Modeling of biocatalytic reactions: A workflow for model calibration, selection and validation using Bayesian statistics

Ina Eisenkolb, Antje Jensch, Kerstin Eisenkolb, Andrei Kramer,
Patrick C. F. Buchholz, Jürgen Pleiss, Antje Spiess, Nicole E. Radde

November 8, 2019

Supporting Information

1 Model calibration and evaluation

In the following we provide details on the model calibration to experimental data and additional analysis methods for model evaluation.

1.1 Optimization details

For model calibration we generated parameter samples from the posterior distribution employing the Markov Chain Monte Carlo (MCMC) method as described in the main paper. In order to evaluate the likelihood function $L_D(\theta)$, all simulations of all three models were performed via `Matlab 2017b` (64 bit). For model handling we used the `SBTOOLBOX2` and `SBPD` toolboxes from the `SBPOP` package (`http://www.sbtoolbox2.org/main.php`)[1], which make use of the CVODE solver from `SUNDIALS` for integration.

As further options we set boundaries for the parameters. We used a logarithmic scale for all parameters, which allows to cover many orders of magnitude. As a prior step we performed maximum likelihood optimization by minimizing the negative log-likelihood in order to find good starting points for the sampling procedure. Initial boundaries were set to $[-5, 5]$. In order to avoid local optima, optimization was performed with 1000 latin hypercube samples as starting values. The initial conditions A_0, P_0 and E were set according to the conditions specified for each experiment in Table 1 in the main paper. For optimization we utilized the interior-point algorithm of `fmincon`, a built-in minimizer of `Matlab`. Settings were set to default values.

For the subsequent MCMC sampling the `mcmcstat` toolbox (`https://mjlaine.github.io/mcmcstat/`) was used. The sampling method was set to `dram`, representing the Delayed Rejection Adaptive Metropolis algorithm[2] with the covariance adaptation interval `MCMCoptions.adaptint=20·number of parameters`. We initialized four independent chains starting from the 1st, 5th, 10th and 15th best parameter vectors of the prior optimization step. To achieve convergence the sampling was split in a warm-up phase of $5 \cdot 10^5$ steps and the main run consisting of 10^6 samples. In order for the final MAP estimator not to lie on one of the boundaries, we repeated the sampling procedure for Model 1 and Model 2 with adjusted bounds as listed in Figure 4A and Figure 5A, respectively. The rejection rates of all chains are between 35-40% and 78-96% for Model 1 and 2, respectively. The exact values are listed in Table **S1**.

Convergence of individual chains was tested with the Geweke convergence diagnostic, which compares the sample variance of the first 10% with the last 50% of the chain. All chains passed the convergene test with a p-value of at least 0.7 after partial sample cut-off for chain 4 in case of Model 1 and chain 1 and 2 in case of Model 2. Overall chain convergence was assessed using the Gelman-Rubin-Brooks diagnostic via the function `mpsrf`, which returns a potential scale reduction factor R. The diagnostic values $R = 1.0010$ and $R = 1.0294$ for Model 1 and 2, respectively, suggest also overall chain convergence of the merged shortened chains. Both diagnostics are implemented in `mcmcstat`.

Table S1: Average rejection rates of the MCMC sampling procedure for Model 1 and Model 2.

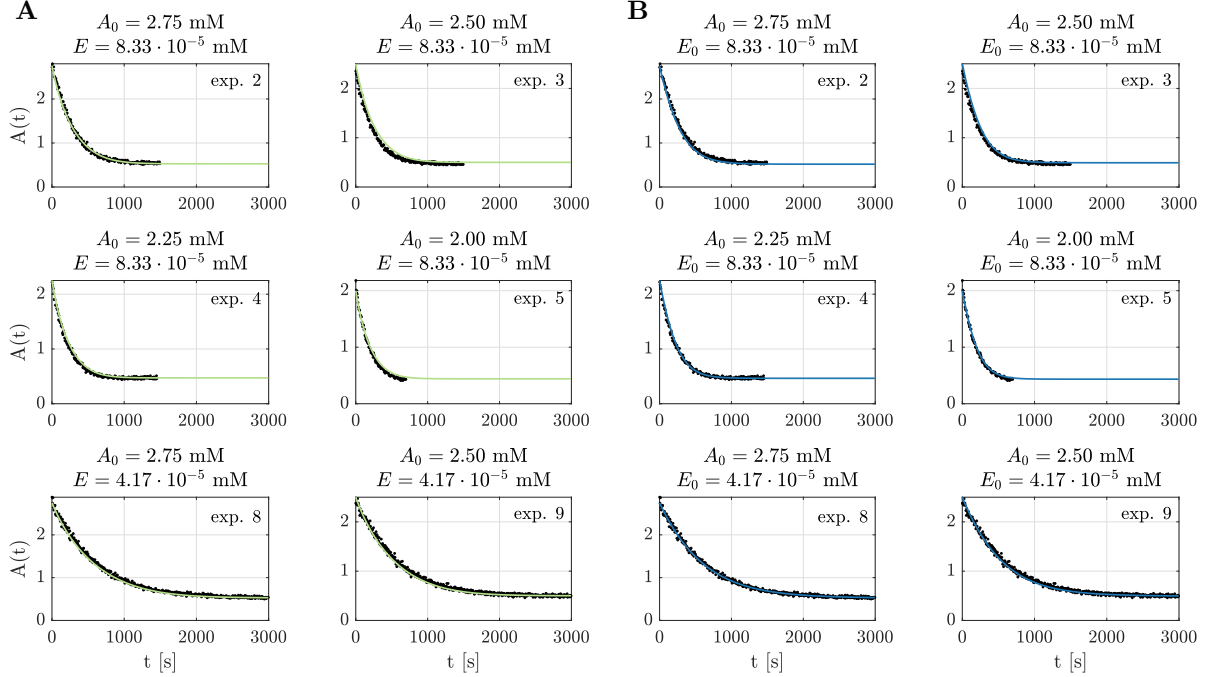|	Rejection rate			
	Chain 1	Chain 2	Chain 3	Chain 4
Model 1	0.3515	0.3655	0.3920	0.3637
Model 2	0.8780	0.9305	0.7803	0.9586

Fig S1: **Additional fits for (A) Model 1 and (B) Model 2.** A comparison of model trajectories simulated with the MAP estimator $\widehat{\theta}^{\mathrm{MAP}}$ and experimental time courses for six initial conditions (experiments 2, 3, 4, 5, 8 and 9).

We used more than $3 \cdot 10^5$ parameter sets, uniformly sampled from the merged MCMC chain, to evaluate the model fit and to have an estimate of the resulting uncertainty for each model. Model fits of Model 1 and 2 for experiments not shown in the main paper are presented in Figure **S1**.

1.2 Profile likelihood analysis

Parameter identifiablity was examined via profile likelihood analysis. In order to obtain the profile likelihood, one parameter is fixed to specific values while the other model parameters are reoptimized with respect to the experimental data. Profile likelihood analysis was performed for all parameters of Model 1 and Model 2 using the Matlab toolbox `d2d`[3], which employs a scaled version of our likelihood function Eq (5). We used default settings with the exception of the number of sampling steps which was adapted for each parameter, such that, when possible, confidence interval lengths could be calculated. Confidence interval lengths listed in Figure 6C represent the range of parameter values such that the likelihood value in that parameter range remain below the 95% quantile of the chi-square distribution based on 1 degree of freedom. That threshold is depicted as the upper dashed black line in the profile likelihoods in Figures 6A,B and **S2**. The MAP, recalculated using the likelihood function of `d2d`, is represented as a black star. Below the profile likelihood for each parameter we also show the recalibrated values of the other remaining parameters.
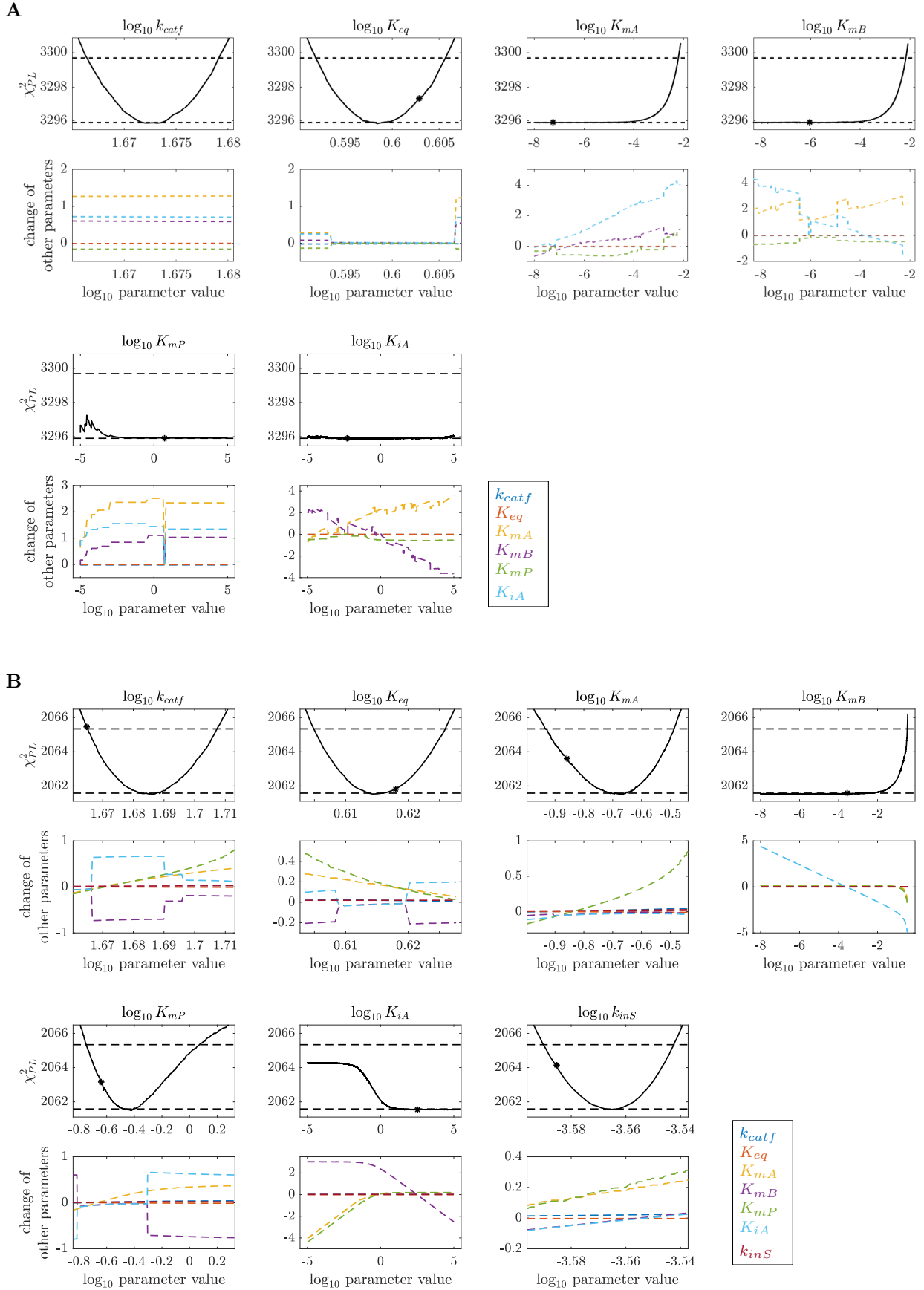
Fig S2: **Additional information on profile likelihood analysis for Model 1 (A) and Model 2 (B).** Shown are the MAP estimate (black star) and changes of fitted parameters.
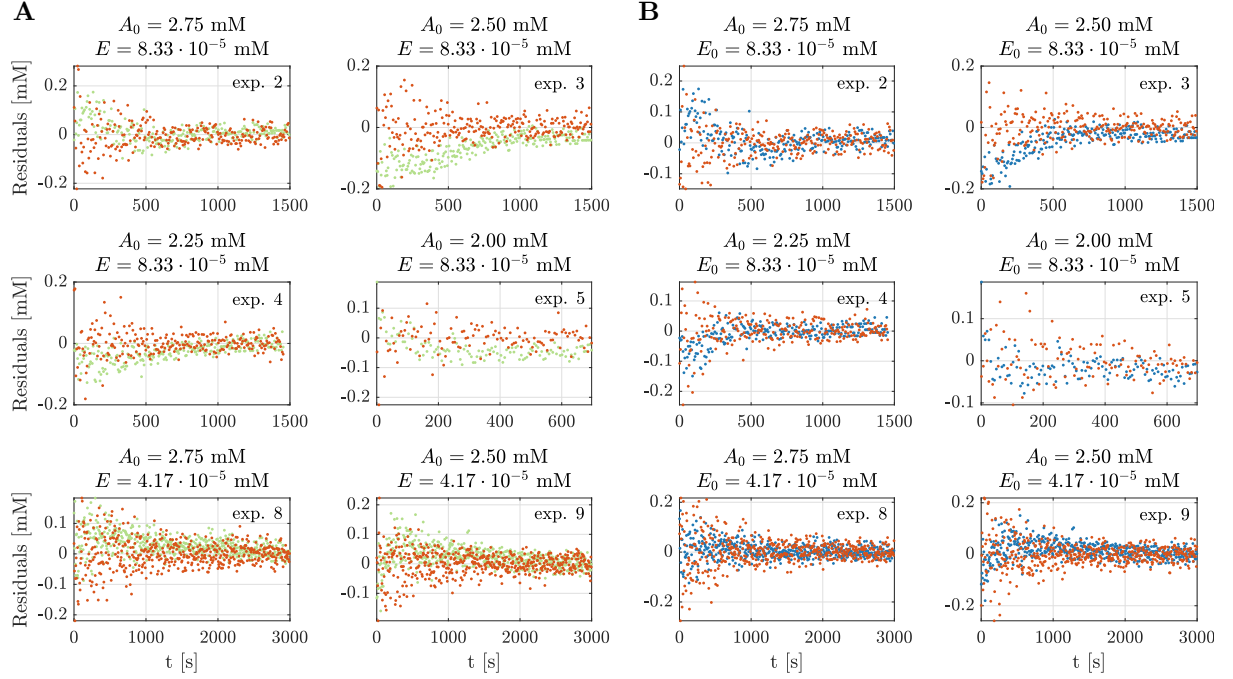
Fig S3: **Additional information on residual analysis for Model 1 (A) and 2 (B).** Red dots indicate residuals simulated with the MAP estimator and the proposed multiplicative noise model for the six experiments not shown in the main paper. (A) Green (Model 1) and (B) blue (Model 2) dots are the residuals resulting from a comparison between the MAP trajectory and the experimental data.

1.3 Residual analysis

Goodness of fit was evaluated by residual analysis, which was performed for trajectories of the MAP estimator $\widehat{\theta}_{\mathrm{MAP}}$. Residuals R_{MAP} were determined for each time point $t_k, k \in 1, \ldots, T$ in all experiments $j \in 1, \ldots, 9$ by

$$R^j_{\mathrm{MAP}}(t_k) = y^j_A(t_k) - z^j_A(t_k, \widehat{\theta}_{\mathrm{MAP}}), \tag{1}$$

where $z^j_A(t_k, \widehat{\theta}_{\mathrm{MAP}})$ refers to the solution of the differential equation for substrate A and $y^j_A(t_k)$ denotes the respective measurement. For comparison, we also drew samples $s^j_A(t_k)$ from the log-normal distribution

$$s^j_A(t_k) \sim \log \ \mathcal{N}(\log \ z^j_A(t_k, \widehat{\theta}_{\mathrm{MAP}}), \ \sigma^2) \tag{2}$$

and evaluated the residuals R_{Sample}

$$R^j_{\mathrm{Sample}}(t_k) = s^j_A(t_k) - z^j_A(t_k, \widehat{\theta}_{\mathrm{MAP}}). \tag{3}$$

The resulting residuals for all experiments for Model 1 (green) and Model 2 (blue) are shown in Figure 8 and Figure **S3**. Residuals of the samples R_{Sample} are depicted in orange.

In order to evaluate the error over all experiments combined, we calculated the mean squared error (MSE) for the data and the samples

$$\mathrm{MSE}_{\mathrm{MAP}}(t_k) = \frac{1}{9} \sum_{j=1}^{9} \left(R^j_{\mathrm{MAP}}(t_k) \right)^2 \tag{4}$$

$$\mathrm{MSE}_{\mathrm{Sample}}(t_k) = \frac{1}{9} \sum_{j=1}^{9} \left(R^j_{\mathrm{Sample}}(t_k) \right)^2. \tag{5}$$

Considering that the measured time frames differ for the experiments, only the first 117 data points of each experiment were used, which corresponds to the longest time frame where measurements for all experiments are available. The result is shown in the fourth row of Figure 7 in the main paper.
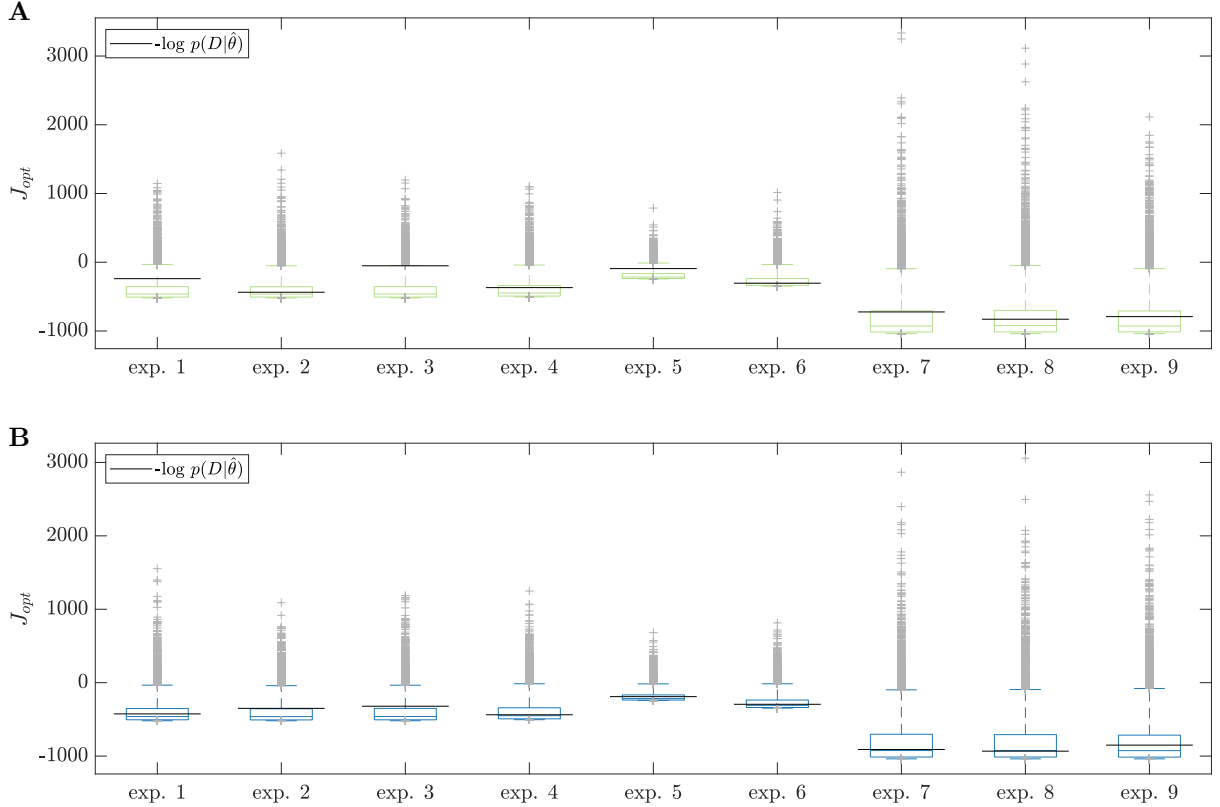
A

B

Fig S4: **Additional information on bootstrap analysis for Model 1 (A) and 2 (B).** For each experiment a boxplot of J_{opt} has been generated by a bootstrap analysis, in which many datasets and respective likelihood function values were simulated with (A) Model 1 and (B) Model 2 for all experiments. Resulting data were used to estimate the median (black line), the 25th and 75th percentiles (blue boxes) and the 5th and 95th percentiles (lower and upper adjacent values, respectively). Outliers are depicted as crosses. Horizontal lines depict J_{opt} values with respect to the real experiments.

1.4 Model validation via bootstrapping

Plausibility of all models was tested with a parametric bootstrapping approach (see e.g. [4]), in which we generated many datasets from the stochastic model. These were subsequently used to calculate the likelihood function values J obtained by comparison with the simulation of the MAP estimator. For this purpose, we resampled D_i, $i = 1, \ldots, 10000$ datasets that mimic experimental data used in the study (i.e. same number of time points, same conditions etc.). Then we calculated the likelihoods $p(D_i|\widehat{\theta}^{\mathrm{MAP}})$ and used these values to infer the median as well as the 5th, 25th, 75th and 95th percentiles. Figure 7 (last row) shows the overall result for all experiments for Model 1 and Model 2. Results for each experiment are presented in Figure **S4**.

2 Model 3: Modeling and model calibration

For model calibration and subsequent evaluation we followed the same procedure as for Model 1 and Model 2. The MAP estimator lies well within the bounds of -5 to 5, hence there was no need to adjust the bounds for Model 3. The average rejection rates of the chains are listed in Table **S2**. Convergence of individual chains was again tested with the Geweke convergence diagnostic, which all chains passed with a p-value of at least 0.7 after partial sample cut-off for all chains. Convergence assessment using the Gelman-Rubin-Brooks diagnostic suggest also overall chain convergence of the merged shortened chains of Model 3 with a diagnostic value $R = 1.0016$.

Figure **S5** summarizes the results for Model 3. Shown are the remaining model trajectories (A), the

Table S2: Average rejection rates of the MCMC sampling procedure for Model 3.

	Rejection rate			
	Chain 1	Chain 2	Chain 3	Chain 4
Model 3	0.2359	0.2644	0.2390	0.9432

Table S3: Best objective function values of the ML method and MCMC sampling together with sampling times as well as AIC and BIC values for Model 1, Model 2 and Model 3.

	Best J_{opt} of ML method	Best J of MCMC sampling	MCMC sampling time [h]	AIC	BIC
Model 1	-3819.4	-3834.1	7.01	-7656.2	-7620.6
Model 2	-4701.4	-4710.2	12.50	-9406.4	-9364.9
Model 3	7200.5	-1561.6	11.40	-3113.2	-3083.6

boxplots for the bootstrap analysis for each experiment (B), remaining residual analysis plots (C) and results of the profile likelihood analysis (D).

3 Model comparison

In order to compare all three models we employed the Akaike Information Criterion (AIC)[5] and the Bayes Information Criterion (BIC)[6]. As can be seen in Table **S3**, both criteria favour Model 2. Additionally, we also compare the best J_{opt} value of the ML method with the best J value of the MCMC sampling for all three models. In all three cases, the sampling procedure is able to find a better optimum than the ML method. Finally, we show the required time for MCMC sampling. Model 2 which showed the best results has the longest sampling time. Model 3 has a similar samling time despite fitting two parameters less. The fastest run time is achieved by Model 1.

References

1. Schmidt H, Jirstrand M. Systems Biology Toolbox for MATLAB: a computational platform for research in systems biology. Bioinformatics 2005 11;22(4):514–515. `https://doi.org/10.1093/bioinformatics/bti799`.

2. Haario H, Laine M, Mira A, Saksman E. DRAM: efficient adaptive MCMC. Statistics and computing 2006;16(4):339–354.

3. Raue A, Steiert B, Schelker M, Kreutz C, Maiwald T, Hass H, et al. Data2Dynamics: a modeling environment tailored to parameter estimation in dynamical systems. Bioinformatics 2015;31(21):3558–3560.

4. Efron B. Bayesian inference and the parametric bootstrap. Ann Appl Stat 2012;6(4):1971–97.

5. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control 1974 December;19(6):716–723.

6. Schwarz G. Estimating the Dimension of a Model. Ann Statist 1978 03;6(2):461–464. `https://doi.org/10.1214/aos/1176344136`.
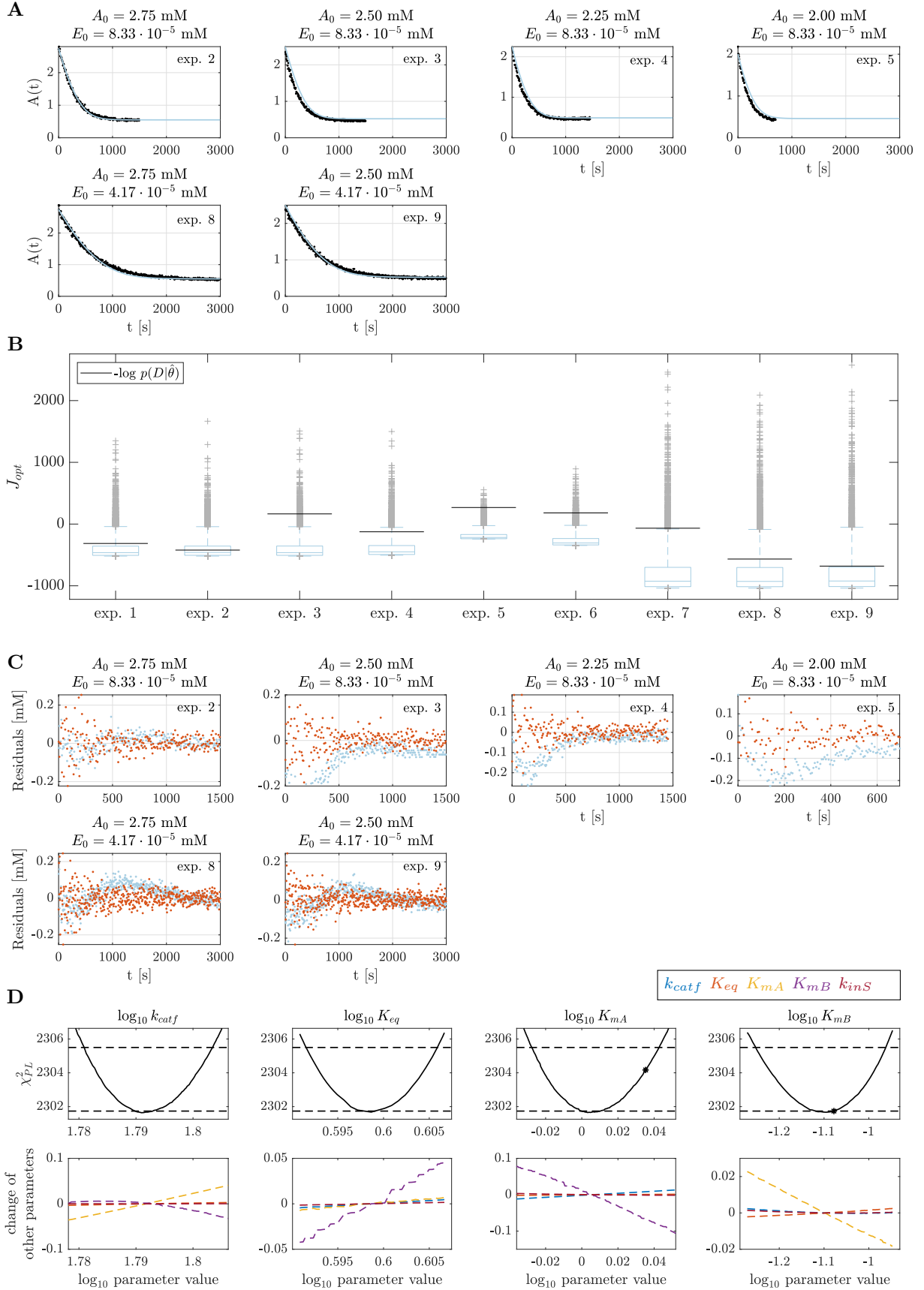
Fig S5: **Model evaluation Model 3.** (A) Model trajectories (light blue) with experimental data (black dots), (B) boxplots of bootstrap results for each experiment (blue horizontal lines represent the 5th, 25th, median, 75th and 95th percentile), (C) residual plots depicting the residuals of the MAP estimator from the data (light blue) and the residuals of the MAP estimator from the samples (red) and (D) profile likelihood analysis with resulting changes in other parameters. The black star denotes the MAP estimate.