Design of Experiments for Model Discrimination
Hybridising Analytical and Data-Driven Approaches

Simon Olofsson 1  Marc Peter Deisenroth 1 2  Ruth Misener 1

Abstract
Healthcare companies must submit pharmaceutical drugs or medical devices to regulatory bodies before marketing new technology. Regulatory bodies frequently require transparent and interpretable computational modelling to justify a new healthcare technology, but researchers may have several competing models for a biological system and too little data to discriminate between the models. In design of experiments for model discrimination, the goal is to design maximally informative physical experiments in order to discriminate between rival predictive models. Prior work has focused either on analytical approaches, which cannot manage all functions, or on data-driven approaches, which may have computational difficulties or lack interpretable marginal predictive distributions. We develop a methodology introducing Gaussian process surrogates in lieu of the original mechanistic models. We thereby extend existing design and model discrimination methods developed for analytical models to cases of non-analytical models in a computationally efficient manner.

1. Introduction
A patient’s health and safety is an important concern, and healthcare is a heavily regulated industry. In the USA and the EU, private and public regulatory bodies exist on federal/union and state levels (Field, 2008; Hervey, 2010). Healthcare companies applying to market a new drug or medical device must submit extensive technical information to the regulatory bodies. In the USA and the EU, the Food & Drug Administration (FDA) and European Medicines Agency (EMA) handle these applications, respectively. The FDA require that the technical information contains e.g. the chemical composition of the drug, how the drug affects the human body (pharmacodynamics), how the body affects the drug (pharmacokinetics), and methods for drug manufacturing, packaging and quality control (U.S. Food & Drug Administration, 2018). Likewise, applying to market a new medical device requires submitting extensive technical information. A medical device can be any appliance, software or material used for non-pharmacological diagnosis, monitoring or treatment of a disease or condition (European Union, 2017). An important part of proving the efficacy and safety of a new drug or medical device is showing that its effects on the patient can be predicted and interpreted, e.g. via mathematical and computational models.

These transparency and interpretability requirements, combined with limited amounts of available experimental data, make models learnt solely from observed data unsuitable for proving the efficacy and safety of a new drug or medical device to the regulatory bodies. Hence researchers use explicit parametric models. For drugs, the pharmacokinetics (how the human body affects the drug) is often modelled using systems of ordinary differential equations. These models elucidate how the drug is absorbed and distributed through the body (e.g. brain, kidneys and liver) under different dosing profiles and methods of drug administration (e.g. orally or intravenously) (see Figure 1). The EMA pharmacokinetic model guidelines state that regulatory submissions should include detailed descriptions of the models, i.e. justification of model assumptions, parameters and their biochemical plausibility, as well as parameter sensitivity analyses (European Medicines Agency, 2011, p. 16; 2016, p. 3).

A key problem is to identify a mathematical model that usefully explains and predicts the behaviour of the pharmacokinetic process (Heller et al., 2018). Researchers will often have several models (i.e. hypotheses) about an underlying mechanism of the process, but insufficient experimental data to discriminate between the models (Heller et al., 2018). Additional experiments are generally expensive and time-consuming. Pharmacokinetic experiments typically take 8–24 hours for mice and rats, and 12–100 hours for simians and humans (Ogungbenro & Aarons, 2008; Tuntland et al.,...
2. Model Discrimination

Model discrimination aims to discard inaccurate models, i.e., hypotheses that are not supported by the data. Assume $M > 1$ models $\mathcal{M}_i$ are given, with corresponding model functions $f_i(x, \theta_i)$. The model is a hypothesis and collection of assumptions about a biological system, e.g. the pharmacokinetic model in the top left of Figure 1, and the function is the model’s mathematical formulation that can be used for predictions, e.g. the drug concentrations on the right in Figure 1. Each model function $f_i$ takes as inputs the design variables $x \in X \subset \mathbb{R}^d$ and model parameters $\theta_i \in \Theta_i \subset \mathbb{R}^{D_i}$. The design variables $x$, e.g. drug administration and dosing profile, specify a physical experiment that can be run on the system described by the models, with $X$ defining the operating system boundaries. The classical setting allows tuning of the model parameters $\theta_i$ to fit the model to data. The system has $E \geq 1$ target dimensions, i.e. $f_i : \mathbb{R}^{d+D_i} \rightarrow \mathbb{R}^E$. We denote the function for each target dimension $e$ as $f_{i,(e)}$. This notation distinguishes between the $e$th target dimension $f_{i,(e)}$ of a model function $f$, and model $\mathcal{M}_i$’s function $f_i$. Table A in the supplementary material summarises the notation used.

We start with an initial set of $N_0$ experimental measurements $y_n = f_{\text{true}}(x_n) + \epsilon_n$, $n = 1, \ldots, N_0$. A common assumption in pharmacokinetics is that the measurement noise term $\epsilon_n$ is i.i.d. zero-mean Gaussian distributed with covariance $\Sigma_{\text{exp}}$ (Steimer et al., 1984; Ette & Williams, 2004; Tsimring, 2014). Skew in the data distribution can be handled, e.g. through log-transformation. The experimental data set is denoted $\mathcal{D}_{\text{exp}} = \{y_j, x_j\}$. The initial $N_0$ experimental data points are insufficient to discriminate

Figure 1. Different aspects of pharmacokinetic models. (Top left) simple pharmacokinetic model with four compartments: digestive system, liver, bloodstream and other tissue. The drug can enter as tablets into the digestive system or as an injection into the bloodstream, and leave the system by being excreted as waste. (Bottom left) the effect on drug concentration in a patient from two different dosing profiles: (i) half the dose twice per day, or (ii) the full dose once per day. (Centre) two different methods of drug administration: orally in the form of tablets, or intravenously as an injection, and (Right) their different effects: an injection has a quicker effect than tablets.

2014). It is key to design experiments yielding maximally informative outcomes with respect to model discrimination.

Contribution We tackle limitations in existing analytical and data-driven approaches to design of experiments for model discrimination by bridging the gap between the classical analytical methods and the Monte Carlo-based approaches. These limitations include the classical analytical methods’ difficulty to manage non-analytical model functions, and the computational cost for the data-driven approaches. To this end, we develop a novel method of replacing the original parametric models with probabilistic, non-parametric Gaussian process (GP) surrogates learnt from model evaluations. The GP surrogates are flexible regression tools that allow us to extend classical analytical results to non-analytical models in a computationally efficient manner, while providing us with confidence bounds on the model predictions.

The method described in this paper has been implemented in a python software package, GPdoemd (2018), publicly available via GitHub under an MIT License.
between the models, i.e. $\forall i : p(M_i \mid D_{\text{exp}}) \approx 1/M$. We are agnostic to the available experimental data, and wish to incorporate all of it in the model discrimination.

### 2.1. Classical Analytical Approach

Methods for tackling the design of experiments for discriminating simple, analytical models have been around for over 50 years. Box & Hill (1967) study the expected change $E[\Delta S] = E[S_{N+1}] - S_N$ in Shannon entropy $S_N = \sum_i \pi_i \log \pi_i$ from making an additional experimental observation, where the posterior model probability $\pi_{i,N+1} = p(y|M_i)\pi_i/N$ from making an additional experimental observation, where the posterior model probability $\pi_{i,N+1} = p(y|M_i)\pi_i,N/p(y)$ and $p(y) = \sum_j p(y|M_i)\pi_{i,N}$. Box & Hill (1967) develop a design criterion $D_{\text{BH}}$ by maximising the upper bound on $E[\Delta S]$. The expression for $D_{\text{BH}}$ can be found in the supplementary material. MacKay (1992a;b) derives a similar expression. Box & Hill (1967) choose the next experiment by finding $x^* = \arg \max_x D_{\text{BH}}$. Design of experiments continues until there exists a model probability $\pi_{N,i} \approx 1$.

Buzzi-Ferraris et al. (1990) propose a new design criterion $D_{\text{BF}}$ (see the supplementary material) from a frequentist point-of-view. They suggest using a $\chi^2$ test with $NE - D_i$ degrees of freedom for the model discrimination. The null hypothesis for each model is that the model has generated the experimental data. Design of experiments continues until only one model passes the $\chi^2$ test. Models are not ranked against each other since Buzzi-Ferraris et al. (1990) argue this simply leads to the least inaccurate model being selected. The $\chi^2$ procedure is more robust against—but not immune to—favouring the least inaccurate model.

Michalik et al. (2010) proceed from the Akaike information criterion (AIC) as the model discrimination criterion to derive a design criterion $D_{\text{AW}}$ from the Akaike weights $w_i$. The expression for $D_{\text{AW}}$ can be found in the supplementary material. Design of experiments continues until there exists an Akaike weight $w_i \approx 1$. Box & Hill (1967) and Michalik et al. (2010) implicitly favour the least inaccurate model.

In order to account for the uncertainty in the model parameters $\theta$, the classical analytical approach (Prasad & Someswara Rao, 1977; Buzzi-Ferraris et al., 1984) is to approximate the model parameters as being Gaussian distributed $N(\theta, \Sigma_\theta)$ around the best-fit parameter values $\hat{\theta}$. The covariance $\Sigma_\theta$ is given by a first-order approximation

$$\Sigma_{\theta}^{-1} = \sum_{n=1}^N \nabla_{\theta} f_i^n \Sigma_{\exp}^{-1} \nabla_{\theta} f_i^n, \quad (1)$$

where $f_i^n = f_i(x_n, \theta)$, $x_n \in D_{\exp}$, and the gradient $\nabla_{\theta} f_i = \partial f_i(x, \theta)/\partial \theta_{i} \mid_{\theta = \hat{\theta}}$. The predictive distribution with $\theta$ approximately marginalised out becomes $p(f_i \mid x, D_{\exp}) = N(f_i(x, \theta), \Sigma_i(x))$, where the covariance is $\Sigma_i(x) = \nabla_{\theta} f_i \Sigma_{\theta} \nabla_{\theta} f_i$.

**Limitations** Most model functions for healthcare applications such as pharmacokinetics or medical devices are not analytical. They are often, from a practical point-of-view, complex black boxes of legacy code, e.g., large systems of ordinary or partial differential equations. We can evaluate the model function $f_i$, but the gradients $\nabla_{\theta} f_i$ are not readily available. Specialist software can retrieve the gradients $\nabla_{\theta} f_i$ from systems of ordinary or partial differential equations, but this requires implementing and maintaining the models in said specialist software packages. Other types of mathematical models may require solving an optimisation problem (Boukouvala et al., 2016). For model discrimination, we wish to be agnostic with regards to the software implementation or model type, since this flexibly (i) allows faster model prototyping and development, and (ii) satisfies the personal preferences of researchers and engineers.

### 2.2. Bayesian Design of Experiments

Methods for accommodating non-analytical model functions have developed in parallel with increasing computer speed. These methods are typically closer to fully Bayesian than the classical analytical methods. Vanlier et al. (2014) approximate the marginal predictive distributions of $M$ models and their Jensen-Shannon divergence using Markov chain Monte Carlo (MCMC) sampling and a $k$-nearest neighbours density estimate. On the downside, the density estimates become less accurate as the number of experimental observations increases (Vanlier et al., 2014) and the method is computationally intensive (Ryan et al., 2016).

Instead of studying design of experiments for model discrimination, statisticians have focused on design of experiments and model discrimination separately (Chaloner & Verdinelli, 1995). Design of experiments can also be used to aid model parameter estimation. They solve problems of the type

$$x^* = \arg \max_x E_{y_{N+1}} \left[ U(y_{N+1}, x, \theta) \right] \quad (2)$$

where the utility function $U(\cdot, \cdot, \cdot)$ serves to either discriminate models or estimate parameters. Criteria for model discrimination are handled separately, usually under the name of model selection or hypothesis testing.

Ryan et al. (2015) use a Laplace approximation of the posterior distribution $p(\theta \mid D_{\text{exp}})$ combined with importance sampling. Drovandi et al. (2014) develop a method based on sequential Monte Carlo (SMC) that is faster than using MCMC. Woods et al. (2017) use a Monte Carlo approximation $\Phi(x) = 1/2 \sum_b U(y_b, x, \theta_b)$ with $(y_b, \theta_b) \sim p(y_b, \theta_b \mid x)$ on which they place a Gaussian process prior.

**Limitations** These methods agnostically accommodate non-analytical models using a Monte Carlo approach but require exhaustive model sampling in the model parameter space. On a case study with four models, each with ten
model parameters, and two design variables, the Vanlier et al. (2014) MCMC method requires six days of wall-clock time to compute two sets of four marginal predictive distributions. This cost is impractical for pharmacokinetic applications, where a physical experiment takes ≈1–4 days. SMC methods can converge faster than MCMC methods (Woods et al., 2017). But SMC methods can suffer from sample degeneracy, where only a few particles receive the vast majority of the probability weight (Li et al., 2014). Also, convergence analysis for MCMC and SMC methods is difficult.

Furthermore, these methods are only for design of experiments. Once an experiment is executed, the model discrimination issue remains. In this case the marginal predictive distributions \( p(f_i | x, D_{\text{exp}}) \) would enable calculating the model likelihoods.

### 3. Method

We wish to exploit results from the classical analytical approach to design of experiments for model discrimination, while considering that we may deal with non-analytical model functions. The key idea is replacing the original (non-analytical) model functions with analytical and differentiable surrogate models. The surrogate models can be learnt from model evaluations. We choose to use GPs to construct the surrogate models because they are flexible regression tools that allow us to extract gradient information.

A GP is a collection of random variables, any finite subset of which is jointly Gaussian distributed (Rasmussen & Williams, 2006, p. 13) and fully specified by a mean function \( m(\cdot) \) and a kernel function \( k(\cdot, \cdot) \). Prior knowledge about the model functions \( f_i \) can be incorporated into the choice of \( m \) and \( k \). To construct the GP surrogates, target data is needed. The purpose of the surrogate model is to predict the outcome of a model evaluation \( f_i(x, \theta_i) \), rather than the outcome of a physical experiment \( f_{\text{true}}(x) \). Hence, for each model \( M_t \) we create a data set \( D_{\text{sim},i} = \{ f_{\ell, t}, \bar{x}_t, \theta_t; \ell = 1, \ldots, N_{\text{sim},i} \} \) of simulated observations, i.e. model evaluations \( f_{\ell} = f_i(\bar{x}_t, \theta_t) \), where \( \bar{x}_t \) and \( \theta_t \) are sampled from \( \mathcal{X} \) and \( \Theta_i \), respectively. Note the difference between the experimental data set \( D_{\text{exp}} \) and the simulated data sets \( D_{\text{sim},i} \); whereas the data set \( D_{\text{exp}} \) has few (noisy, biological) observations, \( D_{\text{sim},i} \) has many (noise-free, artificial) observations.

We place independent GP priors \( f_i(\cdot) \sim \mathcal{GP}(m_{i, (e)}, k_{i, (e)}) \) on each target dimension \( e = 1, \ldots, E \) of the model function \( f_i \), \( i = 1, \ldots, M \). For notational simplicity, let \( f = f_i(\cdot), m = m_{i, (e)}, k = k_{i, (e)} \) and \( z = [x^\top, \theta_i^\top] \).

Given \( D_{\text{sim},i} \) with model evaluations \( f_{\ell} = f_i(\bar{x}_t, \theta_t) \) at locations \( [\mathbf{Z}]_{t} = \bar{z}_t \), the predictive distribution \( p(f | z, f, \mathbf{Z}) = \mathcal{N}(\mu(z_*), \sigma^2(z_*) \) at a test point \( z_* \) is given by

\[
\mu(z_*) = m(z_*) + k^\top_*/K^{-1}(J - m),
\]

where the covariance vector is \( [k_*] = k(z_*, z_*) \), the kernel matrix is \( [K]_{t, t} = k(\bar{z}_t, \bar{z}_t) + \sigma^2_0 \delta_{t, t} \), and the target mean is \( [m]_{t} = m(z_t) \), with a small artificial noise variance \( \sigma^2_0 \ll 1 \) added for numerical stability. Note the simplified notation \( \mu = \mu_{i,(e)} \) and \( \sigma^2 = \sigma^2_{0,e} \).

Pharmacokinetic models are often formulated as systems of ordinary differential equations, e.g. Han et al. (2006), where the solutions are smooth functions. If we wish to model these functions with a GP, it makes sense to use a covariance function that encodes the smoothness of the underlying function, e.g. the radial basis function (RBF) covariance function \( k_{\text{RBF}}(z, z') = \rho \exp\left(-\frac{1}{2}(z - z')^\top \Lambda^{-1} (z - z') \right) \), since it corresponds to Bayesian linear regression with an infinite number of basis functions (Rasmussen & Williams, 2006).

Our method extends to other choices of covariance function. Each function sample \( f \) drawn from a zero-mean GP is a linear combination \( f(x) = \sum_j \omega_j k(x, x_j) \) of covariance function terms (Rasmussen & Williams, 2006, p. 17). Different model characteristics, e.g. smoothness or periodicity, make alternate covariance functions appropriate. The RBF covariance function is common, but its strong smoothness prior may be too unrealistic for some systems. Other choices include: Matérn, polynomial, periodic, and non-stationary covariance functions. Sums or products of valid covariance functions are still valid covariance functions. Covariance functions can also model different behaviour over different input regions (Ambrogioni & Maris, 2016). Choosing the covariance function requires some prior knowledge about the modelled function characteristics, e.g. a periodic covariance function may be appropriate to model periodic drug dosing.

In \( k_{\text{RBF}} \), and other similar covariance functions, \( \rho^2 \) is the signal variance and \( \Lambda = \text{diag}(\lambda_{1:L}^2) \) is a diagonal matrix of squared length scales, with \( z \in \mathbb{R}^D \). Together, \( \rho, \Lambda \) and the artificial noise variance \( \sigma^2_0 \) (in \( K \)) are the GP hyperparameters. Because the data set \( D_{\text{sim},i} \) is arbitrarily large, we use point-estimate hyperparameters learnt through standard evidence maximisation.

We study a single model function \( f = f_i \) and its GP surrogates. Using the GP surrogates we can predict the outcome of a model evaluation \( f(x, \theta) \sim p(f | z, D_{\text{sim},i}) = N(\mu(z_*), \Sigma(z_*)) \), where \( \mu = [\mu_{1:1}]^\top \) and \( \Sigma = \text{diag}(\sigma^2_{0:1}) \). We approximate the (marginal) predictive distribution \( p(f | x, D_{\text{sim},i}, D_{\text{exp}}) \) to account for the model parameter uncertainty \( \theta \sim p(\theta | D_{\text{exp}}) \) due to parameter estimation from the noisy experimental data \( D_{\text{exp}} \).

### 3.1. Model Parameter Marginalisation

We assume that the experimental observations \( y_{1:N} \in D_{\text{exp}} \) have i.i.d. zero-mean Gaussian distributed measurement
We approximate the model parameter posterior distribution which we will detail in the following. 

Inserting (12) into (5) yields the approximate mean of the marginal predictive distribution

$$
\hat{\mu}_{(e)}(x) \approx \mu_{(e)}(\hat{\theta}, \theta) + \frac{1}{2} \text{tr} \left( \nabla_{\theta}^2 \mu_{(e)}(\theta) \Sigma_\theta \right) .
$$

Similarly, inserting (13) into (6) yields the diagonal approximate expectation of the variance with elements

$$
\mathbb{E}_\theta [\sigma^2_{(e)}(x, \theta)] \approx \sigma^2_{(e)}(\hat{\theta}, \theta) + \frac{1}{2} \text{tr} \left( \nabla_{\theta}^2 \sigma^2_{(e)}(\theta) \right) .
$$

The variance of the mean is a dense matrix with elements

$$
\mathbb{V}_\theta [\mu_{(e)}, \mu_{(e)} | x, \theta] \approx \nabla_{\theta} \mu_{(e)} \Sigma_\theta \nabla_{\theta} \mu_{(e)}^\top + \frac{1}{2} \text{tr} \left( \nabla_{\theta}^2 \Sigma_\theta \right) .
$$

(14)–(16) yield a tractable second-order Taylor approximation to $p(f | x, D_{\text{sim}}, D_{\text{exp}}) \approx \mathcal{N}(\hat{\mu}(x), \Sigma(x))$.

Figure 2 illustrates the process of constructing the GP surrogate models and the approximations of the marginal predictive distributions. Firstly, the model parameters are estimated from $D_{\text{exp}}$. Secondly, $D_{\text{sim}}$ is created from model evaluations and used (i) to learn the GP surrogate’s hyperparameters, and (ii) as target data for GP predictions. Thirdly, the approximation $\mathcal{N}(\hat{\mu}(x), \Sigma(x))$ of the marginal predictive distribution is computed. Fourthly and lastly, the marginal predictive distribution is used to design the next experiment and approximate the model likelihood.

4. Results

The following demonstrates that our approach for model discrimination with GP surrogate models exhibits a similar performance to the classical analytical method, but can be used to design experiments to discriminate between non-analytical model functions. The GP surrogate method performance will be studied using two different case studies.

Case study 1 has analytical model functions, so we can compare how well our new method performs compared to the classical analytical methods in Section 2.1. Case study 2 has pharmacokinetic-like models consisting of systems of ordinary differential equations.

In both case studies we compute the following metrics:

- (A) the average number of additional experiments $N - N_0$ required for all incorrect models to be discarded.
- (SE) the standard error of the average (A).
- (S) success rate; the proportion of tests in which all incorrect models were discarded.
- (F) failure rate; the proportion of tests in which the correct (data-generating) model was discarded.
We compare the design criteria (DC) $D_{BH}$, $D_{BF}$ and $D_{AW}$ described in Section 2.1 as well as random uniform sampling (denoted Uni.). We also compare the three different criteria for model discrimination (MD) described in Section 2.1: updated posterior likelihood $\pi_{N,i}$, the $\chi^2$ adequacy test, and the Akaike information criterion (AIC).

### 4.1. Case Study 1: Comparison to Analytical Method

Case study 1 is from the seminal paper by Buzzi-Ferraris et al. (1984). With two measured states $y_1$, $y_2$ and two design variables $x_1$, $x_2 \in [5, 55]$, we discriminate between four chemical kinetic models $M_i$, each with four model parameters $\theta_{i,j}$ that we fix to $\theta_{i,j} \in [0, 1]$:

- $M_1: f(1) = \theta_{1,1} x_1 x_2 / g_1$, $f(2) = \theta_{1,2} x_1 x_2 / g_1$,
- $M_2: f(1) = \theta_{2,1} x_1 x_2 / g_2^2$, $f(2) = \theta_{2,2} x_1 x_2 / h_2^2$,
- $M_3: f(1) = \theta_{3,1} x_1 x_2 / h_3^2$, $f(2) = \theta_{3,2} x_1 x_2 / h_3^2$,
- $M_4: f(1) = \theta_{4,1} x_1 x_2 / g_4$, $f(2) = \theta_{4,2} x_1 x_2 / h_4^2$,

where $g_1 = 1 + \theta_{1,3} x_1 + \theta_{1,4} x_2$ and $h_{i,j} = 1 + \theta_{i,2} x_2$. Buzzi-Ferraris et al. (1984) generate the experimental data $D_{exp}$ from model $M_1$ using $\theta_{1,1} = \theta_{1,3} = 0.1$ and $\theta_{1,2} = \theta_{1,4} = 0.01$ and Gaussian noise covariance $\Sigma_{exp} = \text{diag}(0.35, 2.3e-3)$. We start each test with $N_0 = 5$ randomly sampled experimental observations, and set a maximum budget of $N_{\text{max}} - N_0 = 40$ additional experiments.

Tables 1–3 compare the performance in 500 completed test instances of the classical analytical approach to the GP surrogate method using first- or second-order Taylor approximations. For this case study, the GP surrogate method performs similarly to the classical analytical approach. We see that the average number of required additional experiments (A) increases using $\pi$ for model discrimination, but that the model discrimination success rate (S) also increases. The model discrimination failure rate drops for almost all MD/DC combinations using the GP surrogate method. We also see that even for this simple case study, the difference in performance between the design criteria and random sampling is significant. The low average (A) for $\pi$/Uni. in Table 1 is likely an artefact due to the large failure rate (F).

### 4.2. Case Study 2: Non-Analytical Models

Case study 2 is from Vanlier et al. (2014). There are four different models, described by ordinary differential equations. The supplementary material contains the full model descriptions. The models describe a biochemical network, and are similar to typical pharmacokinetic models in (i) the use of reversible reactions, and (ii) the number of parameters and equations (see e.g. Han et al. (2006)).

We assume two measured states ($E = 2$), with model parameters $\theta_i \in [0, 1]^{10}$, initial concentrations $c \in [0, 1]^E$, measurement time point $t \in [0, 20]$, and design variables $x^T = [t, c^T]$. We follow Vanlier et al. (2014) and let $M_1$ generate the observed experimental data, with random uniformly sampled true model parameters and experimental
measurement noise covariance \( \Sigma_{\exp} = 9.0 \times 10^{-4} I \). Each test is initiated with \( N_0 = 20 \) observations at random design locations \( x_1, \ldots, x_{N_0} \). We set a maximum budget of \( N_{\text{max}} - N_0 = 100 \) additional experiments.

Case study 1 shows no obvious benefit in using the second-order over a first-order Taylor approximation. Hence, in case study 2 we only performance test the GP surrogate method with the first-order Taylor approximation.

The results in Table 4 for 4 models show that the rates of inconclusive results are higher for case study 2 than for case study 1. Furthermore, the ratio of successful model discriminations to failed model discriminations are lower for \( \pi/D_{\text{BH}} \) and \( \text{AIC}/D_{\text{AW}} \) than they are in case study 1. This is because for case study 2 the experimental noise is often larger than the difference between the model predictions. Figure 3 shows the evolutions of the Akaike weights \( w_i \) (Michalik et al., 2010) for all 63 completed tests. The Akaike weight evolutions show that especially models \( \mathcal{M}_1 \) and \( \mathcal{M}_2 \) are difficult to discriminate between, given that we only consider two measured states. Corresponding graphs for \( \pi \) and \( \chi^2 \) are similar-looking. One can also see from the model formulations (see supplementary material) that all models are similar.

To verify that the Table 4 results for 4 models are indeed due to the indiscriminability of model \( \mathcal{M}_1 \) and \( \mathcal{M}_2 \), we carry out another set of tests where we only consider models \( \mathcal{M}_1, \mathcal{M}_3, \mathcal{M}_4 \). The results in Table 4 for 3 models show that removing model \( \mathcal{M}_2 \) enables the GP surrogate method to correctly discriminate between the remaining models more successfully. In practice, researchers faced with apparent model indiscriminability have to rethink their experimental set-up to lower measurement noise or add measured states.

### Table 1. Results from case study 1 using the analytical approach.

<table>
<thead>
<tr>
<th>MD</th>
<th>DC</th>
<th>( \pi )</th>
<th>( \chi^2 )</th>
<th>AIC</th>
<th>( \pi )</th>
<th>( \chi^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td>2.60</td>
<td>2.87</td>
<td>2.08</td>
<td>3.43</td>
<td>10.39</td>
</tr>
<tr>
<td>SE</td>
<td></td>
<td>0.04</td>
<td>0.12</td>
<td>0.04</td>
<td>0.11</td>
<td>0.45</td>
</tr>
<tr>
<td>S [%]</td>
<td></td>
<td>86.4</td>
<td>64.2</td>
<td>62.4</td>
<td>59.0</td>
<td>85.8</td>
</tr>
<tr>
<td>F [%]</td>
<td></td>
<td>13.6</td>
<td>5.0</td>
<td>37.6</td>
<td>40.6</td>
<td>4.4</td>
</tr>
<tr>
<td>I [%]</td>
<td></td>
<td>0.0</td>
<td>30.8</td>
<td>0.0</td>
<td>0.4</td>
<td>9.8</td>
</tr>
</tbody>
</table>

### Table 2. Results from case study 1 using the GP surrogate method with first-order Taylor approximation.

<table>
<thead>
<tr>
<th>MD</th>
<th>DC</th>
<th>( \pi )</th>
<th>( \chi^2 )</th>
<th>AIC</th>
<th>( \pi )</th>
<th>( \chi^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td>4.31</td>
<td>2.23</td>
<td>2.72</td>
<td>10.99</td>
<td>10.02</td>
</tr>
<tr>
<td>SE</td>
<td></td>
<td>0.09</td>
<td>0.06</td>
<td>0.08</td>
<td>0.34</td>
<td>0.45</td>
</tr>
<tr>
<td>S [%]</td>
<td></td>
<td>95.6</td>
<td>47.4</td>
<td>88.6</td>
<td>69.6</td>
<td>83.6</td>
</tr>
<tr>
<td>F [%]</td>
<td></td>
<td>4.4</td>
<td>4.8</td>
<td>11.4</td>
<td>30.4</td>
<td>4.8</td>
</tr>
<tr>
<td>I [%]</td>
<td></td>
<td>0.0</td>
<td>47.8</td>
<td>0.0</td>
<td>0.0</td>
<td>11.6</td>
</tr>
</tbody>
</table>

### Table 3. Results from case study 1 using the GP surrogate method with second-order Taylor approximation.

<table>
<thead>
<tr>
<th>MD</th>
<th>DC</th>
<th>( \pi )</th>
<th>( \chi^2 )</th>
<th>AIC</th>
<th>( \pi )</th>
<th>( \chi^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td>4.14</td>
<td>2.29</td>
<td>2.64</td>
<td>9.53</td>
<td>10.11</td>
</tr>
<tr>
<td>SE</td>
<td></td>
<td>0.09</td>
<td>0.07</td>
<td>0.06</td>
<td>0.30</td>
<td>0.47</td>
</tr>
<tr>
<td>S [%]</td>
<td></td>
<td>96.9</td>
<td>46.6</td>
<td>90.1</td>
<td>64.3</td>
<td>80.2</td>
</tr>
<tr>
<td>F [%]</td>
<td></td>
<td>3.1</td>
<td>9.9</td>
<td>9.9</td>
<td>35.7</td>
<td>7.0</td>
</tr>
<tr>
<td>I [%]</td>
<td></td>
<td>0.0</td>
<td>43.5</td>
<td>0.0</td>
<td>0.0</td>
<td>12.8</td>
</tr>
</tbody>
</table>

### Table 4. Results from case study 2.

<table>
<thead>
<tr>
<th>4 models</th>
<th>3 models</th>
</tr>
</thead>
<tbody>
<tr>
<td>MD</td>
<td>DC</td>
</tr>
<tr>
<td>A</td>
<td></td>
</tr>
<tr>
<td>SE</td>
<td></td>
</tr>
<tr>
<td>S [%]</td>
<td></td>
</tr>
<tr>
<td>F [%]</td>
<td></td>
</tr>
<tr>
<td>I [%]</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3. Results from case study 2 with 4 models using the \( D_{\text{AW}} \) design criterion and AIC for model discrimination. (a)–(d) show the evolution of the Akaike weights \( w_i \) for model \( \mathcal{M}_1, \ldots, \mathcal{M}_4 \), respectively, for 63 completed tests. (e) shows the average Akaike weight evolutions (with one standard deviation) in (a)–(d). The results indicate that models \( \mathcal{M}_1 \) and \( \mathcal{M}_2 \) are indiscriminable.
5. Discussion

The healthcare industry already uses machine learning to help discover new drugs (Pfizer, 2016). But passing the pre-clinical stage is expensive. DiMasi et al. (2016) estimate the average pre-approval R&D cost to $2.56B (in 2013 U.S. dollars) with a yearly increase of 8.5%. For the pre-clinical stage (before human testing starts), DiMasi et al. (2016) estimate the R&D cost to $1.1B.

Model discrimination, which takes place in the pre-clinical stage, may lower drug development cost by identifying a model structure quickly. Heller et al. (2018) explain that a major hurdle for in silico pharmacokinetics is the difficulty of model discrimination. Scannell & Bosley (2016) and Plenge (2016) also attribute a significant part of the total R&D cost for new drugs to inaccurate models passing the pre-clinical stage only to be discarded in later clinical stages. Inaccurate models need to be discarded sooner rather than later to avoid expensive failures.

Design of experiments for model discrimination has typically focussed either on analytical approaches, which cannot manage all functions, or on data-driven approaches, which may have computational difficulties or lack interpretable marginal predictive distributions. We leverage advantages of both approaches by hybridising the analytical and data-driven approaches, i.e. we replace the analytical functions from the classical methods with GP surrogates.

The novel GP surrogate method is highly effective. Case study 1 directly compares it with the analytical approach on a classical test instance. The similar performance indicates successful hybridisation between analytical and data-driven approaches. Case study 2 considers an industrially relevant problem suffering from model indiscriminability. The GP surrogate method successfully avoids introducing overconfidence into the model discrimination despite the assumptions used to derive the marginal predictive distribution.

The parameter estimation step is limited to few and noisy data points, e.g. from time-consuming pharmacokinetic experiments. If the approximation $p(\theta \mid D_{\text{exp}}) \approx N(\theta, \Sigma)$ is not accurate, the Section 3.1 Taylor approximations of $\hat{\mu}$ and $\Sigma$ can be extended to a mixture of Gaussian distributions $p(\theta \mid D_{\text{exp}}) = \sum_j \omega_j N(\theta_j, \Sigma_{\theta,j})$, with $\sum_j \omega_j = 1$, if $\{\theta_j, \Sigma_{\theta,j}\}$ are given. The analytical approximations sacrifice some accuracy for computational tractability: Our computational time for design of experiments in case study 2 was on the order of minutes, whereas the method of Vanlier et al. (2014) requires days. Our results indicate that the approximations work well in practice. GP surrogate methods could sharply reduce the cost of drug development’s pre-clinical stage, e.g. by shrinking the time needed to decide on a new experiment or by finding a good model structure in a reduced number of pharmacokinetic experiments.

Our new method directly applies to neighbouring fields. For example, dual control theory deals with control of unknown systems, where the goal is to balance controlling the system optimally vs. learning more about the system. When rival models predict the unknown system behaviour, learning about the system involves designing new experiments, i.e. control inputs to the system (Cheong & Manchester, 2014). This design of experiments problem, e.g. in Larsson et al. (2013), would be likely to have very strict time requirements and benefit from confidence bounds on the model predictions. Our method also applies to neighbouring application areas, e.g. distinguishing between microalgae metabolism models (Ulmasov et al., 2016) or models of bone neotissue growth (Mehrian et al., 2018).

A very hot topic in machine learning research is learning in implicit generative models (Mohamed & Lakshminarayanan, 2017), e.g. generative adversarial networks (GANs) (Goodfellow et al., 2014), which employ a form of model discrimination. Here the likelihood function $p(y \mid M_i)$ is not given explicitly, so the goal is to learn an approximation $q(y \mid M_i)$ of the data distribution $p(y)$. The model discrimination tries to identify from which distribution $q(y \mid M_i)$ or $p(y)$ a random sample has been drawn. For the case of design of experiments for model discrimination, we have a prescribed probabilistic model, since we explicitly specify an approximation of the likelihood $p(y \mid M_i, x, \theta_i)$. But we have (i) the physical experiments, the true but noisy data generator that is expensive to sample, and (ii) the rival models, deterministic and relatively cheap (but probably inaccurate!) data generators. Design of experiments for model discrimination also ties in with several other well-known problems in machine learning, e.g. density estimation, distribution divergence estimation, and model likelihood estimation.

6. Conclusions

We have presented a novel method of performing design of experiments for model discrimination. The method hybridises analytical and data-driven methods in order to exploit classical results while remaining agnostic to the model contents. Our method of approximating the marginal predictive distribution performs similarly to classical methods in a case study with simple analytical models. For an industrially-relevant system of ordinary differential equations, we show that our method both successfully discriminates models and avoids overconfidence in the face of model indiscriminability. The trade-off between computational complexity and accuracy that we introduce is useful for design of experiments for discrimination e.g. between pharmacokinetic models, where the duration of the experiments is shorter than or one the same time-scale as the computational time for Monte Carlo-based methods.
Acknowledgements

This work has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no.675251, and an EPSRC Research Fellowship to R.M. (EP/P016871/1).

References


