kappa}}_f$ will approximately be the point in the parameter space with the highest posterior probability and, therefore, $\widehat{\boldsymbol{\kappa}}_f = \widehat{\boldsymbol{\kappa}}_f^{\text{mode}}$.

When a parameter does not appreciably affect the system dynamics, the hypercube grows along the corresponding dimension. In this case, the skewness of the posterior density $p_W(\boldsymbol{\kappa}_f \mid \boldsymbol{y})$ may draw the optimal estimator away from the posterior mode along the direction of growth. This however is an acceptable loss of optimality, since the parameter does not appreciably affect the concentration dynamics and finding its optimal value is inconsequential.

References

1. Israel A, Greville TNE: *Generalized Inverses: Theory and Applications*. New York: Springer-Verlag, 2nd edition 2003.
2. Vlad MO, Ross J: **Thermodynamically based constraints for rate coefficients of large biochemical networks**. *WIREs Syst. Biol. Med.* 2009, **1**:348–358.
3. Press WH, Teukolsky SA, Vetterling WT, Flannery BP: *Numerical Recipes: The Art of Scientific Computing*. New York: Cambridge University Press, 3rd edition 2007.
4. Berry RS, Rice SA, Ross J: *Physical Chemistry*. New York: Oxford University Press, 2nd edition 2000.
5. Gillespie DT: **A rigorous derivation of the chemical master equation**. *Physica A* 1992, **188**:404–425.
6. Gillespie DT: **The chemical Langevin equation**. *J. Chem. Phys.* 2000, **113**:297–306.
7. Grönholm T, Annila A: **Natural distribution**. *Math. Biosci.* 2007, **210**:659–667.
8. Vlad MO, Cerofolini G, Oefner P, Ross J: **Random activation energy model and disordered kinetics, from static to dynamic disorder**. *J. Phys. Chem. B* 2005, **109**:21241–21257.
9. Papoulis A, Pillai SU: *Probability, Random Variables and Stochastic Processes*. New York: McGraw Hill, 4th edition 2002.
10. Ederer M, Gilles ED: **Thermodynamically feasible kinetic models of reaction networks**. *Biophys. J.* 2007, **92**:1846–1857.

ADDITIONAL FILE 2

Thermodynamically consistent Bayesian analysis of closed biochemical reaction systems

COMPUTATIONS

Garrett Jenkinson,[1] Xiaogang Zhong,[2] and John Goutsias*[1]

[1]Whitaker Biomedical Engineering Institute, The Johns Hopkins University, Baltimore, MD 21218, USA
[2]Department of Applied Mathematics and Statistics, The Johns Hopkins University, Baltimore, MD 21218, USA

Email: Garrett Jenkinson - jenkinson@jhu.edu; Xiaogang Zhong - xzhong4@jhu.edu; John Goutsias*- goutsias@jhu.edu;

*Corresponding author

The posterior mode and posterior covariance matrix cannot be calculated analytically. For this reason, we need to develop appropriate computational techniques for their numerical evaluation. It turns out that we can effectively compute the posterior mode by employing an optimization algorithm based on stochastic approximation, and estimate the posterior covariance by sampling the posterior density $p_W(\boldsymbol{\kappa}_f \mid \boldsymbol{y})$, given by Equation (24) in the Main text, using an appropriately designed Monte Carlo method. In this document, we provide a detailed discussion on how to do this.

Computing the prior mode

Evaluating the posterior mode via optimization requires an initial value $\boldsymbol{\kappa}_f(0)$ for the "free" log-rate constants $\boldsymbol{\kappa}_f$. A good choice for such value can be obtained by maximizing the "effective" thermodynamically consistent prior density

$$p_W(\boldsymbol{\kappa}_f, \boldsymbol{\kappa}_d) \propto \delta(\boldsymbol{\kappa}_d - \mathbb{W}\boldsymbol{\kappa}_f) \int p(\boldsymbol{\kappa}_f, \boldsymbol{\kappa}_d \mid \boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z},$$

given by Equation (23) in the Main text. As a consequence of Equation (17) in the Main text and the fact that we replace $p(\boldsymbol{z})$ by the conditional density $p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}})$, given by Equation (15) in the Main text, we set

$$\begin{aligned}
\boldsymbol{\kappa}(0) &= \arg\max_{\boldsymbol{\kappa}} p_W(\boldsymbol{\kappa}) \\
&= \arg\max_{\boldsymbol{\kappa} \in \mathcal{W}} \int p(\boldsymbol{\kappa} \mid \boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z} \\
&= \arg\max_{\boldsymbol{\kappa} \in \mathcal{W}} \Big[\prod_{m \in \mathcal{M}} p(\kappa_{2m-1}) \Big] \int \Big[\prod_{m \in \mathcal{M}} \delta(\kappa_{2m} - \kappa_{2m-1} + z_m) \Big] p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}}) d\boldsymbol{z} \\
&= \arg\max_{\boldsymbol{\kappa} \in \mathcal{W}} \Big[\prod_{m \in \mathcal{M}} p(\kappa_{2m-1}) \Big] p(\{z_m = \kappa_{2m-1} - \kappa_{2m}, m \in \mathcal{M}\} \mid \widetilde{\boldsymbol{y}}),
\end{aligned}$$

where $\boldsymbol{\kappa} = \{\boldsymbol{\kappa}_f, \boldsymbol{\kappa}_d\}$ and $\mathcal{W}$ is the thermodynamically consistent region of the parameter space, given by the hyperplane $\boldsymbol{\kappa}_d = \mathbb{W}\boldsymbol{\kappa}_f$. The solution to this problem consists of finding the forward log-rate constants $\{\kappa_{2m-1}(0), m \in \mathcal{M}\}$ that maximize the first term $\prod_{m \in \mathcal{M}} p(\kappa_{2m-1})$, calculating thermodynamically consistent log-equilibrium constants $\{z_m(0), m \in \mathcal{M}\}$ that maximize the second term $p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}})$, and setting $\kappa_{2m}(0) = \kappa_{2m-1}(0) - z_m(0)$, for $m \in \mathcal{M}$. Note that the Wegscheider conditions, given by Equation (11) in the Main text, are equivalent to the following conditions:

$$z_{m'} = \sum_{m \in \mathcal{M}_1} [\mathbb{S}_{11}^{-1} \mathbb{S}_{12}]_{m,m'} z_m, \quad \text{for every } m' \in \mathcal{M}_2, \qquad \text{(S-2.1)}$$

by virtue of Equation (8) in the Main text and Equation (S-1.7) in Additional file 1, where $\mathcal{M}_1 = \{1, 2, \ldots, M_1\}$ and $\mathcal{M}_2 = \{M_1 + 1, M_1 + 2, \ldots, M\}$, with $M_1 = \text{rank}(\mathbb{S})$.

To compute the initial forward log-rate constants $\{\kappa_{2m-1}(0), m \in \mathcal{M}\}$, we must find, for each $m \in \mathcal{M}$, the value that maximizes the prior density $p(\kappa_{2m-1})$, given by Equation (16) in the Main text. This problem can be easily solved by a grid search approach that calculates $p(\kappa_{2m-1})$ on a finely spaced uniform grid of points and by detecting the maximum value [1].

To compute a thermodynamically consistent value $\boldsymbol{z}(0)$ that maximizes the conditional density $p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}})$, given by Equation (15) in the Main text, note that

$$\boldsymbol{z} = \mathbb{G}\boldsymbol{z}^{(1)},$$

by virtue of (S-2.1), where $\boldsymbol{z}^{(1)} = \{z_m, m \in \mathcal{M}_1\}$ and

$$\mathbb{G} = \begin{bmatrix} \mathbb{I}_{M_1} \\ (\mathbb{S}_{11}^{-1}\mathbb{S}_{12})^T \end{bmatrix},$$

with $\mathbb{I}_{M_1}$ being the $M_1 \times M_1$ identity matrix. Maximizing $p(\boldsymbol{z} \mid \widetilde{\boldsymbol{y}})$ with respect to $\boldsymbol{z}$ is now equivalent to maximizing $p(\mathbb{G}\boldsymbol{z}^{(1)} \mid \widetilde{\boldsymbol{y}})$ with respect to $\boldsymbol{z}^{(1)}$. This maximization problem leads to solving the system $\mathbb{U}_0^T\mathbb{G}\boldsymbol{z}^{(1)} = \mathbb{U}_0^T\widetilde{\boldsymbol{y}}$ of linear equations with respect to $\boldsymbol{z}^{(1)}$. A least-squares solution to this problem leads to $\boldsymbol{z}^{(1)}(0) = (\mathbb{U}_0^T\mathbb{G})^\dagger\mathbb{U}_0^T\widetilde{\boldsymbol{y}}$, where $\mathbb{A}^\dagger$ denotes the Moore-Penrose pseudoinverse of matrix $\mathbb{A}$. As a consequence, we have that $\boldsymbol{z}(0) = \mathbb{G}(\mathbb{U}_0^T\mathbb{G})^\dagger\mathbb{U}_0^T\widetilde{\boldsymbol{y}}$.

Finding the posterior mode

It is clear from Equations (21), (22), and (24) in the Main text that, in order to evaluate the mode $\widehat{\boldsymbol{\kappa}}_f^{\text{mode}}$, we need an algorithm for solving the following nonlinear optimization problem:

$$\widehat{\boldsymbol{\kappa}}_f^{\text{mode}} = \arg\max_{\boldsymbol{\kappa}_f} C(\boldsymbol{\kappa}_f \mid \boldsymbol{y}),$$

where $C(\boldsymbol{\kappa}_f \mid \boldsymbol{y}) := D(\boldsymbol{\kappa}_f, \mathbb{W}\boldsymbol{\kappa}_f \mid \boldsymbol{y})$, with

$$\begin{aligned} D(\boldsymbol{\kappa}_f, \boldsymbol{\kappa}_d \mid \boldsymbol{y}) \;=\; & \sum_{m \in \mathcal{M}} \frac{\kappa_{2m-1}}{\tau_m} + \ln\left(\text{erfc}\left[\frac{1}{\sqrt{2}}\left(\frac{\lambda_m}{\tau_m} + \frac{\kappa_{2m-1} - \kappa_m^0}{\lambda_m}\right)\right]\right) \\ & -a \ln\left[\frac{2b}{P+1} + \sum_{m \in \mathcal{M}} \sum_{m \in \mathcal{M}'} \theta_{mm'}(\kappa_{2m-1} - \kappa_{2m} - \widetilde{y}_m)(\kappa_{2m'-1} - \kappa_{2m'} - \widetilde{y}_{m'})\right] \\ & - [a + NQ(P+1)/2] \ln\left(2b + \sum_{n \in \mathcal{N}} \sum_{q \in \mathcal{Q}} \sum_{p \in \mathcal{P}} [y_n^{(p)}(t_q) - \ln x_n^{(p)}(t_q)]^2\right). \qquad \text{(S-2.2)} \end{aligned}$$

Although a number of different optimization approaches can be employed to solve this problem, we will use here a method based on simultaneous perturbation stochastic approximation (SPSA) [2]. SPSA is a gradient-free ascent algorithm, which estimates the gradient using a finite difference of the objective function evaluated at random perturbations around the current parameter values. The most attractive features of this method are robustness to noise, computational efficiency, and scalability.

The SPSA recursion is given by

$$\boldsymbol{\kappa}_f(i+1) = \boldsymbol{\kappa}_f(i) + \gamma_i \, \boldsymbol{g}_i(\boldsymbol{\kappa}_f(i)), \quad \text{for } \; i = 0, 1, \ldots,$$

where $\{\gamma_i, i = 0, 1, \ldots\}$ is a decreasing sequence of nonnegative numbers and $\{\boldsymbol{g}_i(\boldsymbol{\kappa}_f), i = 0, 1, \ldots\}$ is a sequence of estimators of the gradient of the objective function $C(\boldsymbol{\kappa}_f \mid \boldsymbol{y})$ at point $\boldsymbol{\kappa}_f$. The gradient estimator $\boldsymbol{g}_i$ is a $2M \times 1$ random vector with elements $g_{i,m}$, given by

$$g_{i,m}(\boldsymbol{\kappa}_f) = \frac{C(\boldsymbol{\kappa}_f + \delta_i \boldsymbol{\epsilon}_i \mid \boldsymbol{y}) - C(\boldsymbol{\kappa}_f - \delta_i \boldsymbol{\epsilon}_i \mid \boldsymbol{y})}{2\delta_i \epsilon_{i,m}}, \;\; m = 1, 2, \ldots, 2M, \tag{S-2.3}$$

where $\boldsymbol{\epsilon}_i$ is a $2M \times 1$ random vector with statistically independent random elements $\epsilon_{i,m}$ that follow a ± 1 Bernoulli distribution with equal success and failure probabilities, and $\{\delta_i, i = 0, 1, \ldots\}$ is a decreasing sequence of nonnegative numbers. Parameters γ_i and δ_i should be chosen based on standard guidelines provided in [2]. By following these guidelines, we set

$$\gamma_i = \frac{\gamma}{(i+1+A)^{0.602}} \qquad \text{and} \qquad \delta_i = \frac{\delta}{(i+1)^{0.101}} \, . \tag{S-2.4}$$

We take the value of A to be $1/10$ of the total number of SPSA iterations. Parameter δ can be set at a level that is approximately equal to the standard deviation of the noise in measuring the objective function C. In our case, this standard deviation is directly related to the error tolerance associated with the ODE integrator we use to integrate Equation (2) in the Main text (see also our discussion below). For this reason, we take the value of δ to be the same as the ODE error tolerance. Finally, we choose γ to satisfy the following equation:

$$\frac{\gamma}{(1+A)^{0.602}} \, \mathrm{E}[||\boldsymbol{g}_0(\boldsymbol{\kappa}_f(0))||] = s_0 ||\boldsymbol{\kappa}_f(0)||,$$

where $||\boldsymbol{x}||$ denotes the magnitude of vector $\boldsymbol{x}$. This guarantees that, on the average, the initial SPSA step yields log-rate values within a sufficiently large neighborhood around the initial point $\boldsymbol{\kappa}_f(0)$, measured by the initial search size s_0.

The algorithm for finding the posterior mode proceeds as follows:

Initialization

1. Select values for the hyperparameters $\{\kappa_m^0, \lambda_m, \tau_m, m \in \mathcal{M}\}$, associated with the prior densities $p(\kappa_m)$ of the forward log-rate constants, and values for the hyperparameters $\{a, b\}$, associated with the prior density of the error variance. A practical method for determining these values was discussed in the Additional file 1.

2. Select a desirable number I of SPSA iterations, a desirable level *tol* of ODE error tolerance, and an initial search size s_0 (we set $tol = 1 \times 10^{-3}$ and $s_0 = 0.01$).

3. Calculate an initial guess $\boldsymbol{\kappa}_f(0)$ for the "free" log-rate constants by following the approach discussed in the previous section.

4. In (S-2.4), set $A = I/10$, $\delta = tol$, and

$$\gamma = \frac{s_0 ||\boldsymbol{\kappa}_f(0)|| (1 + A)^{0.602}}{\dfrac{1}{L_0} \displaystyle\sum_{l=1}^{L} ||\boldsymbol{g}_0^{(l)}(\boldsymbol{\kappa}_f(0))||},$$

where $\{\boldsymbol{g}_0^{(l)}(\boldsymbol{\kappa}_f(0)), l = 1, 2, \ldots, L_0\}$ are statistically independent realizations of the initial gradient estimator $\boldsymbol{g}_0(\boldsymbol{\kappa}_f(0))$, and L_0 is a sufficiently large integer, so that the denominator in the previous formula provides a sufficiently good approximation of the average initial gradient E$[||\boldsymbol{g}_0(\boldsymbol{\kappa}_f(0))||]$ (we take $L_0 = 10$).

Iteration

For $i = 0, 1, \ldots, I - 1$:

5. Draw $2M$ statistically independent samples $\{\epsilon_{i,m}, m = 1, 2, \ldots, 2M\}$ from a ±1 Bernoulli distribution with equal success and failure probabilities.

6. By using (S-2.3) and (S-2.4), calculate the $2M$ gradient values $\{g_{i,m}(\boldsymbol{\kappa}_f(i)), m = 1, 2, \ldots, 2M\}$ and use them to calculate new log-rate constant values $\boldsymbol{\kappa}_f(i + 1) = \boldsymbol{\kappa}_f(i) + \gamma_i \boldsymbol{g}_i(\boldsymbol{\kappa}_f(i))$.

Each iteration of the previous optimization algorithm requires computation of the response of the biochemical reaction system under consideration $2(P + 1)$ times [for evaluating the objective function twice]. If parallel computation is available, the system evaluations required by Step 6 can be done independently. If only serial implementation is available, then an effort should be made to reduce the time it takes to integrate the system ODE's.

An important computational trick, which we have implemented with large performance gains, comes from the fact that SPSA enjoys superior performance with noisy objective function evaluations. However, our biochemical reaction system is characterized by deterministic ODE's, which can lead to error-free objective function evaluation, provided that exact integration of these ODE's is possible. We can take advantage of the fact that most ODE integrators have a built-in error tolerance setting that controls the accuracy of integration. Small error tolerances improve integration accuracy at the expense of increasing computations, whereas, large error tolerances dramatically decrease computations but produce less accurate integrations. Therefore, we can effectively reduce the required computational time by relaxing the ODE error tolerance at the expense of adding “noise” to the evaluation of the objective function. It is expected however that a reasonable amount of “noise” will not appreciably affect the performance of SPSA due to its robustness against inaccurate objective function evaluations [2].

It is a common practice to consider the mode estimator as being the final product $\boldsymbol{\kappa}_f(I)$ of the previous SPSA iterations. However, the value of the objective function C at $\boldsymbol{\kappa}_f(I)$ may not be the largest value obtained during the course of SPSA, due to the fact that SPSA is a stochastic optimization algorithm. An alternative is to consider the mode estimator as being the point in the parameter space at which the value of the objective function becomes maximum during the SPSA iterations, i.e.,

$$\widehat{\boldsymbol{\kappa}}_f^{\text{mode}} = \arg\max \{C(\boldsymbol{\kappa}_f(i) \mid \boldsymbol{y}), i = 0, 1, \ldots, I\}\,.$$

However, implementation of this equation requires computation of C at each SPSA iteration, which in turn requires an additional number of $I+1$ system evaluations.

To address this problem, note that evaluation of the gradient $\boldsymbol{g}_i(\boldsymbol{\kappa}_f(i))$, for $i = 0, 1, \ldots, I-1$, requires computation of the objective function C at points $\boldsymbol{\kappa}_f(i) \pm \delta_i \boldsymbol{\epsilon}_i$, which are proximal to $\boldsymbol{\kappa}_f(i)$. We can therefore approximate the value of the objective function at $\boldsymbol{\kappa}_f(i)$, for $i = 0, 1, \ldots, I-1$, by averaging the two values $C(\boldsymbol{\kappa}_f(i) \pm \delta_i \boldsymbol{\epsilon}_i \mid \boldsymbol{y})$; i.e., we can set

$$C(\boldsymbol{\kappa}_f(i) \mid \boldsymbol{y}) \simeq \frac{C(\boldsymbol{\kappa}_f(i) + \delta_i \boldsymbol{\epsilon}_i \mid \boldsymbol{y}) + C(\boldsymbol{\kappa}_f(i) - \delta_i \boldsymbol{\epsilon}_i \mid \boldsymbol{y})}{2}, \quad \text{for } \; i = 0, 1, \ldots, I-1.$$

Extensive simulations indicate that this modification consistently outperforms the standard SPSA algorithm presented above without requiring additional cost function evaluations.

Estimating the posterior mean and covariance matrix

A potential technique for estimating the posterior mean and covariance matrix is Monte Carlo sampling. This method can be used to estimate posterior expectations of the form $\mathrm{E}[f(\boldsymbol{\kappa}_f) \mid \boldsymbol{y}]$ by generating a large number L of independent and identically distributed (i.i.d.) samples $\{\boldsymbol{\kappa}_f(1), \boldsymbol{\kappa}_f(2), \ldots, \boldsymbol{\kappa}_f(L)\}$, drawn from the posterior distribution $p_W(\boldsymbol{\kappa}_f \mid \boldsymbol{y})$, and by setting

$$\mathrm{E}[f(\boldsymbol{\kappa}_f) \mid \boldsymbol{y}] = \int f(\boldsymbol{\kappa}_f) p_W(\boldsymbol{\kappa}_f \mid \boldsymbol{y}) d\boldsymbol{\kappa}_f \simeq \frac{1}{L} \sum_{l=1}^{L} f(\boldsymbol{\kappa}_f(l))\,. \qquad \text{(S-2.5)}$$

Since the samples are i.i.d., the law of large numbers dictates that an arbitrary degree of estimation accuracy can be achieved by using a sufficiently large number of samples [3].

Unfortunately, this framework is overly restrictive for our problem, since drawing i.i.d. samples from the posterior distribution is a very difficult, if not an impossible, task. An alternative is to use a Markov chain Monte Carlo (MCMC) method, which uses *dependent* samples generated from an ergodic Markov chain converging to $p_W(\boldsymbol{\kappa}_f \mid \boldsymbol{y})$, to estimate the integral in (S-2.5). Indeed, by constructing an appropriate ergodic Markov Chain that generates dependent samples $\{\boldsymbol{\kappa}_f(1), \boldsymbol{\kappa}_f(2), \ldots, \boldsymbol{\kappa}_f(L)\}$, we can guarantee that the sum in (S-2.5) will converge (usually in a mean-square or an almost sure sense) to the posterior mean of $f(\boldsymbol{\kappa}_f)$, as $L \to \infty$ [3].

Although there are several methods for constructing an ergodic MCMC sampling approach, we utilize here the Metropolis algorithm (MA), primarily due to its ease of implementation and known effectiveness in a Bayesian setting. This algorithm proceeds as follows. Given parameters $\boldsymbol{\kappa}_f(l)$ at step l, a new "tentative" set of parameters $\boldsymbol{\kappa}'_f(l)$ is proposed, drawn from a *symmetric* probability distribution $q(\boldsymbol{\kappa}'_f \mid \boldsymbol{\kappa}_f(l))$, satisfying the condition $q(\boldsymbol{\kappa}'_f \mid \boldsymbol{\kappa}_f) = q(\boldsymbol{\kappa}_f \mid \boldsymbol{\kappa}'_f)$, for every $\boldsymbol{\kappa}'_f$ and $\boldsymbol{\kappa}_f$, known as the *proposal* distribution. Then, if $p_W(\boldsymbol{\kappa}'_f(l) \mid \boldsymbol{y}) \geq p_W(\boldsymbol{\kappa}_f(l) \mid \boldsymbol{y})$, we accept $\boldsymbol{\kappa}'_f(l)$ as being the new parameters [i.e., we set $\boldsymbol{\kappa}_f(l+1) = \boldsymbol{\kappa}'_f(l)$]; otherwise, we accept $\boldsymbol{\kappa}'_f(l)$ with probability $p_W(\boldsymbol{\kappa}'_f(l) \mid \boldsymbol{y})/p_W(\boldsymbol{\kappa}_f(l) \mid \boldsymbol{y})$ and reject $\boldsymbol{\kappa}'_f(l)$ [i.e., we set $\boldsymbol{\kappa}_f(l+1) = \boldsymbol{\kappa}_f(l)$] with probability $1 - p_W(\boldsymbol{\kappa}'_f(l) \mid \boldsymbol{y})/p_W(\boldsymbol{\kappa}_f(l) \mid \boldsymbol{y})$.

Note that evaluation of the acceptance/rejection probability requires knowledge of the posterior distribution only up to a constant. This is one reason why MA-MCMC is favorable in a Bayesian setting where it is usually impossible to calculate the proportionality factor associated with the posterior distribution. Another attractive feature is that the proposal distribution can be any symmetric distribution, although choosing this distribution wisely can ensure faster convergence. We can improve the convergence rate if we choose a proposal distribution that results in moderate acceptance rates [2].

Another point worth mentioning here is the *burn-in period* associated with MCMC sampling. The burn-in period is the initial number of MCMC iterations during which the Markov chain has not yet converged to its stationary distribution $p_W(\boldsymbol{\kappa}_f \mid \boldsymbol{y})$. Theoretically speaking, all samples produced by MCMC can be used in (S-2.5). It is however customary to ignore samples during the burn-in period from the computation, hoping that the sum in (S-2.5) will converge faster to the expected value if only samples drawn from the posterior distribution are used.

Unfortunately, it is not easy to accurately determine the burn-in period. Moreover, a large burn-in period may substantially and unnecessarily increase the overall computational effort. Therefore, it would be more attractive if we could initialize the MCMC algorithm with parameters $\boldsymbol{\kappa}_f(1)$ drawn from the posterior distribution $p_W(\boldsymbol{\kappa}_f \mid \boldsymbol{y})$, in which case the burn-in period would be zero, since the Markov chain would be stationary for every $l = 1, 2, \ldots, L$. Of course, this is not possible. However, we can choose $\boldsymbol{\kappa}_f(1)$ to be the posterior mode, in which case we can approximately consider $\boldsymbol{\kappa}_f(1)$ as being a sample drawn from the posterior distribution with the highest probability. This of course will be a good approximation in the ideal case when $\boldsymbol{\kappa}_f(1)$ is indeed the posterior mode and the posterior distribution is tightly clustered around the mode. In practice however the posterior distribution is spread out and we do not know the posterior mode, so $\boldsymbol{\kappa}_f(1)$ is only a local maximum of the posterior distribution found by optimization. Our experience indicates that, by initializing the MCMC sampling algorithm with a local maximum of the posterior distribution obtained by SPSA, we can substantially reduce the number of MCMC iterations required to obtain sufficiently accurate estimates of the posterior mean and covariance matrix.

As a result of the previous discussion, we will adopt the following algorithm for estimating the posterior mean and covariance matrix:

Initialization

1. Select a desirable number L of MA-MCMC iterations.
2. Set $\boldsymbol{\kappa}_f(1) = \widehat{\boldsymbol{\kappa}}_f^{\text{mode}}$, where $\widehat{\boldsymbol{\kappa}}_f^{\text{mode}}$ is obtained after I iterations of the SPSA algorithm discussed in the previous subsection.
3. Set $\xi = 0.1$. Take the proposal distribution $q(\boldsymbol{\kappa}'_f \mid \boldsymbol{\kappa}_f)$ to be the uniform distribution over the hypercube $[\boldsymbol{\kappa}_f - \xi\boldsymbol{e}, \boldsymbol{\kappa}_f + \xi\boldsymbol{e}]$ centered around $\boldsymbol{\kappa}_f$, where $\boldsymbol{e}$ is a vector with all of its elements being equal to one and ξ is a parameter that controls the size of the hypercube in order to achieve a desirable acceptance rate.

Iteration

For $l = 1, 2, \ldots, L$:

4. Draw $2M$ statistically independent samples $\boldsymbol{\epsilon}(l) = \{\epsilon_m(l), m = 1, 2, \ldots, 2M\}$ from the uniform distribution over $[-1, +1]$ and set $\boldsymbol{\kappa}'_f(l) = \boldsymbol{\kappa}_f(l) + \xi\boldsymbol{\epsilon}(l)$.
5. Use (S-2.2) to calculate $C(\boldsymbol{\kappa}'_f(l) \mid \boldsymbol{y})$ and $C(\boldsymbol{\kappa}_f(l) \mid \boldsymbol{y})$ and set $\rho := p_W(\boldsymbol{\kappa}'_f(l) \mid \boldsymbol{y})/p_W(\boldsymbol{\kappa}_f(l) \mid \boldsymbol{y}) = \exp\{C(\boldsymbol{\kappa}'_f(l) \mid \boldsymbol{y}) - C(\boldsymbol{\kappa}_f(l) \mid \boldsymbol{y})\}$.

6. Generate a uniformly distributed random number u over $[0, 1]$.

7. If $\rho \geq u$, set $\boldsymbol{\kappa}_f(l+1) = \boldsymbol{\kappa}'_f(l)$; otherwise, set $\boldsymbol{\kappa}_f(l+1) = \boldsymbol{\kappa}_f(l)$.

Estimation

8. Set

$$\widehat{\boldsymbol{\kappa}}_f^{\text{mean}} = \frac{1}{L} \sum_{l=1}^{L} \boldsymbol{\kappa}_f(l)$$

$$\widehat{\mathbb{V}} = \frac{1}{L} \sum_{l=1}^{L} \left[\boldsymbol{\kappa}_f(l) - \widehat{\boldsymbol{\kappa}}_f^{\text{mode}}\right] \left[\boldsymbol{\kappa}_f(l) - \widehat{\boldsymbol{\kappa}}_f^{\text{mode}}\right]^T.$$

Computing the posterior mode

The objective function $C(\boldsymbol{\kappa}_f \mid \boldsymbol{y})$ is usually not concave, especially when a limited amount of highly noisy data $\boldsymbol{y}$ is available. As a consequence, there is no optimization algorithm that can find the posterior mode in a finite number of steps. However, the following algorithm, which we refer to as maximization-expectation-maximization (MEM) algorithm, performs quite well in our simulations.

Maximization

1. Calculate an initial guess $\boldsymbol{\kappa}_f(0)$ for the log-rate constants by using the previously discussed approach.

2. Perform I SPSA iterations, initialized by $\boldsymbol{\kappa}_f(0)$, to obtain the posterior mode estimate $\widehat{\boldsymbol{\kappa}}_{f,1}^{\text{mode}}$.

Expectation

3. Perform L MA-MCMC iterations, initialized with $\widehat{\boldsymbol{\kappa}}_{f,1}^{\text{mode}}$, to obtain the posterior mean estimate $\widehat{\boldsymbol{\kappa}}_f^{\text{mean}}$.

Maximization

4. Perform I SPSA iterations, initialized by $\widehat{\boldsymbol{\kappa}}_f^{\text{mean}}$, to obtain the posterior mode estimate $\widehat{\boldsymbol{\kappa}}_{f,2}^{\text{mode}}$.

Final Mode Estimate

5. Set $\widehat{\boldsymbol{\kappa}}_f^{\text{mode}} = \arg\max \left\{C(\widehat{\boldsymbol{\kappa}}_{f,1}^{\text{mode}} \mid \boldsymbol{y}), C(\widehat{\boldsymbol{\kappa}}_{f,2}^{\text{mode}} \mid \boldsymbol{y})\right\}$.

A variation of the previous algorithm, which we found to be effective, is to keep track of all objective function evaluations $C(\boldsymbol{\kappa}_f(l) \mid \boldsymbol{y})$, $l = 2, 3, \ldots, L$, during the MA-MCMC (expectation) step, and take

$$\widehat{\boldsymbol{\kappa}}_f^{\text{mode}} = \arg\max \left\{C(\widehat{\boldsymbol{\kappa}}_{f,1}^{\text{mode}} \mid \boldsymbol{y}), C(\boldsymbol{\kappa}_f(l) \mid \boldsymbol{y}), l = 2, 3, \ldots, L, C(\widehat{\boldsymbol{\kappa}}_{f,2}^{\text{mode}} \mid \boldsymbol{y})\right\}.$$

The advantage of this strategy is that it does not waste the objective values evaluated during the MA-MCMC iterations and accounts for the possibility that MA-MCMC may produce parameter values at some iteration that are closer to the actual posterior mode than the parameters obtained by the two SPSA steps.

References

1. Bazaraa MS, Sherali HD, Shetty CM: *Nonlinear programming: Theory and algorithms*. Hoboken, NJ: John Wiley & Sons, 3rd edition 2006.
2. Spall JC: *Introduction to Stochastic Search and Optimization: Estimation, Simulation and Control*. New York: Wiley-Interscience 2003.
3. Liu JS: *Monte Carlo Strategies in Scientific Computing*. New York: Springer-Verlag 2001.

ADDITIONAL FILE 3

Thermodynamically consistent Bayesian analysis of closed biochemical reaction systems

NUMERICAL EXAMPLE

Garrett Jenkinson,[1] Xiaogang Zhong,[2] and John Goutsias*[1]

[1]Whitaker Biomedical Engineering Institute, The Johns Hopkins University, Baltimore, MD 21218, USA
[2]Department of Applied Mathematics and Statistics, The Johns Hopkins University, Baltimore, MD 21218, USA

Email: Garrett Jenkinson - jenkinson@jhu.edu; Xiaogang Zhong - xzhong4@jhu.edu; John Goutsias*- goutsias@jhu.edu;

*Corresponding author

In this document, we list the reactions associated with the biochemical reaction system depicted in Figure 1 of the Main text and provide thermodynamically consistent "true" values for the associated rate constants, as well as appropriate values for the initial concentrations. The example is based on a subset of a well-established model of the EGF/ERK signal transduction pathway proposed by Schoeberl *et al.* [1]. We have obtained published values for the rate constants and initial concentrations from the BioModels database [2].

Model details

The biochemical reaction system depicted in Figure 1 of the Main text is comprised of the following $N = 13$ molecular species:

X_1 :	Shc*
X_2 :	Grb2
X_3 :	Shc*-Grb2
X_4 :	Sos
X_5 :	Shc*-Grb2-Sos
X_6 :	Grb2-Sos
X_7 :	(EGF-EGFR*)2-GAP
X_8 :	(EGF-EGFR*)2-GAP-Grb2
X_9 :	(EGF-EGFR*)2-GAP-Grb2-Sos
X_{10} :	Ras-GDP
X_{11} :	(EGF-EGFR*)2-GAP-Grb2-Sos-Ras-GDP
X_{12} :	Ras-GTP*
X_{13} :	(EGF-EGFR*)2-GAP-Grb2-Sos-Ras-GTP ,

which interact by means of the following $M = 9$ reversible association-dissociation reactions:

$$\begin{aligned}
X_1 + X_2 &\overset{k_1}{\underset{k_2}{\rightleftharpoons}} X_3 \\
X_3 + X_4 &\overset{k_3}{\underset{k_4}{\rightleftharpoons}} X_5 \\
X_2 + X_4 &\overset{k_5}{\underset{k_6}{\rightleftharpoons}} X_6 \\
X_1 + X_6 &\overset{k_7}{\underset{k_8}{\rightleftharpoons}} X_5 \\
X_2 + X_7 &\overset{k_9}{\underset{k_{10}}{\rightleftharpoons}} X_8 \\
X_4 + X_8 &\overset{k_{11}}{\underset{k_{12}}{\rightleftharpoons}} X_9 \\
X_6 + X_7 &\overset{k_{13}}{\underset{k_{14}}{\rightleftharpoons}} X_9 \\
X_9 + X_{10} &\overset{k_{15}}{\underset{k_{16}}{\rightleftharpoons}} X_{11} \\
X_9 + X_{12} &\overset{k_{17}}{\underset{k_{18}}{\rightleftharpoons}} X_{13} \,.
\end{aligned}$$

Published values for the rate constants can be found in the BioModels database [2]. In particular,

$$\begin{array}{llll} k_1 & = 1.0000 \times 10^{-3} & k_2 & = 33.0000 \\ k_3 & = 3.0000 \times 10^{-3} & k_4 & = 3.8400 \\ k_5 & = 4.5000 \times 10^{-4} & k_6 & = 0.0900 \\ k_7 & = 2.1000 \times 10^{-3} & k_8 & = 12.0000 \\ k_9 & = 1.0000 \times 10^{-3} & k_{10} & = 16.5000 \\ k_{11} & = 1.0000 \times 10^{-3} & k_{12} & = 3.6000 \\ k_{13} & = 4.5000 \times 10^{-4} & k_{14} & = 1.8000 \\ k_{15} & = 1.5000 \times 10^{-3} & k_{16} & = 78.0000 \\ k_{17} & = 2.1000 \times 10^{-4} & k_{18} & = 24.0000 \end{array} \tag{S-3.1}$$

where the forward reaction rates (i.e., the reaction rates with odd subscripts) are measured in $\text{cell}/(\text{molecules} \cdot \text{min})$, whereas, the reverse reaction rates (i.e., the reaction rates with even subscripts) are measured in $1/\text{min}$. Unfortunately, these values do not correspond to a thermodynamically feasible biochemical reaction system, since they do not satisfy the Wegscheider conditions, given by Equation (11) in the Main text.

To determine the Wegscheider conditions associated with the previous model [i.e., to determine matrix $\mathbb{W}$ in Equation (S-1.8) of Additional file 1], we must focus on the stoichiometry matrix $\mathbb{S}$, given by

$$\mathbb{S} = \begin{bmatrix} -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & -1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

We want to find an $M \times M$ permutation matrix $\mathbb{P}_M$ and an $N \times N$ permutation matrix $\mathbb{P}_N$, such that

$$\mathbb{P}_N \mathbb{S}\, \mathbb{P}_M = \begin{bmatrix} \mathbb{S}_{11} & \mathbb{S}_{12} \\ \mathbb{S}_{21} & \mathbb{S}_{22} \end{bmatrix},$$

where $\mathbb{S}_{11}$ is an $M_1 \times M_1$ invertible matrix, as we have discussed in Additional File 1, with $M_1 = \text{rank}(\mathbb{S})$. Clearly, these permutation matrices are not unique. To find appropriate $\mathbb{P}_M$ and $\mathbb{P}_N$, we first use the reduced row echelon form of the stoichiometry matrix $\mathbb{S}$ and discover that the $M_1 = 7$ columns $\{1, 2, 3, 5, 6, 8, 9\}$

are linearly independent, whereas, the remaining two columns $\{4, 7\}$ linearly dependent on the independent columns of $\mathbb{S}$. Therefore, we are looking for a permutation matrix P_M to rearrange $\mathbb{S}$ so that the first M_1 columns of the resulting matrix are linearly independent. To do so, we must set

$$\mathbb{P}_M = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

By following a similar procedure on the rows of $\mathbb{S}\,\mathbb{P}_M$, we find

$$\mathbb{P}_N = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

As a consequence, we can show that

$$(\mathbb{S}_{11}^{-1}\mathbb{S}_{12})^T = \begin{bmatrix} 1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 1 & 0 & 0 \end{bmatrix},$$

which, together with Equations (S-1.8) and (S-1.9) in Additional file 1, implies the Wegscheider conditions $\boldsymbol{\kappa}_f = \mathbb{W}\boldsymbol{\kappa}_d$, where the "free" log-rate constants are given by $\boldsymbol{\kappa}_f = \{\kappa_1, \kappa_3, \kappa_5, \kappa_9, \kappa_{11}, \kappa_{15}, \kappa_{17}, \kappa_7, \kappa_{13}, \kappa_2, \kappa_4, \kappa_6, \kappa_{10}, \kappa_{12}, \kappa_{16}, \kappa_{18}\}$, the "dependent" log-rate constants are given by $\boldsymbol{\kappa}_d = \{\kappa_8, \kappa_{14}\}$, and

$$\mathbb{W} = \begin{bmatrix} -1 & -1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 1 & 1 & 0 & 0 \end{bmatrix}. \qquad \text{(S-3.2)}$$

Since the rate constant values in (S-3.1) are not thermodynamically feasible, we do not use them as the "true" values. Instead, we would like to find thermodynamically feasible parameter values that produce a dynamic behavior that is similar to the one produced by the published infeasible values. The simplest solution would be to use the published values for $\boldsymbol{\kappa}_f$ and calculate new values for $\boldsymbol{\kappa}_d = \{k_8, k_{14}\}$ by means of $\boldsymbol{\kappa}_d = \mathbb{W}\boldsymbol{\kappa}_f$, with $\mathbb{W}$ given by (S-3.2). Unfortunately, this leads to molecular dynamics that are very different from the dynamics produced by the published system. Since the published values produce dynamics that have been validated on experimental data, we must find a more accurate way for determining thermodynamically feasible rate constant values from a set of infeasible published values.

We can address the previous problem by finding a set of free parameters $\boldsymbol{\kappa}_f$ such that

$$\begin{bmatrix} \mathbb{I}_{M+M_1} \\ \mathbb{W} \end{bmatrix} \boldsymbol{\kappa}_f = \begin{bmatrix} \boldsymbol{\kappa}_f^{\text{pub}} \\ \boldsymbol{\kappa}_d^{\text{pub}} \end{bmatrix},$$

where $\mathbb{I}_{M+M_1}$ is the $(M + M_1) \times (M + M_1)$ identity matrix, whereas, $\boldsymbol{\kappa}_f^{\text{pub}}, \boldsymbol{\kappa}_d^{\text{pub}}$ are the published "free" and "dependent" log-rate constant values, respectively. Unfortunately, no such $\boldsymbol{\kappa}_f$ exists since we know that the published values are thermodynamically infeasible. However, we can calculate the best solution to this problem, in a least-squares sense, given by

$$\boldsymbol{\kappa}_f^{\text{true}} = \begin{bmatrix} \mathbb{I}_{M+M_1} \\ \mathbb{W} \end{bmatrix}^{\dagger} \begin{bmatrix} \boldsymbol{\kappa}_f^{\text{pub}} \\ \boldsymbol{\kappa}_d^{\text{pub}} \end{bmatrix},$$

where $\mathbb{A}^{\dagger}$ is the Moore-Penrose pseudoinverse of matrix $\mathbb{A}$, and compute the remaining "dependent" values by setting $\boldsymbol{\kappa}_d^{\text{true}} = \mathbb{W}\boldsymbol{\kappa}_f^{\text{true}}$. As a result, we obtain the following thermodynamically feasible values for the reaction rate constants:

$$\begin{aligned}
k_1 &= 1.4018 \times 10^{-3} & k_2 &= 23.5420 \\
k_3 &= 4.2053 \times 10^{-3} & k_4 &= 2.7394 \\
k_5 &= 2.0388 \times 10^{-4} & k_6 &= 0.1987 \\
k_7 &= 1.4981 \times 10^{-3} & k_8 &= 16.8210 \\
k_9 &= 1.5746 \times 10^{-3} & k_{10} &= 10.4790 \\
k_{11} &= 1.5746 \times 10^{-3} & k_{12} &= 2.2863 \\
k_{13} &= 2.8579 \times 10^{-4} & k_{14} &= 2.8343 \\
k_{15} &= 1.5000 \times 10^{-3} & k_{16} &= 78.0000 \\
k_{17} &= 2.1000 \times 10^{-4} & k_{18} &= 24.0000
\end{aligned} \tag{S-3.3}$$

which we treat as the "true" values.

In Fig. S-3.1, we depict the dynamics of selected molecular species obtained by the published (red curves) and thermodynamically feasible rate values (blue curves). Note that the dynamics do not match perfectly, nor would we expect them to, since the published parameters produce thermodynamically impossible concentration dynamics that a physical system could never produce.

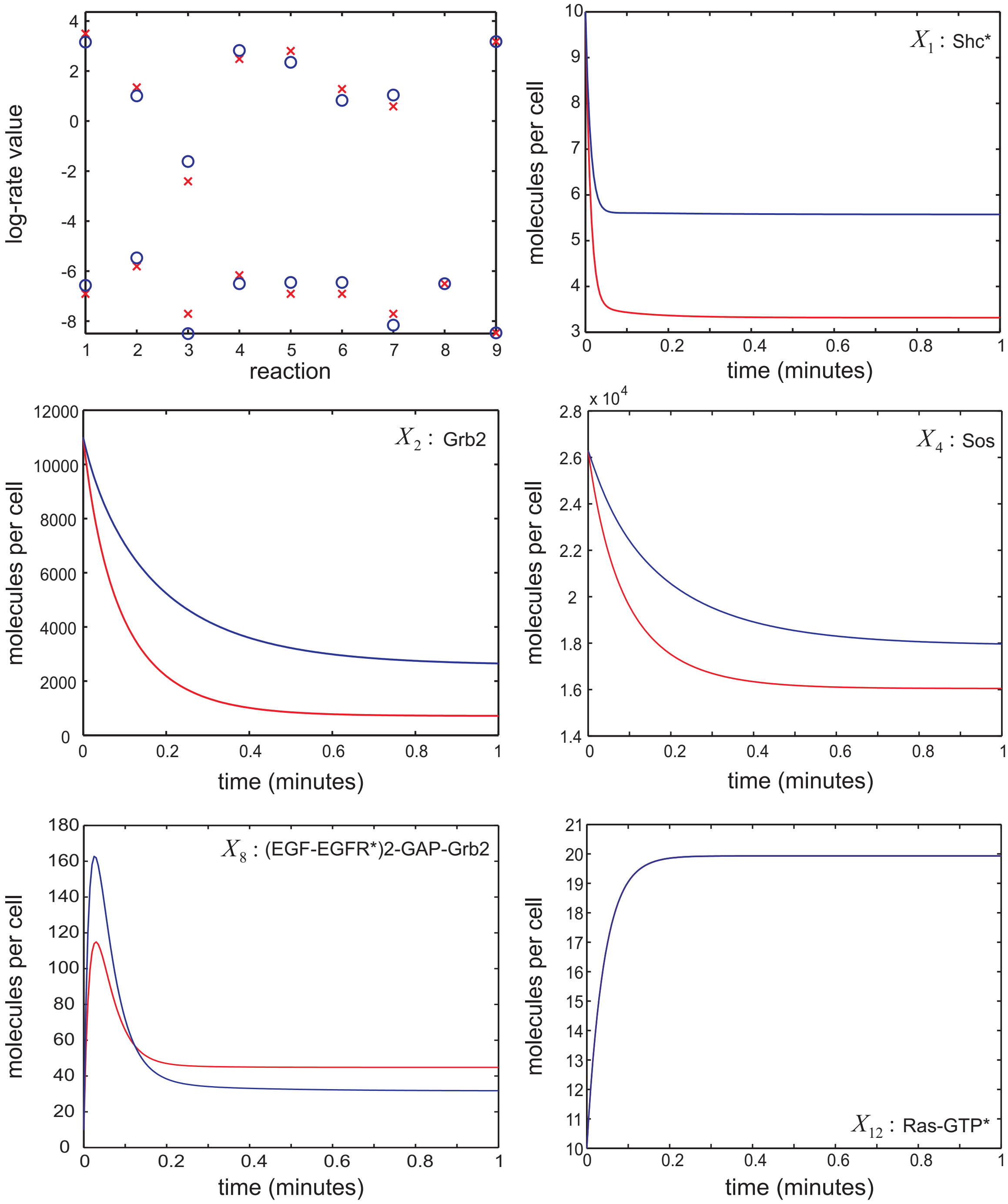

Figure S-3.1: *Published (red) vs. thermodynamically feasible (blue) log-rate values and selected molecular dynamics. Since the published rate values are thermodynamically infeasible, we expect they will result in molecular dynamics that could not be possibly produced by a real biological system. As a consequence, we do not expect perfect match between the "red" and "blue" curves.*

Published values for the initial concentrations of the molecular species can also be found in [2]. Based on these values, we set

$$\begin{aligned}
c_1 &= 10 \\
c_2 &= 11{,}000 \\
c_3 &= 10 \\
c_4 &= 26{,}300 \\
c_5 &= 10 \\
c_6 &= 40{,}000 \\
c_7 &= 1{,}000 \\
c_8 &= 10 \\
c_9 &= 10 \\
c_{10} &= 72{,}000 \\
c_{11} &= 10 \\
c_{12} &= 10 \\
c_{13} &= 10\ ,
\end{aligned} \tag{S-3.4}$$

measured in molecules/cell. To compensate for the fact that our biochemical reaction system does not model the entire EGF/ERK signaling cascade, we must account for the upstream EGF stimulus. To do so, we increase the initial concentration of the most upstream molecular species in our model, namely $X_7 =$ (EGF-EGFR*)2-GAP, from 0 in [2] to 1,000 molecules/cell. Finally, we increase the initial concentrations of X_1, X_3, X_5, X_8, X_9, X_{11}, X_{12}, and X_{13} from 0 in [2] to 10, to take into account that, in a real cellular system, these molecular species are constitutively expressed.

References

1. Schoeberl B, Eichler-Jonsson C, Gilles ED, Müller G: **Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors**. *Nat. Biotechnol.* 2002, **20**:370–375.

2. Li C, Donizelli M, Rodriguez N, Dharuri H, Endler L, Chelliah V, Li L, He E, Henry A, Stefan MI, Snoep JL, Hucka M, Novère NL, Laibe C: **BioModels database: An enhanced, curated and annotated resource for published quantitative kinetic models**. *BMC Syst. Biol.* 2010, **4**:92.