Proceedings of the 5th International Conference on Statistics: Theory and Applications (ICSTA'23)
Brunel University, London, United Kingdom – August 03-05, 2023
Paper No. 153
DOI: 10.11159/icsta23.153

Bayesian Model Selection and Emulation for Protein Fluorescence

William Ryan[1], Dirk Husmeier[1], Olaf Rolinski[2], Vladislav Vyshemirsky[1]
[1]University of Glasgow
G12 8QQ, Glasgow, United Kingdom
w.ryan.1@research.gla.ac.uk
[2]University of Strathclyde
G1 1XQ, Glasgow, United Kingdom

Abstract **-** Fluorescence decay of amino acids in protein is a complex process for which multiple models have been proposed. Likelihood function evaluation for certain models can be computationally expensive, and as such surrogate models may be introduced to speed up inference. In this paper, Gaussian processes are implemented in likelihood estimation of a range of models defined by convolutions of an initial excitation input and a decay function using both synthetic and real world data. Parameter inference and model selection using the surrogate models are performed and compared against the exact results. Model selection when incorporating surrogate models into the inference process is shown to be consistent.

Keywords**:** Protein fluorescence, Bayesian model selection, emulation, Kronecker product

1. Introduction

Protein fluorescence is a topic under study which displays properties potentially very useful for medical diagnoses [1][2]. Multiple competing models exist which explain the decay of intrinsic fluorescence of proteins, ranging from computationally cheap exponential function to more expensive lifetime models of fluorescence decay involving computationally intensive methods [3][4]. In this paper, multi-exponential models of increasing complexity are considered with the goal of comparing reliability of model selection using exact likelihood functions against surrogate-assisted models.

2. Model

Mathematically, the fluorescence decay of protein in amino acids is modelled by a convolution function of an excitation input, $\mathrm{E}(t)$, with a delay parameter, Δ, which takes into account the delay between emitting the light pulse and the decay beginning, and a decay function, $I\,(t)$:

$$\mathrm{F}(t, \Delta) = \mathrm{E}(t + \Delta) * \mathrm{I}(t) \tag{1}$$

The fluorescence decay functions considered here are multi-exponential, namely 2-, 3- and 4-exponential:

$$I(t) = \sum_{i=1}^{n} \alpha_i \exp\left(-\frac{t}{\tau_i}\right) \tag{2}$$

where α_i and τ_i indicate the intensity scaling factor and lifetime decay rate, respectively

In order to perform inference, we consider the observed photon counts to be independently Poisson distributed, with rate at each observation defined to be the value of the fitted model at each of the n observed points. Using a normal approximation to the Poisson distribution and the shorthand $\mathrm{F}_i = \mathrm{F}(i, \theta)$, the likelihood of observing data $\mathrm{y} = (y_i)_{i=1,\dots,n}$ given parameters θ (decay function parameters and the input delay parameter) is of the form:

$$p(\mathrm{y}|\theta) = \prod_{i=1}^{n} \mathcal{N}\,(y_i|\mathrm{F}_i, \mathrm{F}_i + \sigma^2) \tag{3}$$

An additional variance term σ^2 was introduced to account for model mismatch. Using the likelihood above and applying priors to the model parameters allows us to perform Bayesian inference to sample from the parameters' posterior distributions. The marginal likelihood given a selected model $p(\mathrm{y}|\mathcal{M})$ is also of interest, as a means of comparing models.

In this paper, results of inference using exact likelihood values are compared against the same inference procedures performed utilizing emulators, and the results are shown to be reliable by a large scale simulation process.

3. Methods

3.1. Model selection

The industry standard model comparison procedure is to find parameters which minimize the reduced χ^2 statistic [5], but performing inference on the model parameters allows for the opportunity to use all available information from the parameters' posterior distributions. For this reason, we make use of the Watanabe-Akaike information criterion (WAIC) in place of a statistic based on a single vector of parameters [6]. WAIC is defined as the within-sample predictive accuracy with an additional correction term that accounts for effective number of parameters [7].

Further, marginal likelihood comparisons are another viable Bayesian method of ranking hypotheses, in this case the decay models under consideration. While WAIC tests the predictive capabilities of a model, marginal likelihoods act as a form of explanatory model selection and are defined as the probability of observing data **y** given a model $\mathcal{M}$,

$$p(\mathrm{y}|\mathcal{M}) = \int p(\mathrm{y}|\boldsymbol{\theta}, \mathcal{M})\, p(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{4}$$

and are more sensitive to choice of prior placed on $\boldsymbol{\theta}$.

Competing models may be compared by Bayes factor, given by the ratio of marginal likelihoods for models $\mathcal{M}_1$ and $\mathcal{M}_2$,

$$\text{Bayes factor}(\mathcal{M}_2, \mathcal{M}_1) = \frac{p(\mathrm{y}|\mathcal{M}_2)}{p(\mathrm{y}|\mathcal{M}_1)} \tag{5}$$

which can then be used as evidence of one model over another [8].

Often, the integral in Equation 4 is intractable, and for this reason numerous methods of approximating the marginal likelihood have been proposed. In this paper, Chib's method of estimating marginal likelihoods $p(\mathrm{y}|\mathcal{M})$ is utilised [9]. It is derived by computing a posterior ordinate $\pi(\theta^*|\mathbf{y})$, where π is the posterior density of θ, and θ^* is a sampled point from the posterior distribution, typically from a high density region for computational efficiency. The log marginal likelihood for some model M may then found by rearranging Bayes' Theorem:

$$\log p(\mathrm{y}|\mathcal{M}) = \log p(\mathrm{y}|\theta^*, \mathcal{M}) + \log p(\theta^*) - \log \pi(\theta^*|\mathrm{y}, \mathcal{M}) \tag{6}$$

3.2. Bayesian inference

The prior information available consisted of expert knowledge of lower and upper limits for parameter spaces, upon which uniform prior distributions were placed [5]. Additionally, since the exponential decay functions were invariant to label switching of the pairs of decay lifetimes and scaling factors in Equation 2, the parameters were permuted post-sampling to place the mean of the scaling factors in ascending order, focusing on the unimodal posterior instead of all $n!$ modes of the n-exponential model.

3.3. Surrogate models in likelihood functions

The motivation for utilising surrogate models in fluorescence decay stems from the computationally demanding lifetime decay processes that have been proposed in the literature. In particular, fluorescence decay has been previously described by non-Debye models requiring numerical integration [3] and systems of differential equations. Emulators have previously been utilised to reduce computational costs by several orders of magnitude [10].

Two methods of incorporating surrogate models into inference were considered, both using Gaussian Processes (GPs): emulating the log-likelihood surface of the parameters involved in the convolution function and emulating the signal of the convolution function itself [11].

Briefly, a Gaussian process prior on a function f gives it a distribution defined by a mean function $m(x)$ and covariance function $k(x, x')$ of its inputs,

$$f \sim \mathcal{GP}\big(m(x), k(x, x')\big) \tag{7}$$

Observing values $\mathcal{Y}$ at training locations $\mathcal{X}$ of the underlying function f allows us to update the GP. The new posterior predictive distribution of $f(x)$ given $\mathcal{Y}$ and covariance matrix $K = k(\mathcal{X}, \mathcal{X})$ at a new single input location x^* is then given by standard conditional multivariate Gaussian distribution updating rules [12],

$$f(x^*)|\mathcal{Y} \sim \mathcal{N}\big(\mu(x^*), \sigma^2(x^*)\big) \tag{8}$$

where $\mu(x*) = m(x*) - k(x*, \mathcal{X})K^{-1}\mathcal{Y} - m(\mathcal{X})$ and $\sigma^2(x*) = k(x*, x*) - k(x*, \mathcal{X})K\ k(\mathcal{X}, x*)$. Gaussian processes can be used for emulation of a wide variety of functions thanks to the versatility provided by the choice of covariance function, or kernel, and its associated hyperparameters.

Directly emulating the log-likelihood function meant implementing a fairly straightforward GP with one output and all model parameters involved as inputs, whereas emulating the convolution function required emulating all $4,096$ data points for computing the likelihood function. To this end, time was included as an input variable which allowed for some computational shortcuts. To ensure positive count values, the natural logarithm of the convolution function was emulated.

Treating time as an input variable allowed us to use a smaller subset of time points in the training data to make predictions at any of the points and avoided having to build a multivariate output GP with 4,096 outputs. We allow time to have a separate covariance function from the rest of the input variables by imposing separability upon the GP's covariance function, that is, for vectors of decay function parameters θ, θ' with covariance function k_θ and time points t, t', with kernel k_t, we have

$$k\big((t,\theta),(t',\theta')\big) = k_t(t,t')k_\theta(\theta,\theta') \tag{9}$$

The computational cost of inverting the $nm \times nm$ covariance matrix, K, for n vectors of decay function parameters and m time points would normally be $\mathcal{O}(n^3m^3)$, but we are able to make use of the Kronecker product structure of K,

$$K(\mathrm{t}, \boldsymbol{\theta}) = K_t(\mathrm{t}) \otimes K_\theta(\boldsymbol{\theta}) \tag{10}$$

and the properties of inverting a Kronecker product of matrices,

$$\big(K_t(\mathrm{t}) \otimes K_\theta(\boldsymbol{\theta})\big)^{-1} = K_t(\mathrm{t})^{-1} \otimes K_\theta(\boldsymbol{\theta})^{-1} \tag{11}$$

to reduce the complexity to $\mathcal{O}(n^3 + m^3)$ [13].

3.4. Experiment set-up

When building the dataset to train the emulator for the log-likelihood surface, an initial Latin hypercube space filling design of the parameters' prior domains consisting of $1,000$ points was used followed by a history matching algorithm to refine the training dataset [14]. For each of the multi-exponential decay models, separable Matérn 5/2 kernels were placed on the parameters, the hyperparameters of which were optimised to maximise marginal log-likelihood of the multivariate Gaussian function given the data. Matérn 5/2 kernels were used for their versatility while being twice differentiable compared to the infinitely differentiable RBF kernel, which may have imposed too large an assumption on smoothness of the physical systems being modelled. Additionally, the mean function placed on the GP was constant at the maximum observed log-likelihood in the training dataset to avoid the emulated surface reverting to 0.

For the GP emulating the log output of the convolution function, the same Matérn kernels were again placed on the model parameters. Given the sudden spike and gradual decay of the convolution function (see Figure 1), a non-stationary neural network covariance function was placed on the time variable [12]. Despite making use of the property of an inverse of a Kronecker product (11) to reduce computational complexity, emulating the output required making predictions at all available time points instead of the singular output of the log-likelihood emulator. The initial space filling design of the parameter space was increased to $5,000$ points, and roughly 200 time points were used across the $4,096$ in the model, with points clustered closer together around the spike of the convolution function. For each new prediction at some vector $\boldsymbol{\theta}$, 50 of the nearest distance vectors in the training dataset based on standardised Euclidean distance were used to build a new GP model to perform the prediction as a local GP approximation [15]. This contributed to a considerable time save while providing satisfactorily accurate results.

Both emulators were used as part of a Metropolis-Hastings sampler implemented for each decay model. For the first $10,000$ iterations of the algorithm at intervals of 100 iterations, the GP predictions were compared against the true evaluations at the current step. If the prediction variance was too high given a predetermined threshold, the parameter

values and corresponding exact convolution function output were added to the GP's dataset and the hyperparameters re-optimised [16].

The methods described were implemented in Python, and GPs were built using libraries GPflow and GPy. It is important to note that performing inference by making predictions using GPs was similar in speed or slower than running the exact convolution function (roughly 4 seconds to perform 100 likelihood iterations using the exact function, the same when emulating the log-likelihood surface and half a minute when emulating the entire output). Instead, the goal has been to compare whether introducing emulation is reliable for model selection for the computationally easy models considered in this paper before applying the methods with more demanding decay response curves.

3.5. Data

The multi-exponential decay models were tested against synthetic data generated using the 3-exponential model and fixed parameters. Model selection was also carried out on an available real-world dataset of measured photon counts of the fluorescence decay of the amino acid tryptophan (Trp) in the protein human serum albumin (HSA) at fixed temperature and observation wavelength. The human serum albumin and the phosphate buffer were purchased from Sigma-Aldrich (Poole, UK). The $3x10^{-5}$ M HSA solution in 0.01 M phosphate buffer, pH7.4 has been prepared on the day of measurements. The time-correlated single photon counting (TCSPC)-based Fluorocube fluorescence lifetime system (Horiba Jobin Yvon IBH, Glasgow, UK) has been used to record the fluorescence decays. An AlGaN version of a pulse light emitting diode working at 295 nm and 1 MHz repetition rate has been used to excite the Trp in HSA directly. The sample was stored in an oven at 37 °C for the duration of the experiments. All measurements were carried out using 4 × 1 × 1 cm quartz cuvettes.

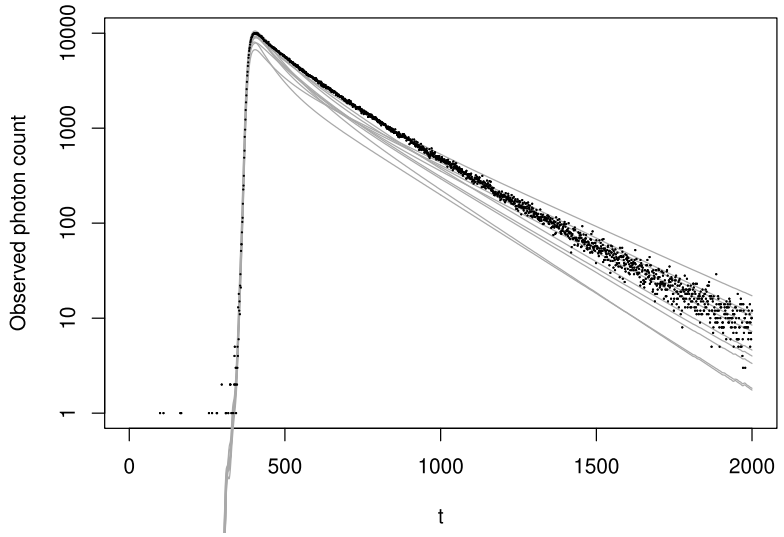

Fig. 1: Photon counts at 2, 000 of the 4, 096 real world datapoints on the log scale. Mean predicted values from the GP emulating the 3-exponential model with Kronecker product kernel for 10 vectors of decay parameters sampled from the parameters' priors are plotted in grey.

4. Results

4.1 Synthetic data

Synthetic data was created using the 3-exponential model with fixed decay parameters in order to test the model selection criteria when the true underlying model was known. Figure 2 displays boxplots of negative WAIC and log marginal likelihood estimates for the 3 models obtained from 20 simulations carried out. WAIC was multiplied by -1 for ease of comparison with log marginal likelihood, so bigger outputs indicate better scores.

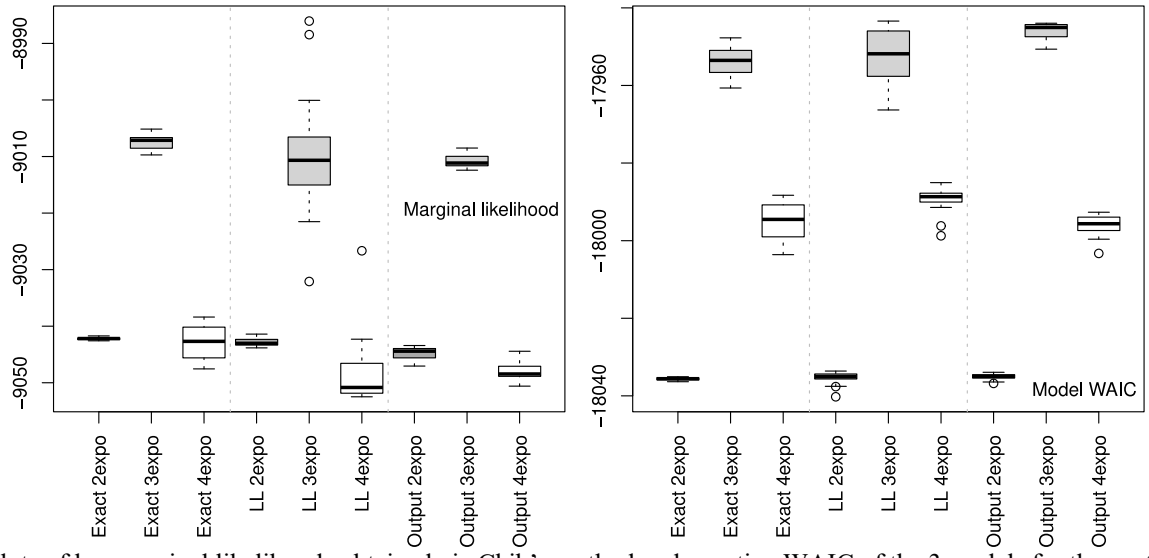

Fig. 2: Boxplots of log marginal likelihoods obtained via Chib's method and negative WAIC of the 3 models for the synthetic data. The labels on the x-axis indicate whether the exact likelihood, emulated log-likelihood or emulated output were used in inference.

The model selection criteria used in this paper both correctly identified the 3-exponential model as the data-generating model, and we arrive at the same conclusion when using surrogate-assisted likelihood functions. While WAIC indicated a better predictive fit for the more complex model, the log marginal likelihoods of the 2- and 4-exponential models found using both the exact likelihood and emulation are relatively similar, which may be a product of Occam's razor intrinsic to marginal likelihood.

4.1 Real-world data

40 simulations were carried out for each method of inference for the real-world dataset. In contrast to the synthetic data where the ground truth was known, in the case of real-world data, selection criteria in the case of model mismatch where the true model wasn't included in the pool of candidate models was of interest [17]. The results in Figure 3a indicate that the 2-exponential model isn't able to adequately model the fluorescence decay present in the data as well as the other two models. Choosing between the 3- and 4-exponential model to represent the decay represented in the data is not quite as clear cut as ruling out the 2-exponential model. Looking more closely at the comparison in Figure 3b reveals that using the exact likelihood function in MCMC outputs a higher estimated marginal likelihood via Chib's method for the less complex 3-exponential model but that WAIC highlights the better predictive performance of the 4-exponential model. The output from the MCMC runs using surrogate-assisted likelihood functions indicates the same.

Table 1: 95% credible intervals for scaling parameter posterior distributions.

Model	α_1	α_2	α_3	α_4
3-exponential	(0.0087, 0.014)	(0.016, 0.021)	(0.023, 0.025)	-
4-exponential	$(3.3 \times 10^{-5}, 1.2 \times 10^{-4})$	(0.012, 0.014)	(0.019, 0.022)	(0.021, 0.023)

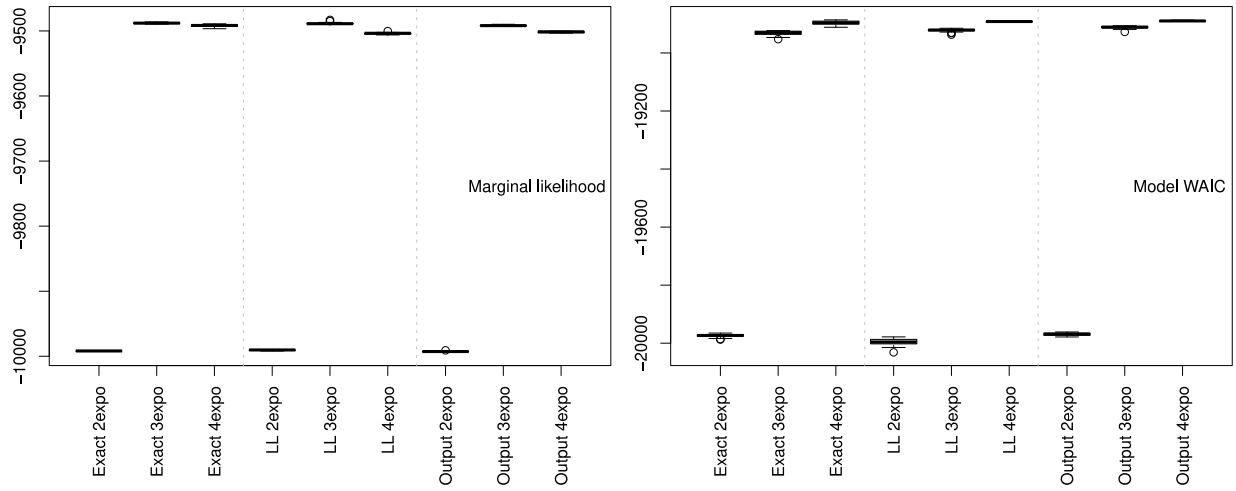

Fig. 3a: Boxplots of model selection criteria for the all exponential models side by side for the real-world data. It is clear that the 2-exponential model doesn't perform as well as either of the other 2.

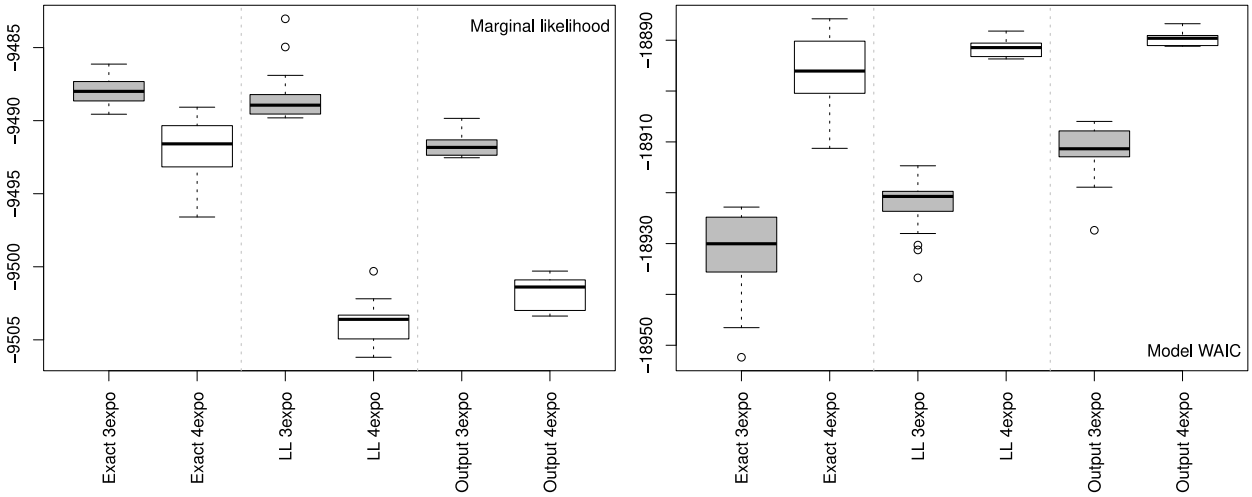

Fig. 3b: Comparison of model selection criteria of the 3-exponential (in grey) and 4-exponential (in white) models for the real-world data. The labels on the x-axis indicate whether the exact likelihood, emulated log-likelihood or emulated output were used in inference. The emulated outputs in particular were able to provide clear results for which model was preferred for either model selection criterion

Looking closer at the posterior values of the parameters reveals the decay parameter introduced in the 4-exponential model may play a less significant role in explaining the decay. Table 1 displays 95% credible intervals for the scaling parameters found across all methods from one of the simulations for the 3- and 4-exponential models. The smallest parameter in the 4-exponential is two orders of magnitude smaller than the smallest parameter in the 3-exponential model, which may indicate an extra parameter being unnecessary in modelling the decay. This is in agreement with the marginal likelihood, which supports the notion of picking the simpler model of the two. Using Jeffreys's original scale in [8] of the Bayes' factor between the 3- and 4-exponential model, we would conclude there is substantial to strong evidence for the simpler of the two models using the mean of the marginal likelihoods, depending on inference method.

5. Conclusion

Gaussian processes are a versatile tool for emulating a target function, and have here been applied to the likelihood functions in parameter inference and model selection of competing fluorescence decay models. Two methods of utilising emulation were applied, directly emulating the log-likelihood function and emulating the decay output. Given the large amounts of outputs, implementing a full multivariate output GP would have been particularly onerous, and emulating each time point individually would have naively overlooked the correlation structure of the time points. A GP using both the decay function parameters and time as input variables with a Kronecker product structured covariance function for computational feasibility was introduced as a middle ground between the two extremes.

Using synthetic and real-world data, the model selection criteria obtained using both emulators were shown to be accurate and consistent for decision making purposes. In particular, the emulators were able to perform well even when using little prior knowledge of the parameters' values and without extensive optimisation before performing MCMC. However, emulating the output allows for building the GP before any data collection occurs, whereas emulating the likelihood directly requires constructing the GP with a large dataset for predictive accuracy or performing optimisation post-data collection. Depending on applications of the emulator, this may be an attractive property of using the emulated output over the emulated likelihood.

References

[1] A. Alghamdi, V. Vyshemirsky, D. Birch, and O. Rolinski, "Detecting beta-amyloid aggregation from the time-resolved emission spectra," *Methods and Applications in Fluorescence*, Dec. 2017.

[2] L. H. C. Chung, D. J. S. Birch, V. Vyshemirsky, M. G. Ryadnov, and O. J. Rolinski, "Insulin aggregation tracked by its intrinsic tres," *Applied Physics Letters*, vol. 111, no. 26, p. 263701, 2017.

[3] O. J. Rolinski, T. Wellbrock, D. J. S. Birch, and V. Vyshemirsky, "Tyrosine photophysics during the early stages of β-amyloid aggregation leading to alzheimer's," *The Journal of Physical Chemistry Letters*, vol. 6, no. 15, pp. 3116–3120, 2015. PMID: 26267211.

[4] O. J. Rolinski and V. Vyshemirsky, "Fluorescence kinetics of tryptophan in a heterogeneous environment," *Methods and Applications in Fluorescence*, vol. 2, p. 045002, 12 2014.

[5] J. R. Lakowicz, ed., *Introduction to Fluorescence*, pp. 1–26. Boston, MA: Springer US, 2006.

[6] S. Watanabe, "Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory," 2010.

[7] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. ed., 2004.

[8] H. Jeffreys, *Theory of Probability*. Oxford, England: Clarendon Press, 1939.

[9] S. Chib and I. Jeliazkov, "Marginal likelihood from the metropolis–hastings output," *Journal of the American Statistical Association*, vol. 96, no. 453, pp. 270–281, 2001.

[10] U. Noè, A. Lazarus, H. Gao, V. Davies, B. Macdonald, K. Mangion, C. Berry, X. Luo, and D. Husmeier, "Gaussian process emulation to accelerate parameter estimation in a mechanical model of the left ventricle: a critical step towards clinical end-user relevance," *Journal of The Royal Society Interface*, vol. 16, no. 156, p. 20190114, 2019.

[11] V. Davies, U. Noè, A. Lazarus, H. Gao, B. Macdonald, C. Berry, X. Luo, and D. Husmeier, "Fast parameter inference in a biomechanical model of the left ventricle by using statistical emulation," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 68, no. 5, pp. 1555–1576, 2019.

[12] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning.* Adaptive computation and machine learning, MIT Press, 2006.

[13] M. A. Alvarez, L. Rosasco, and N. D. Lawrence, "Kernels for vector-valued functions: a review," 2012.

[14] R. D. Wilkinson, "Accelerating abc methods using gaussian processes," 2014.

[15] R. B. Gramacy and D. W. Apley, "Local gaussian process approximation for large computer experiments," *Journal of Computational and Graphical Statistics*, vol. 24, no. 2, pp. 561–578, 2015.
[16] A. J. Chowdhury and G. Terejanu, "Approximate sampling using an accelerated metropolis-hastings based on bayesian optimization and gaussian processes," *CoRR*, vol. abs/1910.09347, 2019.
[17] L. M. Paun, M. J. Colebank, M. S. Olufsen, N. A. Hill, and D. Husmeier, "Assessing model mismatch and model selection in a bayesian uncertainty quantification analysis of a fluid-dynamics model of pulmonary blood circulation," *Journal of The Royal Society Interface*, vol. 17, no. 173, p. 20200886, 2020.