Trajectory-oriented Bayesian experiment design versus Fisher A-optimal design: an in depth comparison study

Patrick Weber[1,*], Andrei Kramer[1], Clemens Dingler[1] and Nicole Radde[1]

[1]Institute for Systems Theory and Automatic Control, University of Stuttgart, Pfaffenwaldring 9, Stuttgart 70550, Germany

ABSTRACT

Motivation: Experiment design strategies for biomedical models with the purpose of parameter estimation or model discrimination are in the focus of intense research. Experimental limitations such as sparse and noisy data result in unidentifiable parameters and render-related design tasks challenging problems. Often, the temporal resolution of data is a limiting factor and the amount of possible experimental interventions is finite. To address this issue, we propose a Bayesian experiment design algorithm to minimize the prediction uncertainty for a given set of experiments and compare it to traditional A-optimal design.

Results: In an in depth numerical study involving an ordinary differential equation model of the *trans*-Golgi network with 12 partly non-identifiable parameters, we minimized the prediction uncertainty efficiently for predefined scenarios. The introduced method results in twice the prediction precision as the same amount of A-optimal designed experiments while introducing a useful stopping criterion. The simulation intensity of the algorithm's major design step is thereby reasonably affordable. Besides smaller variances in the predicted trajectories compared with Fisher design, we could also achieve smaller parameter posterior distribution entropies, rendering this method superior to A-optimal Fisher design also in the parameter space.

Availability: Necessary software/toolbox information are available in the supplementary material. The project script including example data can be downloaded from http://www.ist.uni-stuttgart.de/%7eweber/BayesFisher2012.

Contact: patrick.weber@ist.uni-stuttgart.de

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Regulation models are preferably used to describe intra- and inter-cellular interactions of biomolecules. The Biomodel Database of the European Bioinformatics Institute comprises over 400 published and curated models in its repository, most of them being ordinary differential equation (ODE) models (Li *et al.*, 2010). The usefulness of a biomedical ODE model is often assessed through its capability to predict possible scenarios of interest. In order to calibrate a biological ODE model, expensive and time-consuming experiments are performed to gain the needed experimental training data. Despite continuously improving measurement methods, the data for most models remain scarce, resulting in practically non-identifiable parameters (Gutenkunst *et al.*, 2007). To support the

learning process, several experiment design methods have been successfully developed including classical Fisher information matrix (FIM) design (Bandara *et al.*, 2009), Bayesian methods (Chaloner and Verdinelli, 1995; Kramer and Radde, 2010; Wilkinson, 2007) and optimization-based methods (Kreutz *et al.*, 2011). Claiming that ODE model predictions improve when reducing parameter uncertainty intervals, these methods design optimal experiments for parameter estimation (OED/PE). However, models can be used to make precise predictions despite sloppy parameters with large confidence intervals (Brown *et al.*, 2004; Klinke, 2009). This is rendering OED/PE an indirect method to improve predictions. Methods have been developed to directly address experiment design for better model predictions (Casey *et al.*, 2007; Vanlier *et al.*, 2012), which we here refer to as OED/MP. In this work, we also propose to go this direct way, by predefining sets of experimentally feasible prediction scenarios of interest. We use a Bayesian posterior inference method and predict experimentally feasible scenarios to purposively reduce uncertainty in the model trajectories. In addition, we define a reasonable stopping criterion to avoid further experimentation when the desired predictions meet a certain precision compared with the expected data quality. The method is successfully applied to reduce prediction uncertainty of an ODE model of secretory pathway control at the *trans*-Golgi network. In an intense comparison study, we were able to outperform A-optimal designed experiments in both prediction uncertainty and posterior distribution entropy, with affordable computational effort.

2 METHODS

System

In this work, we address the problem of efficiently selecting experiments to improve the prediction capabilities of a given ODE model of the following form:

$$\dot{x}(t; u, \theta) = f(x(t; u, \theta), u(t), \theta) \qquad x(0; u, \theta) = x_0(u, \theta)$$
$$y(t; u, \theta) = h(x(t; u, \theta), u(t), \theta), \qquad (1)$$

with state $x(t; u, \theta) \in \mathbb{R}_+^{n_x}$, input $u(t) \in \mathbb{R}_+^{n_u}$, parameter $\theta \in \mathbb{R}_+^{n_\theta}$ and output $y(t; u, \theta) \in \mathbb{R}_+^{n_y}$. The vector field f is assumed to be continuously differentiable to guarantee the existence of a unique solution. For intracellular models, f is often derived by using chemical reaction kinetics. The output mapping h defines the measurable outputs as functions of model states. To include typical experimental pretreatments, we allow the initial conditions $x_0(u, \theta)$ to depend on the system parameters θ and initial experimental influences u. The system parameter vector θ consists of positive rate coefficients, Michaelis Menten parameters or degradation and synthesis rates. Given positivity, these constants are assumed to be $\log_{10}$-transformed, $\theta = \{\theta_1, \ldots, \theta_{n_\theta}\} = \{\log_{10}(k_1), \ldots, \log_{10}(k_{n_\theta})\}$, to guarantee numerical stability in subsequently introduced numerical computations.

---

*To whom correspondence should be addressed.

Experiments The goal of improving the model prediction is restricted to a predefined set of experiments $\{\mathcal{E}^i\}_{i=1}^{n_\mathcal{E}}$. In this work, we consider each prediction scenario also a potential experiment. Each experiment index i is connected to an experimental setup. An experiment $\mathcal{E}^i$ is defined by an experimentally feasible input vector $u^i(t)$ and a set of measurable quantities $y^i = (y_1^i, \ldots, y_{n_y^i}^i)$. Typical experiments include for example initial changes of $u^i(t)$ at t_0 describing vector overexpression or siRNA experiments. Also popular are specific changes at a predefined time instance that refer to culture medium changes or external stimuli. Each measurable quantity $y_j^i, j = 1, \ldots, n_y^i$ can be measured at one discrete time point t_j^i within the set $t_j^i \in \mathcal{T} = \{t_0, t_0 + \Delta t, t_0 + 2\Delta t, \ldots, t_{\mathrm{end}}\}$.

Data and likelihood function We assume that the measurement process is prone to measurement noise, and we choose a log-normal error model here, in accordance with recent studies (Gassmann *et al.*, 2009; Kreutz *et al.*, 2007). Thus, whenever a measurement is performed, the true system is assumed to provide a data point of the form

$$\forall i, j : z_j^i(t_j^i) = y_j^i(t_j^i; u^i, \theta^*)\eta, \tag{2}$$

with $\eta \sim \log\mathcal{N}(0, \sigma^{*2})$, θ^* denoting the true system parametrization and σ^* denoting the true error model parameter. For simplicity, we assume the same parameter σ for all outputs and experiments, which is no general restriction for our proposed method. We collect the data from experiment i in the set $\mathcal{D}^i = \{z_j^i(t_j^i)\}_{j=1,\ldots,n_y^i}$. Given this error model and assuming independence of all measured data points, the likelihood function for the system states

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{n_\mathcal{E}} p(\mathcal{D}^i|\theta)$$
$$= \prod_{i=1}^{n_\mathcal{E}} \prod_{j=1}^{n_y^i} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2} \left(\frac{\tilde{z}_j^i(t_j^i) - \tilde{y}_j^i(t_j^i; u^i, \theta)}{\sigma} \right)^2 \right\}, \tag{3}$$

with the log transformations $\tilde{y}_j^i(t_j^i) = \log(y_j^i(t_j^i; u^i, \theta))$ and $\tilde{z}_j^i(t_j^i) = \log(z_j^i(t_j^i))$. The maximum-likelihood estimate (MLE) for the parameters is given by maximizing the likelihood function:

$$\hat{\theta}^{\mathrm{ml}} = \arg\max_\theta p(\mathcal{D}|\theta). \tag{4}$$

2.1 Trajectory-oriented Bayesian design

The objective function in the Bayesian framework is the posterior distribution $p(\theta|\mathcal{D})$, which is a distribution over the parameters θ after having seen the data. Values of the posterior distribution can be calculated using Bayes' theorem:

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}, \tag{5}$$

where $P(\theta)$, the prior distribution, reflects our prior assumptions about the parameters. $P(\theta)$ can be used to restrict the search space to regions of plausible values. $P(\theta|\mathcal{D})$ can be investigated through Markov chain Monte Carlo sampling, and we use an adaptive Metropolis algorithm for this purpose.

In the following, we denote the representative sample for the current posterior distribution $\{\theta^s : s = 1, \ldots, N\}$ or short $\{\theta^S\}$. Each sample step requires evaluation of the likelihood function, for which the ODE system has to be solved numerically. Thus, together with a sample $\{\theta^S\}$, we get automatically for each experiment i and each measurable output y_j^i a set of trajectories $\{y_j^i(t; \theta^s, u^i) : s = 1, \ldots, N\}$. From those trajectories, we can estimate variances according to

$$\widehat{V}_j^i(t_j^i) = \frac{1}{N-1} \sum_{\theta \in \{\theta^S\}} \left(y_j^i(t_j^i; \theta, u^i) - \hat{\bar{y}}_j^i(t_j^i; u^i) \right)^2, \tag{6}$$

with estimated mean

$$\hat{\bar{y}}_j^i(t_j^i; u^i) = \frac{1}{N} \sum_{\theta \in \{\theta^S\}} y_j^i(t_j^i; \theta, u^i) \tag{7}$$

for each i, j and each time point $t_j^i \in \mathcal{T}$. In the first step, we determine the set of measurement times $\hat{t}_j^i$ that maximize the expected variances in the trajectories

$$\hat{t}_j^i = \arg\max_{t_j^i \in \mathcal{T}} \widehat{V}_j^i(t_j^i), \tag{8}$$

and collect the set of respective variance estimates in the sets $\widehat{V}^i = \{V_j^i(\hat{t}_j^i)\}_{j=1,\ldots,n_y^i}$. These sets are the basis for our experiment selection procedure. At this point, we introduce a stopping criterion: measurement on y_j^i in experiment i is performed only if the respective variance $\widehat{V}_j^i$ is still above a certain threshold, which we set to the pooled empirical variance estimate $\hat{\sigma}^2$ here. This means that we propose to measure only if we can still expect a decrease in $\widehat{V}_j^i$ even when we take the precision $\hat{\sigma}^{-2}$ of the measurement process into account. To address this stopping criterion mathematically, we set $\widehat{V}^{i,*} = \{\widehat{V}_j^{i,*} := \max(\widehat{V}_j^i - \hat{\sigma}^2, 0)\}$. The successor experiment $\hat{\imath}$ is selected by maximizing the sum of expected variances within each experiment

$$\hat{\imath} = \arg\max_i \sum_{j=1,\ldots,n_y^i} \widehat{V}_j^{i,*}. \tag{9}$$

Once $\hat{\imath}$ has been determined, we suggest to measure at time instances $\hat{t}_j^{\hat{\imath}}$ for which $\widehat{V}_j^{\hat{\imath},*} \neq 0$.

To illustrate this approach, we refer to a fetch-ahead of our numerical study given in Figure 1, which shows the sets of trajectories $\{y_j^i(t; \theta^s, u^i)\}$ of the posterior sample $\{\theta^S\}$ for each measurable quantity y_j^i after a first experiment was performed (upper row), along with the proposed measurement time instances $\hat{t}_j^i$, which are indicated with vertical lines here, the estimated variances $\widehat{V}_j^i$ and the current values of the pooled measurement error estimates $\hat{\sigma}^2$. The second line shows predictions subject to the current sample $\{\theta^S\}$ for a different experiment. In this case, the algorithm would decide to measure all quantities y_j^i once again, since the stopping criterion is fulfilled for none of them.

2.2 FIM experiment design

To compare our proposed Bayesian experiment selection to a computationally inexpensive and well-established experiment design routine, we shortly introduce the well-known FIM based experiment design routine (Atkinson and Donev, 1992). The improvement of FIM optimal experiment design for time-series measurements in biomedical model discrimination (OED/MD), OED/PE or the combination of both disciplines is in the focus of recent publication series (Donckels *et al.*, 2012, 2009), rendering it a solid reference method. In order to apply the method, the FIM must be evaluated at the current MLE $\hat{\theta}^{\mathrm{ml}}$ for the current index set $\mathcal{I}_p$ of all already performed experiments:

$$F(\hat{\theta}^{\mathrm{ml}}) = \sum_{i \in \mathcal{I}_p} F^i(\hat{\theta}^{\mathrm{ml}}) \tag{10}$$

$$= \sum_{i \in \mathcal{I}_p} \sum_{j=1}^{n_y^i} \frac{1}{\sigma^2} \left[\nabla_\theta \tilde{y}_j^i(t_j^i; u^i, \theta) \right]^T \Big|_{\hat{\theta}^{\mathrm{ml}}} \left[\nabla_\theta \tilde{y}_j^i(t_j^i; u^i, \theta) \right] \Big|_{\hat{\theta}^{\mathrm{ml}}}.$$

The superposition principle of the FIMs holds for experiments, outputs and measurement time points. In the next step, anticipatory update FIMs $\bar{F}_j^i(\hat{\theta}^{\mathrm{ml}}, t_j^i)$ for all candidate experiments $i = 1, \ldots, n_\mathcal{E}$ and $j = 1, \ldots, n_y^i$ are evaluated at the current MLE. The update FIMs $\bar{F}_j^i(\hat{\theta}^{\mathrm{ml}}, t_j^i)$ are added individually to the current FIM, forming the overall predicted FIMs $\hat{F}_j^i(\hat{\theta}^{\mathrm{ml}}, t_j^i) = F(\hat{\theta}^{\mathrm{ml}}) + \bar{F}_i^j(\hat{\theta}^{\mathrm{ml}}, t_j^i)$. A design criterion $\Phi(\hat{F}_j^i(\hat{\theta}^{\mathrm{ml}}, t_j^i))$ has to be applied to the update FIMs to decide upon the successor experiment for the current design step. Under the threat of non-identifiable parameters, we decided to apply the modified A-optimal criterion $\Phi(\hat{F}_j^i(\hat{\theta}^{\mathrm{ml}}, t_j^i)) =$
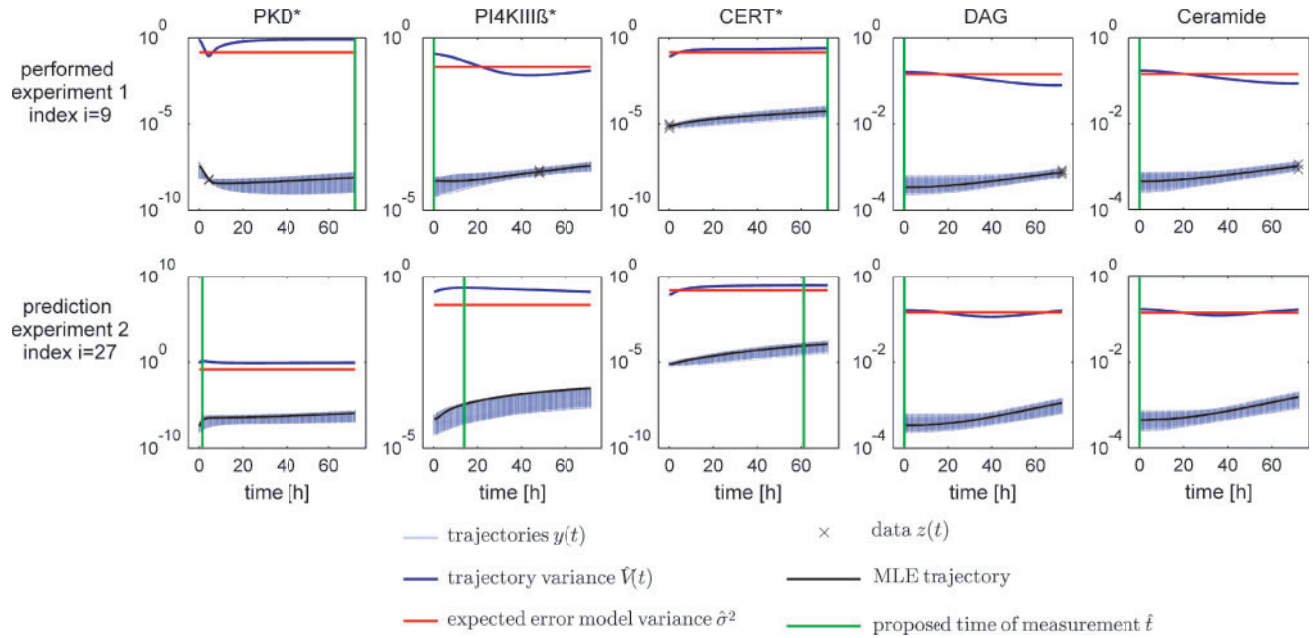
Fig. 1. Predicted trajectories for the experiment outcome after an initial training experiment is performed. Trajectories are depicted together with the trajectories' empirical variance. The first row depicts the model fit for the initial experiment. The measurement time instances suggested by our design algorithm are depicted by vertical lines. The second row evinces the proposed measurements for the second successor experiment. For experimental inputs corresponding to experiment index i, we refer to Supplementary Table S3

$\mathrm{tr}(\hat{F}_j^i(\hat{\theta}^{\mathrm{ml}}, t_j^i))$ to completely avoid the inversion of close to singular FIMs. For fairness of comparison, we need to have the same amount of overall measurements m for the next experiment, which is determined by the Bayesian design. If m outputs are to be measured at yet unknown t_j^i, these are chosen according to

$$\hat{t}_j^i = \arg\max_{t_j^i} \mathrm{tr}(\hat{F}_j^i(\hat{\theta}^{\mathrm{ml}}, t_j^i)), \tag{11}$$

$$\mathcal{A}(K_m^i) = \sum_{j \in K_m^i} \mathrm{tr}(\hat{F}_j^i(\hat{\theta}^{\mathrm{ml}}, \hat{t}_j^i)), \tag{12}$$

$$\widehat{K}_m^i = \arg\max_{K_m^i} \mathcal{A}(K_m^i), \tag{13}$$

$$\hat{i} = \arg\max_i \mathcal{A}(\widehat{K}_m^i), \tag{14}$$

where K_m^i denotes the index subset of the set of all measurable outputs y_j^i with m elements. The successor output set $\hat{K}_m^{\hat{i}}$ and time points $\hat{t}_j^{\hat{i}}$ contain all needed information to plan the experiment. We note that other FIM design criteria such as minimizing the determinant of the inverse matrix are more involved computationally, since the superposition principle does not hold any more and maximization cannot be decoupled any longer.

3 DESIGN OF THE STUDY

3.1 A model for secretion control at the *trans*-Golgi network

Protein secretion in mammalian cells is a highly regulated process, in which the regulation of the formation of transport vesicles at the *trans*-Golgi network plays a crucial role. Understanding these regulatory processes is for example important for further optimization of producer cell lines, where it has been shown that for

high messenger RNA (mRNA) copy numbers, there are bottlenecks further downstream, and one example is the *trans*-Golgi network (Becker *et al.*, 2008). Vesicle formation is mediated by a network of interacting proteins and lipids, involving the protein kinase D (PKD), the ceramide tranfer protein CERT, the lipid kinase PI4KIIIβ, ceramide and diacylglycerol (DAG). Figure 2 shows a scheme of our regulatory network model. DAG is located in the Golgi membrane and recruits and thereby activates PKD (Bard and Malhotra, 2006). Active PKD has a dual effect on CERT transport activity: on the one hand, it directly phosphorylates and thereby deactivates CERT (Fugmann *et al.*, 2007), on the other hand, it activates CERT indirectly through activating PI4KIIIβ, which triggers synthesis of phosphatidyl-inositol-4-phosphate (PI4P). PI4P enables binding of CERT to the TGN and release of ceramide (Fugmann *et al.*, 2007). The feedback loop between CERT and PKD is closed by conversion of phosphatidylcholine (PC) and ceramide to sphingomyelin (SM) at the Golgi, where DAG occurs as a byproduct. Both PKD and DAG have also been shown to play a direct role in vesicle formation (Bard and Malhotra, 2006; Hausser *et al.*, 2005). Although the main interactions are known, the kinetics of this system has not yet been investigated quantitatively, and there are currently no appropriate experimental data available for model fitting. Thus, we use artificial data here.

A differential equation model for the regulation process was established which involves 9 molecular species, 17 chemical reactions, 12 interaction parameters and 17 further parameters such as basal production and degradation rates. Up to five model outputs are measurable upon request, namely active PKD (PKD*), active PI4KIIIβ (PI4KIIIβ*), active CERT (CERT*), overall DAG and overall ceramide. The four remaining model status variables are
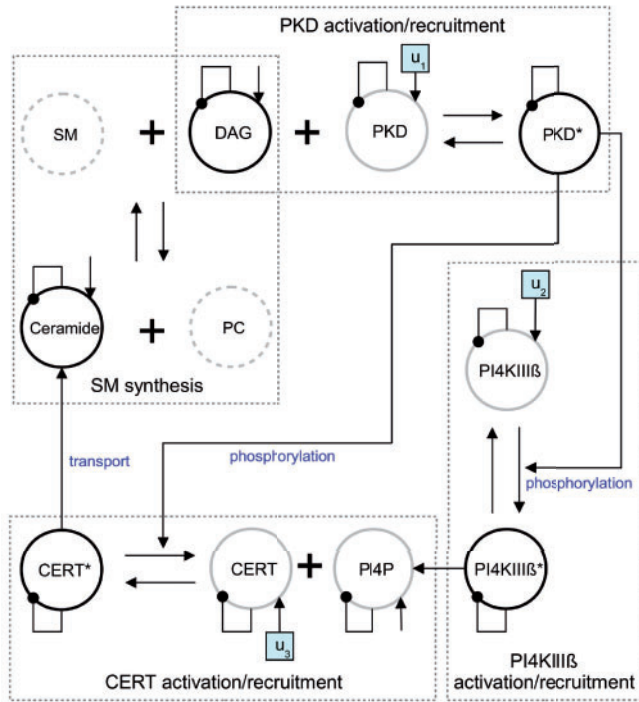
Fig. 2. Model for secretion control at the *trans*-Golgi network. Grey nodes denote latent variables, black nodes measurable species. Active species are denoted with an asterisk. Self-loops describe linear degradation terms. Dashed boxes depict turnover reactions. The inputs 'u' indicate experimentally possible perturbations of synthesis rates. Blue text explains biological processes depicted by corresponding edge. SM and PC are not contained in the model but shown for completeness of the SM synthesis reaction

defined as latent variables. Equations and parameter values are listed in Supplementary Appendix.

3.2 Experiment design testing

In order to test our newly suggested algorithm, we performed a biologically motivated *in silico* study. Our Bayesian and a classical A-optimal design routine are allowed to make suggestions from a set of 27 experiments, which describe all possible perturbation permutations of the three species that can experimentally be accessed through overexpression or siRNA silencing, namely PKD, CERT and PI4KIIIβ. The valid experiment time horizon has been set to $t_{end} = 72h$, which is a realistic time frame for a cell culture experiment with proliferation and confluence effects. In the case of silencing, the respective basal expression rate is damped by a factor of 10, while in the vector overexpression case, the basal rate is 10-fold higher. These input perturbation parameter values for siRNA silencing and vector overexpression are experimentally motivated by typical suppression rates (Pillai *et al.*, 2005). Although the Bayesian approach can in principle incorporate input uncertainty, this is not considered in this comparison study to avoid shifting the focus.

We assume all 12 interaction parameters to be initially unknown and choose a bounded log uniform prior supporting four orders of magnitude for each of them. Parameters are estimated and sampled

in logarithmic space, while basal production and degradation rates are set to biologically plausible values (Eden *et al.*, 2011). To account for limitations in the quantity of sampling time instances in protein quantification (e.g. western blotting: 18 lanes per gel), both algorithms were allowed to demand for a maximum of one data triplet at one sampling time instance for each output. Thus, if for example data for all five outputs are requested this would result in 15 data points in total. The requested data are generated by the unknown true system and noise corrupted using a log-normal error model with $\sigma^* = 0.15$. The exercise is stopped as soon as one algorithm trained the model in a way that it fulfills our stopping criterion. Details on the choice of the joint initial experiment are given in the Supplementary Material. To make the study more realistic, we estimate the variance $\hat{\sigma}^2$ in each simulation scenario using a pooled variance estimate. In order to compare the two algorithms, the 12-dimensional parameter posterior distributions have been estimated through Markov chain Monte Carlo (MCMC) sampling after each newly obtained dataset. For their computation, the sampling algorithms of the MCMC toolbox (Haario *et al.*, 2006) have been combined with the fast ODE integrators of the SBtoolbox2 (Schmidt and Jirstrand, 2006) and the parallel computing capabilities of Matlab. Model prediction and entropy estimation in the following sections are based on these posterior distributions. For the prediction of our proposed Bayesian method, 1000 parameters were drawn from the posterior distribution according to the description in 'Methods' section, and all 27 experiments are simulated. The time horizon of 72 h was divided into 105 equidistant discrete sampling times, leaving realistic 40 min of reaction time for an experimentalist between two subsequent wet lab sample preparation procedures. The whole design exercise has been repeated five times for each algorithm.

4 RESULTS

4.1 Bayesian experiment design results in twice the prediction quality than A-optimal design

Figure 3 depicts the expected uncertainty of the predictions, represented here as the overall mean of all expected variances $\hat{V}_j^i(t_j^i)$ after being trained with experiments selected by the competing algorithms. We additionally averaged over five runs. After each new training experiment, a prediction for all experiments has been performed. The stopping criterion was reached by the Bayesian routine after four experiments in all five runs, while requesting a total of 12 data points. This means that we expect the trained model to be able to make predictions for all 27 experiments in which all trajectory variances are below the threshold.

The evolution of the mean trajectory variance $\bar{v}$, averaged over all 27 experiments, outputs and time points, is depicted in Figure 3. It reaches a value of 0.06 after four experiments for the Bayesian design, making the model prediction twice as good as the A-optimal trained model with a value of 0.12. After the terminal experiment $\hat{\sigma} = 0.15 \pm 0.2$ was close to σ^* (see Supplementary Appendix).

4.2 Bayesian experiment design gains more information about parameters than A-optimal design

Since A-optimal design reduces the overall spread of the posterior distribution, while the proposed method reduces the observable
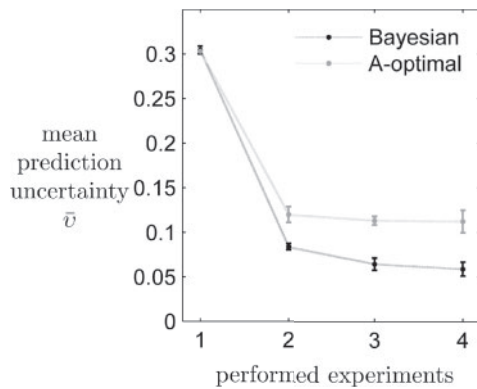
Fig. 3. Mean expected prediction uncertainty for all 27 experiments after the model has been trained with up to four experiments chosen by the Bayesian or the A-optimal design criterion. We additionally averaged over five different runs
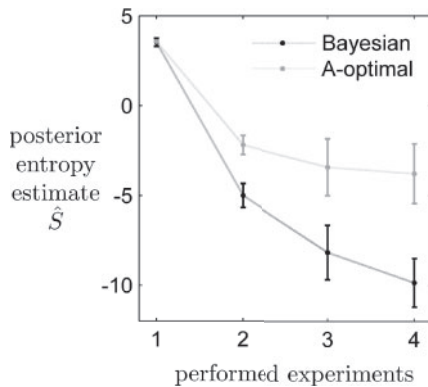

Fig. 4. Mean posterior distribution entropy estimates after the model has been trained with up to four experiments chosen by the Bayesian or the A-optimal design criterion, respectively. Five runs of the overall experiment design exercise have been performed

trajectory variance, we should also consider the posterior size in a fair comparison. The entropy S reflects the spread of a probability distribution regardless of shape complexity. Therefore, we include a numerical approximation of the entropy given by

$$\hat{S} = -\frac{1}{N} \sum_{\theta \in \{\theta^S\}} \ln P(\theta|\mathcal{D}) \approx S = E[-\ln P(\theta|\mathcal{D})] \qquad (15)$$

in this analysis as a measure for the information content of the posterior sample of the parameters. Figure 4 depicts the evolution of the Shannon entropy estimated from the posterior samples using a parallel multivariate kernel density estimator. Our prediction-based Bayesian design routine achieved an average entropy estimate of -10 nats, which is 8 nats lower than the A-optimal experiments with an average value of -2 nats. Our design is superior to A-optimal design already after the second experiment.

As entropy is an abstract measure, an additional visual insight is given by Figure 5, which shows a comparison of the marginal posterior distributions for three parameters of one exemplary design run. We have exemplary chosen a parameter that is expected to be very well identifiable after the design exercise with four
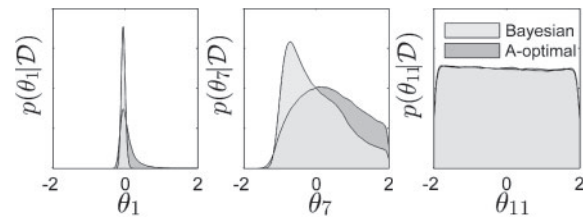

Fig. 5. Estimated marginal distributions of three parameters of the posterior sample after the model has been trained with four experiments chosen by the Bayesian or the A-optimal design criterion. Y-axis limits denote the boundaries of the log uniform prior

Table 1. Qualitative properties of the model parameters

Algorithm	Two boundaries	One boundary	Prior bounded
Bayes design	5.6 ± 0.5	4.2 ± 0.7	2.2 ± 0.4
A-optimal design	4.6 ± 1.4	3.6 ± 1	3.8 ± 1.2

Average number of parameter boundaries distinct from the initially set prior boundaries established after the fourth experiment in five runs. The categories two boundaries, one boundary and prior bounded parameters are determined by visually examining the marginal distributions (Figure 5).

successive experiments, one that is vaguely identifiable and one that is expected to be not identifiable at all. Axis ranges have been set equal to the prior ranges. We have also visually inspected all marginal distributions of five different design runs and categorized the established parameter boundaries in Table 1. Although this is not an established quantitative measure as, e.g. entropy, it is a good visualization for the differences of the two designs in the parameter space.

In our example, the parameter θ_1 describes recruitment and thereby activation of PKD by DAG. The flux for this reaction depends in our model linearly on both variables. Furthermore, PKDa and DAG are both measurable, such that it is intuitive that θ_1 becomes identifiable. In the five runs, the Bayesian design routine established an average of one more double bounded parameters such as θ_1, than A-optimal design. Double bounded parameters with tight confidence intervals are most useful to draw biological conclusions, as they restrict model fluxes. As each flux stands for a different effect (e.g. transport or activation), the relative importance of these effects becomes comparable.

The parameter θ_7 describes recruitment of CERT to the TGN through PI4P and its activation. Although CERT can be disturbed and is also directly measurable, both are not the case for PI4P, and so the experiments are less informative for this parameter. Still, an average of 0.6 more one-sided parameter boundaries such as the lower bound of θ_7 are identified by the proposed method. With the same line of arguments, upper or lower bounded parameters are useful to identify, e.g. a tendency towards higher or lower influence of a modelled effect.

Finally, θ_{11} describes conversion of DAG to ceramide. Although both DAG and ceramide are observables, this parameter is hardly identifiable, since the flux of this reaction is small in our model and both variables cannot be directly perturbed. Bayesian design ended up with an average of 1.6 less of these uninformative parameters. An example plot of all 12 marginal distributions of the parameter vector
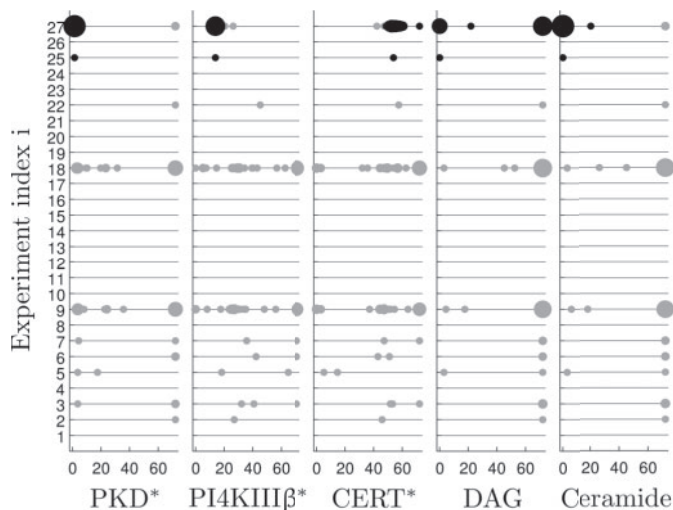
Fig. 6. Repeated decision making ($n = 100$) of the Bayesian (black) and the A-optimal Fisher (grey) experiment design. Big spots indicate more frequently chosen combinations of experiments and sampling times

is given in Supplementary Figure S2. Summarizing it can be said that the Bayesian design enables a more detailed interpretation of the model with the same amount of data measured by more efficiently restricting the models parameter values.

4.3 Bayesian design makes more consistent decisions

Figures 3 and 4 indicate that the highest incremental improvement in predictions and entropy is achieved after the second experiment. We used this particular design step to further investigate differences in decision making of the competing algorithms. We tested two potential weak points of each method. Fisher design is assumed to be highly dependent on the MLE. Therefore, we repeated the second design step 100 times, while each time re-estimating the MLE. Bayesian design is limited by the amount of representative posterior trajectories that can be simulated for decision making; therefore, we redraw 100 times the $N = 1000$ prediction trajectories from the posterior distribution. The statistic for the proposal of the second experiment is depicted in Figure 6. Trajectory-based Bayesian design prefers (beside one single outlier) experiment 27, overexpression of all three species, while merely altering in sampling time. A-optimal decisions are mostly bi-modal between experiments 9 and 18, in which PI4KIIIβ and CERT are overexpressed, while about 15% of the decisions are scattered over various experiments and time points. It is clearly visible that the suggestions of the Bayesian design are much more consistent.

4.4 Computational effort

The main computational effort of this study stems from the sampling of the posterior distribution, which serves as a common criterion to compare the two different methods. This involved over a quarter of a billion ODE integrations and a detailed Markov chain convergence analysis. Parallel Markov chains have been merged for each posterior distribution estimate containing at least 3.5 million sample points each. Not converged sub-chains have been discharged. Note that experiments are typically scheduled on a weekly basis, so that differences in computation time might not

be crucial. The posterior sampling has to be performed only after new data is received from the wet lab. For a fair comparison of the computational effort of the two design routines, we focus on the number of necessary ODE integrations needed for the actual design or proposal steps which are, respectively, used to assess new potential experiments. The calculation of the A-optimality of a candidate experiment with a 12×12 FIM (simple finite differences) requires 48 model simulations; the Bayesian concept required 1000 simulations per prediction, rendering it 20 times more simulation intensive than the Fisher design in this example. Still, any state of the art Desktop PC should be able to assess 100–1000 experiments per hour using the Bayesian method.

5 DISCUSSION AND CONCLUSION

In this work, we introduced a trajectory-based Bayesian experiment design approach and compared this with A-optimal Fisher design for a regulatory network of transport vesicle formation at the *trans*-Golgi network in mammalian cells with 12 parameters. In this example, which contains several sloppy parameters, Bayesian design clearly outperformed Fisher design with respect to reducing uncertainty in predictions and also in the parameter space.

Our results demonstrate that design criteria that rely on local approximations of the posterior distribution are not suited in case of sloppy or non-identifiable parameters. Fisher design, for example, assumes that the posterior distribution can be approximated by a multivariate Gaussian about its maximum. In our study, we could further identify vulnerable dependence of FIM design decision making to the obviously non-reliable MLE. Since the appearance of sloppy parameters is ubiquitous for intracellular regulation networks, we strongly encourage to use and further develop methods that are based on output or trajectory optimization rather than parameter identification.

One of the reviewers brought up the important point of model errors, which leads to further discrepancies between model simulations and 'real' observations. This study uses artificial data that are generated by the model under consideration, which has of course the advantage that all true parameters are known and results are controllable. However, we are aware that this simplifies analyses a lot compared with more realistic settings with real experimental data. For example, if the model is erroneous and does not really capture the dynamics of the underlying processes, this might lead to likelihood functions for different datasets that are not consistent with respect to the model and favour different regions in the parameter space, which complicates the PE step and thus also the whole experiment design method. This fact has for sure to be taken into account when working with real data and can for example be addressed by a design strategy that is optimized to discriminate between different model hypotheses, which is one of the main issues for our future work in this project.

When summarizing our numerical results, we became aware of a very recent study on prediction-based experiment design (Vanlier *et al.*, 2012). The authors focus on posterior predictive distributions, which assign probabilities to new simulation scenarios. Although this is an elegant approach in general, calculating this distribution requires the evaluation of further integrals and thus becomes rapidly computationally expensive for high-dimensional design spaces, e.g. if multiple candidate measurements are planned within a single experiment.

Overall, model-based experiment design that optimizes prediction uncertainty is a challenging new field, which is very suitable especially for intracellular network models with sparse data and should further be investigated.

6 AUTHOR CONTRIBUTIONS

P.W. and A.K. developed the trajectory-based Bayesian design algorithm. P.W. implemented the algorithm and created the results with support from A.K. C.D. created results for the Fisher design. P.W., A.K. and N.R. developed the model and wrote the manuscript. All authors read and approved the final manuscript.

REFERENCES

Atkinson,A. and Donev,A. (1992) *Optimum Experimental Designs*. Oxford Statistical Science Series, vol. 8, Oxford: Clarendon Press.

Bandara,S. *et al.* (2009) Optimal experimental design for parameter estimation of a cell signaling model. *PLoS Comput. Biol.*, **5**, e1000558.

Bard,F. and Malhotra,V. (2006) The formation of TGN-to-plasma-membrane transport carriers. *Annu. Rev. Cell Dev. Biol.*, **22**, 439–455.

Becker,E. (2008) An xbp-1 dependent bottle-neck in production of igg subtype antibodies in chemically defined serum-free chinese hamster ovary (cho) fed-batch processes. *J. Biotechnol.*, **135**, 217–223.

Brown,K.S. *et al.* (2004) The statistical mechanics of complex signaling networks: nerve growth factor signaling. *Phys. Biol.*, **1**, 184–195.

Casey,F.P. *et al.* (2007) Optimal experimental design in an epidermal growth factor receptor signalling and down-regulation model. *IET Syst. Biol.*, **1**, 190–202.

Chaloner,K. and Verdinelli,I. (1995) Bayesian experimental design: a review. *Stat. Sci.*, **10**, 273–304.

Donckels,B.M. *et al.* (2012) Performance assessment of the anticipatory approach to optimal experimental design for model discrimination. *Chemomet. Intell. Lab. Syst.*, **110**, 20–31.

Donckels,B.M.R. *et al.* (2009) A kernel-based method to determine optimal sampling times for the simultaneous estimation of the parameters of rival mathematical models. *J. Comput. Chem.*, **30**, 2064–2077.

Eden,E. *et al.* (2011) Proteome half-life dynamics in living human cells. *Science*, **331**, 764–768.

Fugmann,T. *et al.* (2007) Regulation of secretory transport by protein kinase d-mediated phosphorylation of the ceramide transfer protein. *J. Cell Biol.*, **178**, 15–22.

Gassmann,M. *et al.* (2009) Quantifying western blots: pitfalls of densitometry. *Electrophoresis*, **30**, 1845–1855.

Gutenkunst,R.N. *et al.* (2007) Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput. Biol.*, **3**, 1871–1878.

Haario,H. *et al.* (2006) Dram: efficient adaptive mcmc. *Stat. Comput.*, **16**, 339–354.

Hausser,A. *et al.* (2005) Protein kinase D regulates vesicular transport by phosphorylation and activation of phosphatidylinositol-4 kinase III β at the Golgi. *Nat. Cell Biol.*, **7**, 880–886.

Klinke,D.J. (2009) An empirical bayesian approach for model-based inference of cellular signaling networks. *BMC Bioinformatics*, **10**, 371.

Kramer,A. and Radde,N. (2010) Towards experimental design using a Bayesian framework for parameter identification in dynamic intracellular network models. *Procedia Comput. Sci.*, **1**, 1639–1647.

Kreutz,C. *et al.* (2007) An error model for protein quantification. *Bioinformatics*, **23**, 2747–2753.

Kreutz,C. (2011) Likelihood based observability analysis and confidence intervals for predictions of dynamic models. ArXiv e-prints, 1107.0013.

Li,C. *et al.* (2010) Biomodels database: an enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst. Biol.*, **4**, 92.

Pillai,R.S. *et al.* (2005) Inhibition of translational initiation by let-7 microrna in human cells. *Science*, **309**, 1573–1576.

Schmidt,H. and Jirstrand,M. (2006) Systems biology toolbox for matlab: a computational platform for research in systems biology. *Bioinformatics*, **22**, 514–515.

Vanlier,J. *et al.* (2012). A bayesian approach to targeted experiment design. *Bioinformatics*, **28**, 1136–1142.

Wilkinson,D.J. (2007) Bayesian methods in bioinformatics and computational systems biology. *Brief. Bioinform.*, **8**, 109–116.