# Considering discrepancy when calibrating a mechanistic electrophysiology model

Chon Lok Lei[1], Sanmitra Ghosh[2], Dominic G. Whittaker[3], Yasser Aboelkassem[4], Kylie A. Beattie[5], Chris D. Cantwell[6], Tammo Delhaas[7], Charles Houston[6], Gustavo Montes Novaes[8], Alexander V. Panfilov[9,10], Pras Pathmanathan[11], Marina Riabiz[12], Rodrigo Weber dos Santos[8], Keith Worden[13], Gary R. Mirams[3] and Richard D. Wilkinson[14]

[1] Computational Biology & Health Informatics, Dept. of Computer Science, University of Oxford, UK.
[2] MRC Biostatistics Unit, University of Cambridge, UK
[3] Centre for Mathematical Medicine & Biology, School of Mathematical Sciences, University of Nottingham, UK.
[4] Department of Bioengineering, University of California San Diego, USA.
[5] Systems Modeling and Translational Biology, GlaxoSmithKline R&D, Stevenage, UK.
[6] ElectroCardioMaths Programme, Centre for Cardiac Engineering, Imperial College London, UK.
[7] CARIM School for Cardiovascular Diseases, Maastricht University, the Netherlands.
[8] Graduate Program in Computational Modeling, Universidade Federal de Juiz de Fora, Brazil.
[9] Department of Physics and Astronomy, Ghent University, Belgium.
[10] Laboratory of Computational Biology and Medicine, Ural Federal University, Ekaterinburg, Russia.
[11] U.S. Food and Drug Administration, Center for Devices and Radiological Health, Office of Science and Engineering Laboratories, USA.
[12] Department of Biomedical Engineering King's College London and Alan Turing Institute, UK.
[13] Dynamics Research Group, Department of Mechanical Engineering, University of Sheffield, UK.
[14] School of Mathematics and Statistics, University of Sheffield, UK.

**Abstract**

Uncertainty quantification (UQ) is a vital step in using mathematical models and simulations to take decisions. The field of cardiac simulation has begun to explore and adopt UQ methods to characterise uncertainty in model inputs and how that propagates through to outputs or predictions. In this perspective piece we draw attention to an important and under-addressed source of uncertainty in our predictions — that of uncertainty in the model structure or the equations themselves. The difference between imperfect models and reality is termed *model discrepancy*, and we are often uncertain as to the size and consequences of this discrepancy. Here we provide two examples of the consequences of discrepancy when calibrating models at the ion channel and action potential scales. Furthermore, we attempt to account for this discrepancy when calibrating and validating an ion channel model using different methods, based on modelling the discrepancy using Gaussian processes (GPs) and autoregressive-moving-average (ARMA) models, then highlight the advantages and shortcomings of each approach. Finally, suggestions and lines of enquiry for future work are provided.

**Keywords:** Model discrepancy, Uncertainty quantification, Cardiac model, Bayesian inference

## 1 Introduction

This perspective paper discusses the issue of model discrepancy — the difference between a model's equations and reality, even with the best possible set of values given for parameters within these equations. The concepts and issues we highlight are widely applicable to any modelling where governing equations are approximations or assumptions; thus our perspective paper is intended for computational, mathematical and statistical modellers within many other fields as well as within and outside biological modelling. The focus of the examples is in cellular cardiac electrophysiology, a well-developed area of systems biology [1].

### 1.1 Cardiac modelling

Generally speaking, cardiac models are sets of mathematical functions governed by continuous sets of ordinary and/or partial (when spatial dimensions are considered) differential equations, integrated using

computational discrete techniques which produce responses that depend on the model inputs. Inputs can include model parameters, initial conditions, boundary conditions and cellular, tissue or whole organ geometrical aspects. Such inputs are often calibrated using experimental data, and those which have physiological meaning can sometimes be obtained by direct measurement.

Cardiac mathematical modelling and computational simulations have been remarkably successful tools that have provided insights into cardiac physiological mechanisms at cellular, tissue and whole organ scales. In the majority of these quantitative efforts, models are derived based on simplified representations of complex biophysical systems and use *in vitro* and *in vivo* experimental data for calibration and validation purposes. Quantitative cardiac models have been an essential tool not only for basic research, but have been proposed for transition into clinical and safety-critical applications [2, 3, 4, 5, 6]. The translation of cardiac mathematical models for such applications will require high levels of credibility of predictive model outputs. One important phase of cardiac model credibility assessment is the study of uncertainty.

Parameters in cardiac models are often uncertain, mainly due to measurement uncertainty and/or natural physiological variability [7]. Thus, uncertainty quantification (UQ) methods are required to study uncertainty propagation in these models and help to establish confidence in model predictions. Parametric UQ is the process of determining the uncertainty in model inputs or parameters, and then estimating the resultant uncertainty in model outputs. This tests the robustness of model predictions given our uncertainty in their inputs.

There has been increasing recent interest and research into UQ of cardiac models to outline their predictive capability and credibility [8, 9, 10, 11, 12, 13]. However, another major source of uncertainty in modelling is uncertainty in the model structure itself. There is always a difference between the imperfect model used to approximate reality, and reality itself; this difference is termed model discrepancy. What has received little attention in this field (and mathematical/systems biology more generally) is assessment of robustness of model predictions given our uncertainty in the model structure, and methods to characterise model discrepancy.

We have found only one published explicit treatment of discrepancy in cardiac electrophysiology models, in papers by Plumlee *et al.* [14, 15]. In these studies, the assumption of ion channel rate equations following an explicit form (such as that given, as we will see later, by Eq. (8)) was relaxed, and rates were allowed to be Gaussian processes (GPs) in voltage. A two-dimensional GP (in time and voltage) was then also added to the current prediction to represent discrepancy in current for a single step to any fixed voltage.

## 1.2 Notation and terminology

Before discussing model discrepancy in detail, we introduce some notation and terminology. As the concepts introduced here are intended to be understood not just by a cardiac modelling audience, we provide a non-exhaustive list of terminology we have encountered in different fields to describe useful concepts relating to calibration and model discrepancy (and mathematical/computational modelling in general) in Table 1.

We here delve into some of those concepts in more detail. Suppose a physiological system is modelled as $y = f(\boldsymbol{\theta}, u)$, where $f$ represents all governing equations used to model the system (also referred to as model form or model structure), $\boldsymbol{\theta}$ is a vector of parameters characterising the system, and $u$ are known externally applied conditions or control variables applied in the particular experimental procedure. In a cardiac modelling context, these might represent a stimulus protocol, a drug concentration, or the applied voltage protocol in a simulated voltage clamp experiment. In general, $\boldsymbol{\theta} = \{\boldsymbol{\theta}_D, \boldsymbol{\theta}_C\}$, where values of $\boldsymbol{\theta}_D$ are directly measured (or determined using a different model, as is often the case when inheriting already-parameterised equations for particular currents into cardiac action potential models [16]) and are not determined using the model $f$, and where values of $\boldsymbol{\theta}_C$ are determined by calibration using the model $f$. Here, for simplicity of exposition, we assume $\boldsymbol{\theta}_D$ is fixed (and known) and $\boldsymbol{\theta} = \boldsymbol{\theta}_C$.

We can distinguish between external conditions used for calibration, validation, and prediction (that is, the application of the model, or context of use (CoU)), $u_C$, $u_V$, $u_{CoU}$, say. Suppose we have experimental data $Y_C$ for calibration and $Y_V$ for validation. A typical workflow, without UQ, is:

- **Calibration**: find $\boldsymbol{\theta}^* = \text{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} d_C\left(L(f(\boldsymbol{\theta}, u_C)), L(Y_C)\right)$, using some calibration metric $d_C(\cdot,\cdot)$ (e.g. it can be as simple as a vector norm: $d_C(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$), some postprocessing operator $L$, and some subset of parameter space $\boldsymbol{\Theta}$;

Table 1: Terminology used in different fields to refer to inverse problem concepts.

| CONCEPT | TERMINOLOGIES | |
|---|---|---|
| Fitting parameters in a given model to data | Calibration<br>Parameter inference<br>Parameter estimation<br>Parameterisation<br>Parameter fitting | Inverse problem<br>Parameter identification<br>Parameter tuning<br>Parameter optimisation<br>Model matching/fitting |
| Do data from given experiment provide sufficient information to identify the model parameters? | Parameter identifiability<br>Structural identifiability | Practical identifiability<br>Well-posedness |
| Altering experiments to help with inverse problems | Optimal experimental design | Protocol design |
| Choosing model equations | Model selection<br>System identification | Model choice |
| The difference between model and reality | Model discrepancy<br>Model misspecification<br>Model inadequacy<br>Structural error | Model uncertainty<br>Model mismatch<br>Model form error<br>Model structure error |
| The observable measurements (data) | Observables<br>Quantities of Interest (QoIs) | Observable outputs |
| A faster simplified version of the simulator/model | Surrogate model<br>Proxy<br>Look-up table | Metamodel<br>Emulator |
| Checking the performance of the fitted model | Validation<br>Qualification | Certification<br>Performance estimation |

- **Validation**: compare $y_V = f(\boldsymbol{\theta}^*, u_V)$ against $Y_V$, either qualitatively or using a suitable validation metric $d_V(f(\boldsymbol{\theta}^*, u_V), Y_V)$;
- **Context of use**: compute $y_{CoU} = f(\boldsymbol{\theta}^*, u_{CoU})$, or some quantity derived from this, to learn about the system or to make a model-based decision.

The calibration stage has many different names, see Table 1 for different terms.

Consideration of uncertainty in the parameters (parametric UQ) introduces additional complexity. There are various possibilities that should be distinguished. One potential source of uncertainty in the 'correct' value of $\boldsymbol{\theta}$ is measurement error in $Y_C$. Computing the uncertainty about $\boldsymbol{\theta}$ based on measurement error in $Y_C$ is referred to as 'inverse UQ', and requires a model of the experimental error to be specified. For example, a common choice is to assume independent zero-mean Gaussian measurement error which is the same on all data points, in which case (neglecting model discrepancy; see later) our model for the $i^{\text{th}}$ data point ($i^{\text{th}}$ realisation of the vector $Y_C$) is:

$$(Y_C)_i = f_i(\boldsymbol{\theta}, u_C) + \epsilon_i, \tag{1}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. There are many methods available to solve inverse UQ problems, which we cannot review in any detail here, but most are based on inferring or approximating a probability distribution of possible parameter sets that would be in agreement with the available data. Though a number of methods to solve inverse UQ problems have been applied in cardiac electrophysiology [7], the most common is a Bayesian approach, which combines prior information about the parameters, $\pi(\boldsymbol{\theta})$, with the likelihood of the data given each parameter $\pi(Y_C \mid \boldsymbol{\theta})$, to find the posterior distribution over the parameters:

$$\pi(\boldsymbol{\theta} \mid Y_C) = \frac{\pi(Y_C \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(Y_C)}. \tag{2}$$

For an introduction to Bayesian methods, see [17, 18]. For the above i.i.d. Gaussian error model (Eq. (1)), the likelihood is given by

$$\pi(Y_C \mid \boldsymbol{\theta}) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{||Y_C - f(\boldsymbol{\theta}, u_C)||_2^2}{2\sigma^2}\right), \tag{3}$$

where $||x||_2^2 = \sum_i x_i^2$, and $n$ is the number of data points.

A second potential source of uncertainty about $\boldsymbol{\theta}$ can occur when the parameter varies across the (or a) population. Estimating population variability in $\boldsymbol{\theta}$ requires multiple $Y_C$ recordings, $\{Y_C^{(1)}, Y_C^{(2)}, \ldots\}$. Multilevel or hierarchical models can then be used: we assume the parameters for population $i$ are drawn from some distribution $\boldsymbol{\theta}^{(i)} \sim \pi(\boldsymbol{\theta} \mid \psi)$, and infer the population parameters $\psi$, see [19].

Once uncertainty in $\boldsymbol{\theta}^*$ has been determined, the impact of this uncertainty on validation simulations $y_V$ or CoU simulations $y_{CoU}$ can be computed by propagating the uncertainty through the model $f$ in the validation/CoU simulations. This is referred to as 'uncertainty propagation' or 'forward UQ'. Uncertainty in $y_V$ helps provide a more informed comparison to $Y_V$ in the validation stage (especially if experimental error in $Y_V$ is also accounted for). Uncertainty in $y_{CoU}$ enables a more informed model-based decision-making process.

### 1.3 Model discrepancy

UQ as outlined above does not account for the fact that the model is always an imperfect representation of reality, due to limited understanding of the true data-generating mechanism and perhaps also any premeditated abstraction of the system. The model discrepancy is the difference between the model and the 'true' data-generating mechanism, and its existence has implications for model selection, calibration and validation, and CoU simulations.

For calibration, the existence of model discrepancy can change the meaning of the estimated parameters. If we fail to account for the model discrepancy in our inference, our parameter estimates, instead of being physically meaningful quantities, will have their meaning intimately tied to the model used to estimate them. The estimated parameter values depend on the chosen model form, and the uncertainty estimates obtained during inverse parameter UQ tell us nothing about where the 'true' value is. In other words, there is no guarantee the obtained $\boldsymbol{\theta}$ will match 'true' physiological values of any parameters which have a clear physiological meaning.

We can try to restore meaning to the estimated parameters by including a term to represent the model discrepancy in our models. Validation in particular, provides an opportunity for us to identify possible model discrepancy. In fact, validation, rather than being considered as an activity for confirming a 'model is correct', is better considered as a method for estimating the model discrepancy. To maximise the likelihood that the validation can discern model discrepancy, the validation data should ideally be 'far' from the calibration data, and as close to the CoU as possible.

## 2 A motivating example of discrepancy

To illustrate the concept of model discrepancy and some of its potential consequences we have created a cardiac example inspired by previous work [20]. We will assume that the Ten Tusscher *et al.* ventricular myocyte electrophysiology model [21] (Model T) represents the ground truth. We use Model T to generate some data traces in different situations: (i) the action potential under 1 Hz pacing; (ii) under 2 Hz pacing; and (iii) under 1 Hz pacing with 75% $I_{\mathrm{Kr}}$ block ($g_{Kr}$ is multiplied by a scaling factor of 0.25). We split the data traces into calibration (i), validation (ii), and 'context of use' (CoU) (iii). We might construct a candidate model based on comparison with only the calibration and validation dataset, and then use it to predict the CoU situation.

So we assume we do not know the ground truth and instead fit an alternative model, the Fink *et al.* model [22] (Model F), to the synthetic data generated from Model T. Since both models were built for human ventricular cardiomyocytes, comparing these two models is an example which highlights the potential problems associated with fitting any such model under model discrepancy. Note, that Model F is a modification of Model T which improves the descriptions of repolarising currents (see Supplementary Section S1), especially of hERG (or $I_{\mathrm{Kr}}$), as this channel is a major focus for Safety Pharmacology.

We use Model T to generate synthetic current clamp experiments by simulating the different protocols then adding i.i.d. Gaussian noise $\sim \mathcal{N}(0, \sigma^2)$ to the resulting voltage traces, with $\sigma$ chosen to be 1 mV. We then fit eight maximal conductance/current density parameters for $I_{\mathrm{Na}}$, $I_{\mathrm{CaL}}$, $I_{\mathrm{Kr}}$, $I_{\mathrm{Ks}}$, $I_{\mathrm{to}}$, $I_{\mathrm{NaCa}}$, $I_{\mathrm{K1}}$, and $I_{\mathrm{NaK}}$ to the synthetic data. Such an exercise assumes that the parameters and equations describing the kinetics/gating of the ion currents do not need to be re-calibrated to this particular dataset and have been fitted previously (a common assumption in electrophysiology modelling [23, 24, 25, 26, 27]). Here, we know that this assumption is incorrect, as we purposefully introduced discrepancy between the current kinetics.

A comparison of the differences in current kinetics between Models T and F is shown in Figure 1, also see equations for each current in Supplementary Section S1 for reference. Only five currents have kinetics

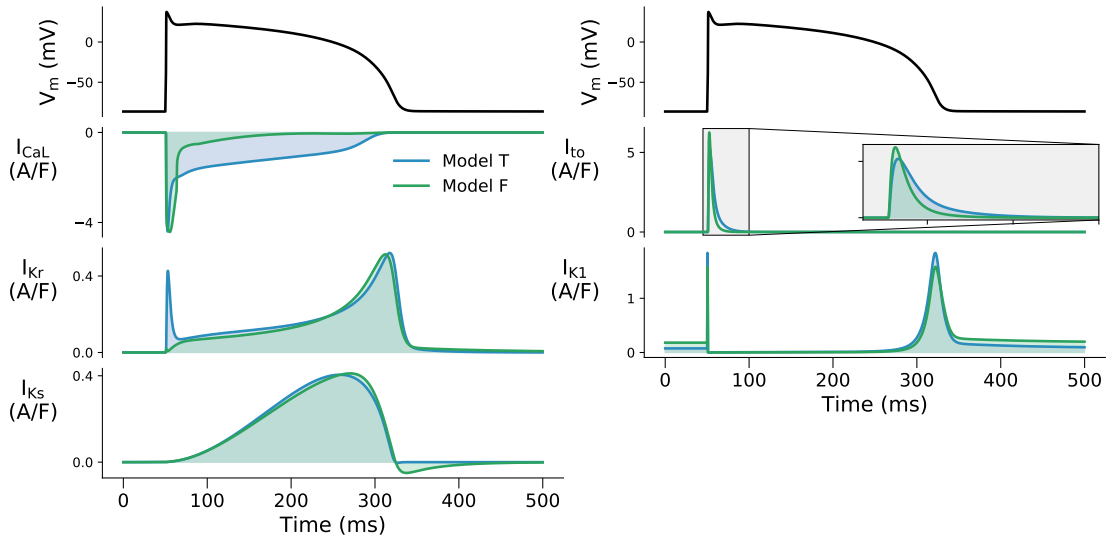


Figure 1: A comparison of Ten Tusscher (Model T [21], blue) and Fink (Model F [22], green) kinetics. These currents are voltage clamp simulations under the same action potential clamp, shown on the top panels. Here only those currents with different kinetics are shown; the kinetics of $I_{\mathrm{Na}}$, $I_{\mathrm{NaCa}}$, and $I_{\mathrm{NaK}}$ are identical in both models. Two of the gates in $I_{\mathrm{CaL}}$ are identical in the two models, one gate has a different formulation, and Model F has one extra gate compared to Model T. The two models use different formulations for $I_{\mathrm{Kr}}$, different parameterisations of the kinetics for $I_{\mathrm{Ks}}$ and $I_{\mathrm{to}}$, and different equations for $I_{\mathrm{K1}}$ steady state. Currents are normalised in this plot by minimising the squared-difference between the two models' currents such that we emphasise the differences in kinetics rather than the conductances (which are rescaled during the calibration). Only $I_{\mathrm{CaL}}$ shows what we would typically consider to be a 'large' difference in kinetics, with the rest of the currents apparently being close matches between Model T and Model F.

that vary between the two models, and importantly no currents or compartments are missing (unlike when attempting to fit a model in reality). We will investigate whether we can trust the calibrated Model F to make good predictions in new situations.

The code to reproduce the results in this example is available at https://github.com/CardiacModelling/fickleheart-method-tutorials.

## 2.1 Model calibration

We calibrate the model using a train of five action potentials stimulated under a 1 Hz pacing protocol as the calibration data. Before attempting to do this fitting exercise, the appropriately sceptical reader might ask whether we are attempting to do something sensible. Will we get back information on all the parameters we want, or will we just find one good fit to the data amongst many equally plausible ones, indicating non-identifiability of parameters?

To address these questions we first look at inferring the parameters of the original Model T. We use Eq. (1) with Gaussian noise giving the likelihood in Eq. (2), together with a uniform prior distribution from 0.1× to 10× the original parameters of Model T (additionally inferring the noise model parameter, $\sigma$). We perform both non-UQ based and UQ-based calibration: for the first, we use a global optimisation algorithm [28] to find the optimal model parameters; for the second, we generate samples to approximate the full posterior distribution using Markov chain Monte Carlo (MCMC). All inference is done using an open source Python package, PINTS [29], and simulations are performed in Myokit [30].

The results are shown in Supplementary Figure S1. This exercise results in a narrow plausible distribution of parameters very close to the ones that generated the data. Therefore, the 'true' model's parameters are identifiable with the given data. Additionally, Supplementary Figure S1 shows that when using samples of these distributions to make predictions, all of the forward simulations are very closely grouped around the synthetic data for the $I_{\mathrm{Kr}}$ block CoU.

We now move on to attempt the fitting exercise using Model F. The fitting result, the maximum a-

posteriori probability (MAP) estimate, is shown in Figure 2 (Top). The agreement between the calibrated model output and the synthetic data would be considered excellent if these were real experimental data. Therefore, it would be tempting for modellers to conclude that this calibrated model were a good model, due to small model discrepancy. However, can we truly trust the predictive power of the model based on the result we see in Figure 2 (Top)?

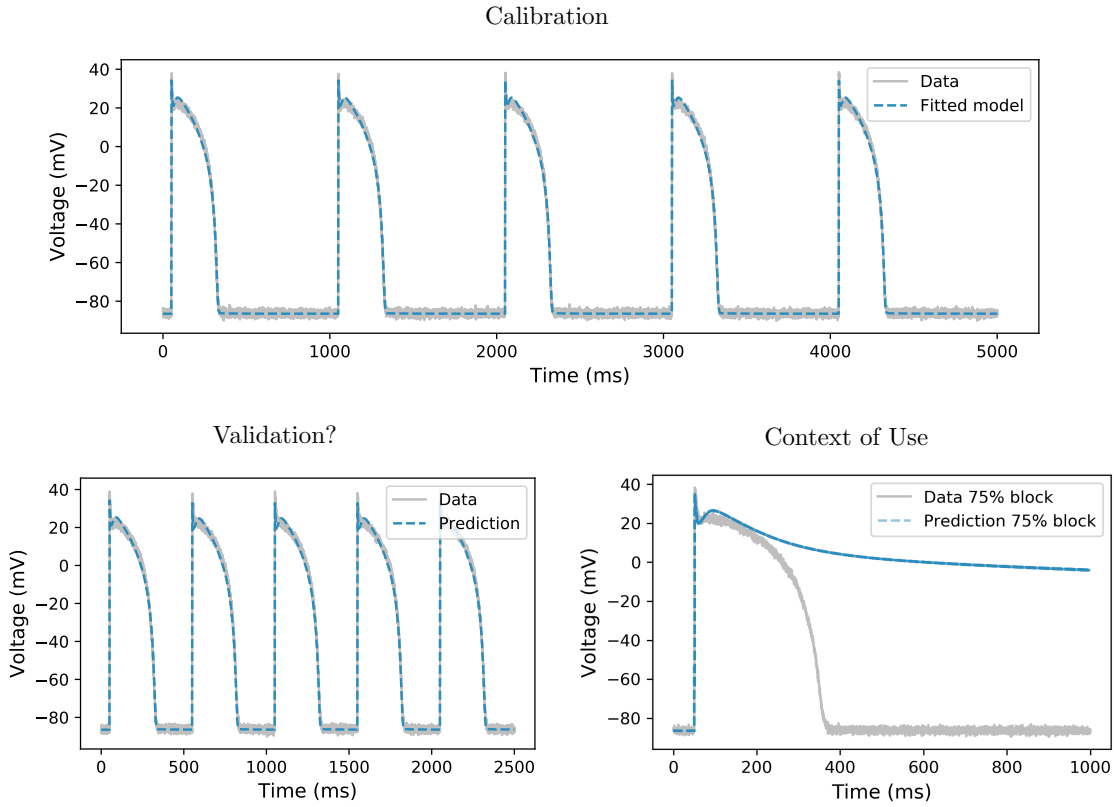


Figure 2: **Model F fitting and validation results**. (**Top**) The model is fitted to the ground truth synthetic data (generated from Model T), using a five action potential recording under a 1 Hz pacing protocol. The calibrated Model F (blue dashed line) shows an excellent fit to the data (grey solid line). (**Bottom**) Model F validation and prediction for a different context of use (CoU). (*Left*) The calibrated Model F can match some validation data (2 Hz pacing) very well, giving us a (false) confidence in the model performance. (*Right*) Notably, despite the calibrated Model F giving an excellent fit and validation, it gives a catastrophic (posterior sample) prediction for the $I_{\mathrm{Kr}}$ block (CoU) experiments (suggesting in this example (*Left*) is not a good validation experiment!). The posterior predictions are model predictions simulated using the parameter samples from the posterior distribution (Figure 3); here, 200 samples/predictions are shown.

## 2.2 Discrepant model predictions

Interestingly, the calibrated Model F gives an excellent validation prediction for a 2 Hz pacing protocol, as shown in Figure 2 (Left). Such rate-adaptation predictions are used commonly as validation evidence for action potential models. At this stage we might be very tempted to say we have a good model of this system's electrophysiology.

But if one now uses the model to predict the effect of drug-induced $I_{\mathrm{Kr}}$ block, the catastrophic results are shown in the bottom panel of Figure 2 (Right). The calibrated Model F fails to repolarise, completely missing the 'true' $I_{\mathrm{Kr}}$ block response of a modest APD prolongation. This example highlights the need for thorough validation and the CoU-dependence of model validation, but also the difficulty in choosing appropriate validation experiments.

We quantify uncertainty in parameters and predictions whilst continuing to ignore the discrepancy in Model F's kinetics. Again, we use Eq. (2) together with a uniform prior to derive the posterior of the parameters. The posterior distributions estimated by MCMC and the point estimates obtained by optimisation, are shown in Figure 3. The posterior distribution is very narrow (note the scale), which suggests that we can be confident about the parameter values. The resulting posterior predictions, shown in Figure 2 (Right), give a very narrow bound. Therefore, by ignoring model discrepancy we could become very (and wrongly) certain that the catastrophically bad predictions are correct.

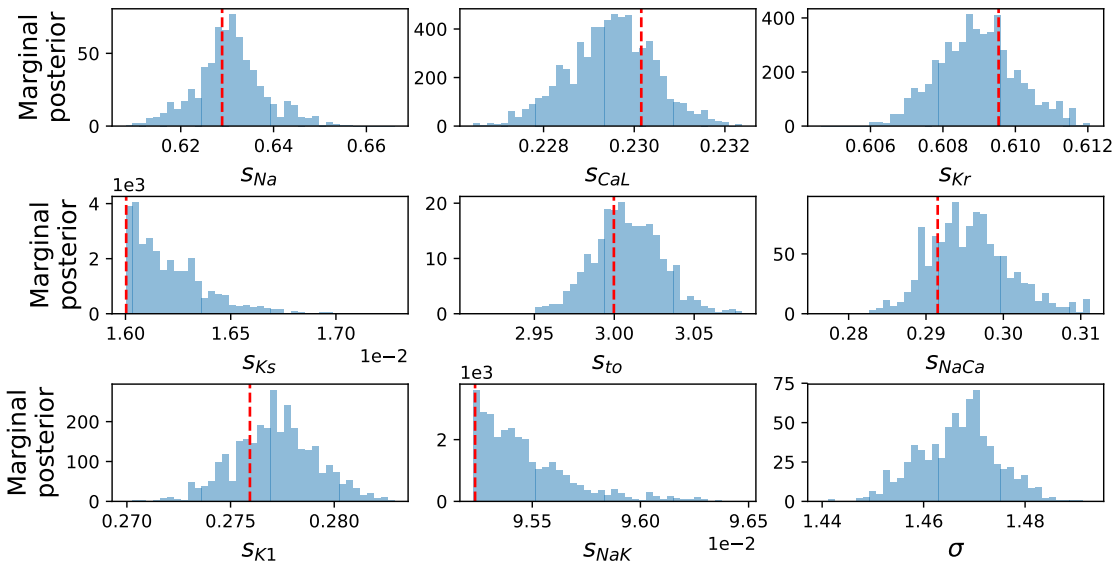


Figure 3: Marginal posterior distribution of Model F parameters, in terms of scaling factors for the conductances in Model T ($s_i = g_i^{\text{Model F}}/g_i^{\text{Model T}}$). Values of 1 would represent the parameters of Model T that generated the data; notably, none of the inferred parameters for Model F is close to a value of 1. The red dashed lines indicate the result of the global optimisation routine. Note that two of these parameters, $S_{Ks}$ and $S_{NaK}$, have distributions hitting the lower bound that was imposed by the prior, indicating that the calibration process is attempting to make them smaller than 10% of the original Model F parameter values.

It is worth noting that all the issues above arise from the fact that model discrepancy was ignored. In the scenario of no model discrepancy, i.e. when fitting Model T to the data, all of the issues above were solved, as shown in Supplementary Figure S1.

To conclude our motivation of this paper, we can see that neglecting discrepancy in the model's equations is dangerous and can lead to false confidence in predictions for a new context of use, so we will discuss methods that have been suggested to remedy this.

## 2.3 A statistical explanation of the issue

To understand what is happening, consider the well-specified situation where the data generating process (DGP, the process that produces the experimental data in reality) has probability density function (pdf) $g(y)$, and for which we have data $y_i \sim g(\cdot)$ for $i = 1, \ldots, n$. Then suppose we are considering the models $\mathcal{F} = \{f_{\boldsymbol{\theta}}(y) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, i.e., a collection of pdfs parameterized by unknown parameter $\boldsymbol{\theta}$. If the DGP $g$ is in $\mathcal{F}$, i.e., we have a well-specified model so that for some $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, we have $g = f_{\boldsymbol{\theta}_0}$, then asymptotically, as we collect more data (and under suitable conditions [31]), the maximum likelihood estimator converges to the true value $\boldsymbol{\theta}_0$ almost surely:

$$\hat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{n} \log f_{\boldsymbol{\theta}}(y_i) \longrightarrow \boldsymbol{\theta}_0, \text{ almost surely as } n \longrightarrow \infty,$$

or equivalently $f_{\hat{\boldsymbol{\theta}}_n}$ converges to $g$. Similarly, for a Bayesian analysis (again under suitable conditions [32]), the posterior will converge to a Gaussian distribution centered around the true value $\boldsymbol{\theta}_0$, with

variance that shrinks to zero at the asymptotically optimal rate (given by the Cramér-Rao lower bound), i.e.

$$\pi(\boldsymbol{\theta} \mid y_{1:n}) \overset{d}{\Rightarrow} \mathcal{N}(\boldsymbol{\theta}_0, \frac{1}{n}\mathcal{I}(\boldsymbol{\theta}_0)^{-1}),$$

where $y_{1:n} = (y_1, \ldots, y_n)$, and $\mathcal{I}(\boldsymbol{\theta}_0)$ is the Fisher information matrix.

However, when our model is misspecified, i.e., $g \notin \mathcal{F}$ (there is no $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ for which $g = f_{\boldsymbol{\theta}}$), if we do inference for $\boldsymbol{\theta}$ ignoring the discrepancy, then we usually still get asymptotic convergence of the maximum likelihood estimator and Bayesian posterior [33, 34]. However, instead of converging to a true value (which does not exist), we converge to the *pseudo-true* value

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmin}} \operatorname{KL}(g||f_{\boldsymbol{\theta}})$$

where $\operatorname{KL}(g||f) = \int g(x) \log \frac{g(x)}{f(x)} \mathrm{d}x$ is the Kullback-Leibler divergence of $f$ from $g$ (a measure of the difference between two distributions). In other words, we converge upon the model, $f_{\boldsymbol{\theta}^*}$, which is closest to the DGP as measured by the Kullback-Leibler divergence (see Figure 4).

Perhaps more importantly from a UQ perspective, as well as getting a point estimate that converges to the wrong value, we still usually get asymptotic concentration at rate $1/n$, i.e., the posterior variance shrinks to zero. That is, we have found model parameters that are wrong, and yet we are certain about this wrong value. The way to think about this is that the Bayesian approach is not quantifying our uncertainty about a meaningful physical parameter $\boldsymbol{\theta}_0$, but instead, it gives our uncertainty about the pseudo-true value $\boldsymbol{\theta}^*$. Consequently, we can not expect our calibrated predictions

$$\pi(y' \mid y) = \int f_{\boldsymbol{\theta}}(y')\pi(\boldsymbol{\theta} \mid y_{1:n})\mathrm{d}\boldsymbol{\theta}$$

to perform well, as we saw in the action potential example above.

This leaves us with two options. We can either extend our model class $\mathcal{F}$ in the hope that we can find a class of models that incorporates the DGP (and which is still sufficiently simple that we can hope to learn the true model from the data), or we can change our inferential approach.

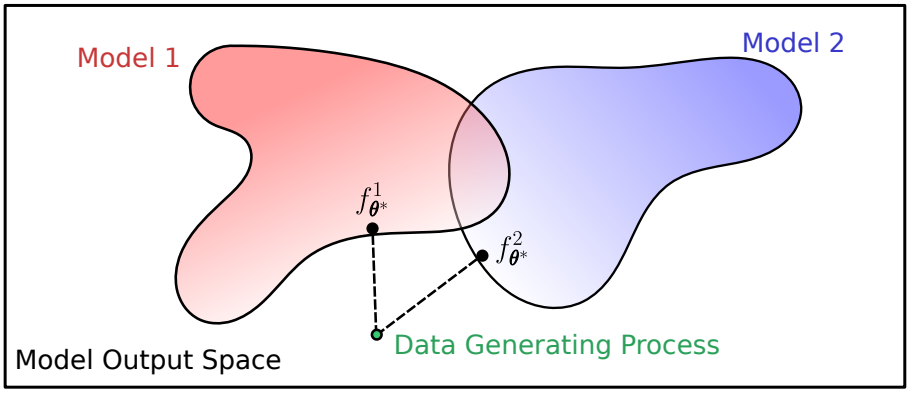


Figure 4: **A cartoon to illustrate the effect of model discrepancy on parameter fits in different models**. Each cloud represents a range of possible outputs from each model, which they can reach with different parameter values. The true data generating process (DGP) lies outside either of our imperfect model classes 1 and 2, and neither can fit the data perfectly due to model discrepancy. When we attempt to infer parameters, we will converge upon models that generate outputs closest to the true DGP under the constraint of being in each model. Adding more data just increases the confidence in being constrained to model parameterisations on the boundary of the particular model. This, however, does not move the model parameters closer to the true parameters used in the data generating process, i.e. we become certain about $f_{\boldsymbol{\theta}^*}$, the model using the pseudo-true parameter value. Note that a different experiment will lead to a different model output, shifting these clouds to a different shape, and then different parameter sets may give outputs that are closer to the data generating process.

# 3 Accounting for model discrepancy

Once we have acknowledged that a model is misspecified, we are then faced with the challenge of how to handle the misspecification. The approach taken should depend upon the aim of the analysis. Using the model to predict independent events, for example, a current time-series for some experimental protocol, will require a different approach to if our aim is inference/calibration, i.e., if interest lies in the physical value of a particular parameter. In the first case (prediction), it can often suffice to fit the model to the data ignoring discrepancy, and then to correct the predictions in some way[1], although this may not work well if the prediction involves extrapolating into a regime far away from the data. The latter case (calibration) is more challenging, as we need to jointly fit the model and the discrepancy model, which can lead to problems of non-identifiability.

The most common approach for dealing with discrepancy is to try to correct the simulator by expanding the model class. The simplest approach is simply to add a flexible, non-parametric term to the simulator output, i.e. instead of assuming the data arose from Eq. (1), to assume

$$y = f(\boldsymbol{\theta}, u_C) + \delta(v_C) + \epsilon. \tag{4}$$

Here, $\delta(v_C)$ is the model discrepancy term, and $\epsilon$ remains an unstructured white noise term. Note that $v_C$ is used as the input to $\delta$ as it is not necessary to have the same input as the mechanistic model. To train this model, one option is to first estimate $\theta^*$ assuming Eq. (1), and then to train $\delta$ to mop up any remaining structure in the residual. However, a better approach is to jointly estimate $\delta$ and $\theta$ in a Bayesian approach [35]. Unfortunately, as demonstrated below, this often fails as it creates a non-identifiability between $\boldsymbol{\theta}$ and $\delta$ when $\delta$ is sufficiently flexible: for any $\boldsymbol{\theta}$, there exists a functional form $\delta(\cdot)$ for which Eq. (4) accurately represents the data generating process. Brynjarsdóttir et al. [20] suggested that the solution is to strongly constrain the functional form of $\delta(\cdot)$ using prior knowledge. They present a toy situation in which $\delta(0) = 0$ and $\delta(x)$ is monotone increasing, and show that once armed with this knowledge, the posterior $\pi(\boldsymbol{\theta} \mid y)$ more accurately represents our uncertainty about $\boldsymbol{\theta}$. However, knowledge of this form is not available in many realistic problems.

## 3.1 Ion channel model example

We now illustrate the difficulty of accounting for model discrepancy, in a tutorial example. We demonstrate that it can be hard to determine the appropriate information to sufficiently constrain $\delta$, and that different functional forms can lead to different parameter estimates.

The code to reproduce the results in the tutorials is available at https://github.com/CardiacModelling/fickleheart-method-tutorials.

We consider three structurally different models: Models A, B, and C. We take Model C as the ground truth model in this particular example, and use it to perform synthetic voltage clamp experiments and generate synthetic data. The goal is to use Models A and B to explain the generated synthetic data, assuming we have no knowledge about the ground truth Model C. This tutorial aims to demonstrate the importance of consideration of model discrepancy, jointly with model selection, to represent given data with unknown true DGP.

In this tutorial, we use the hERG channel current as an example. Models A, B, and C are various model structures proposed in the literature. Model A is a variant of the traditional Hodgkin-Huxley model, described in Beattie *et al.* [36]; Model B is a Markov model structure used in Oehmen *et al.* [37]; and Model C is a Markov model structure adapted from Di Veroli *et al.* [38]. The model structures are shown in Figure 5.

All three ion channel models can be expressed using a Markov model representation. For a model with a state vector, $\boldsymbol{y} = (y_1, y_2, \cdots)^T$, then $\boldsymbol{y}$ evolves according to

$$\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}t} = \mathbf{M}\boldsymbol{y}, \tag{5}$$

where $\mathbf{M}$ is the Markov matrix describing the transition rates between states. Markov models are linear coupled ordinary differential equations (ODEs) with respect to time, $t$, and states, $\boldsymbol{y}$; whilst typically the components in the Markov matrix, $\mathbf{M}$, are nonlinear functions of voltage, $V$. The observable, the

[1] Note that jointly fitting model and discrepancy can make the problem easier, for example, by making the discrepancy a better behaved function more amenable to being modelled.

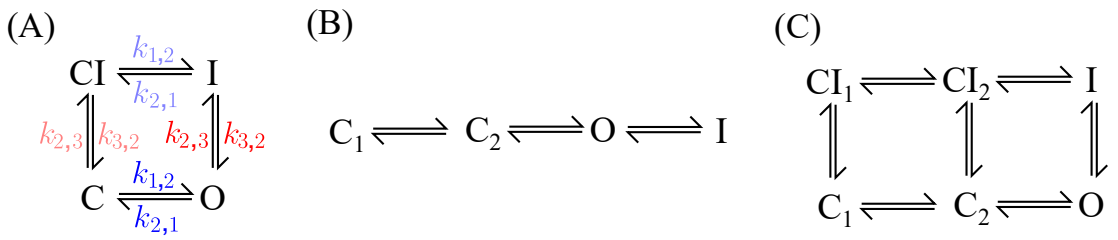


Figure 5: Markov model representation of Models A, B, and C used in the ion channel model tutorial where Model C is taken as ground truth and used to generate synthetic data whilst Models A and B are candidate models that we attempt to fit and use for predictions, demonstrating both the model discrepancy and model selection challenge.

macroscopic ionic current, $I$, measured under $V(t)$, which in these voltage-clamp experiments is an externally prescribed function of time known as the 'voltage clamp protocol' (i.e. $u_C$ in Eq. (1)), is

$$I(t, V) = g \cdot \mathcal{O} \cdot (V - E), \tag{6}$$

where $g$ is the maximum conductance, $E$ is the reversal potential, and $\mathcal{O}$ is the sum of all 'open states' in the model (frequently, and in our examples, this is just one state, but more than one open state is possible).

Take Model B as an example. Its state vector, $\boldsymbol{y}$, and Markov matrix, $\mathbf{M}$, can be written as

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} \mathrm{C}_2 \\ \mathrm{C}_1 \\ \mathrm{O} \\ \mathrm{I} \end{pmatrix}; \qquad \mathbf{M} = \begin{pmatrix} -k_{1,2} & k_{2,1} & 0 & 0 \\ k_{1,2} & -k_{2,1} - k_{2,3} & k_{3,2} & 0 \\ 0 & k_{2,3} & -k_{3,2} - k_{3,4} & k_{4,3} \\ 0 & 0 & k_{3,4} & -k_{4,3} \end{pmatrix}, \tag{7}$$

where $k_{i,j}$ represents the transition rate from state $y_i$ to state $y_j$. For all models each transition rate, $k_{i,j}$, is voltage dependent, and takes the form

$$k_{i,j}(V) = A_{i,j} \exp(B_{i,j} V), \tag{8}$$

with two parameters $(A_{i,j}, B_{i,j})$ to be inferred. This yields a total of 12 parameters for Model B which we denote as $\{p_1, \ldots, p_{12}\}$, together with the maximum conductance, $g$, to be found. Similarly for Model A, it has 8 parameters $\{p_1, \ldots, p_8\}$ together with $g$, to be inferred.

### 3.2 Synthetic experiments

We say that Model C represents our (hidden) ground truth and simulate data from it under multiple voltage clamp protocols, using parameters fitted to room temperature data from Beattie *et al.* [36] (where $g = 204\,\mathrm{nS}$). We introduce i.i.d. Gaussian noise with a standard deviation $\sigma = 25\,\mathrm{pA}$ to the simulated data. We generate data under three different voltage clamp protocols, $V(t)$. These are a sinusoidal protocol (see top plot in Figure 6) and an action potential series protocol from Beattie *et al.* [36] (see Figure S5 in Supplementary Material), and the staircase protocol from Lei *et al.* [19, 39] (see bottom plot in Figure 6).

### 3.3 Standard calibration ignoring model discrepancy

To calibrate the model (without considering any model discrepancy), we assume a statistical model of the form of Eq. (1), which has the same observation noise model as our synthetic data. The likelihood of observing the data, $\mathbf{y} = y_{1:n}$, given model parameters $\boldsymbol{\theta}$ is given by Eq. (3).

We use the sinusoidal protocol as the calibration protocol; the action potential series protocol and the staircase protocol are used as validation. We employ a global optimisation algorithm [28] to fit the model parameters, and all inference is done using PINTS [29].

The fitting results of Models A and B are shown in Figure 6. Repeats of fitting with different starting points gave almost the same parameter sets. Although both models fit the calibration data reasonably well, neither match perfectly, due to model discrepancy. While the exact forms of the model discrepancy

differ between the two models, both models notably fail to reproduce the correct form of the current decay following the step to −120 mV shortly after 2000 ms.

The validation predictions for the staircase protocol are also shown in Figure 6. Unlike in the sinusoidal protocol, where Model A generally gives a better prediction than Model B, in the staircase protocol different traits of model discrepancy arising from different models are more evident. For example, whereas Model B appears to give slightly better predictions of the current during the first 10 000 ms, after this point Model A begins to give better predictions.

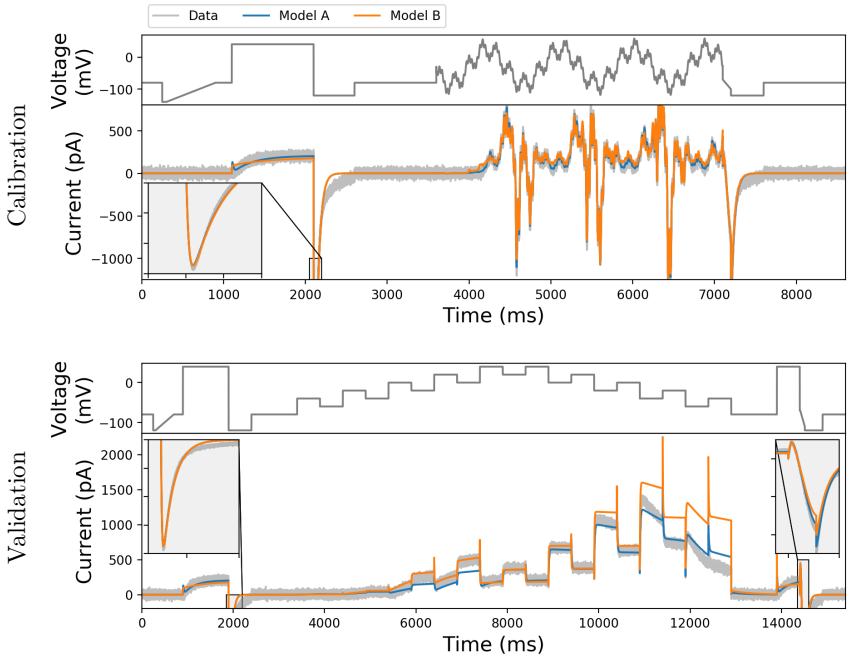


Figure 6: **(Top)** Models A (blue) and B (orange) fitting results for the ion channel model example. The two models are fitted to the same synthetic data (grey) generated using a third model, Model C, with i.i.d. Gaussian noise added to it. The voltage clamp protocol for calibration is the sinusoidal protocol [36]. **(Bottom)** Models A (blue) and B (orange) validation results for the ion channel model tutorial. We show predictions for the two fitted models for the unseen ground truth (grey), generated from Model C under the staircase protocol [19]. Note that there are significant discrepancies around 12 000 ms.

### 3.4 Calibration with model discrepancy

Next we consider methods that allow us to incorporate or acknowledge the model discrepancy when doing parameter inference and model predictions. First, we adapt the method proposed in [35] and instead of assuming Eq. (1), we consider an additive discrepancy model of the form given by Eq. (4). We consider three different choices for the discrepancy $\delta(v_C)$, and jointly infer $\theta$ and $\delta$. Note that we allow for a different choice of input $v_C$, compared to the input of model $f$, $u_C$.

First, we model $\delta$ as a sparse-Gaussian process (GP), for which we adapted the implementation in PyMC3 [40] using Theano [41]. We explored two possibilities, choosing $v_C$ to be either (i) $t$ (time); or (ii) $O$,$V$ (the open probability, $\mathcal{O}$ in Eq. (6), and the voltage, $V$). For details of the method, please refer to Supplementary Section S2.

Second, we model discrepancy $\delta$ and the white noise error $\epsilon$, as an autoregressive-moving-average (ARMA) model of order $p, q$ [42]. If $e_t = \delta_t(v_c) + \epsilon_t$ is the residual at time $t$, then an ARMA$(p, q)$ model for $e_t$ is

$$e_t = \nu_t + \sum_{t'=1}^{p} \varphi_t e_{t-t'} + \sum_{t'=1}^{q} \zeta_{t'} \nu_{t-t'} \tag{9}$$

where $\nu \sim \mathcal{N}(0, \tau^2)$, and $\varphi_t, \zeta_t$ are, respectively, the coefficients of the autoregressive and moving-average

part of the model. We used the StatsModels [43] implementation, and assumed $p = q = 2$ throughout. Note that the ARMA model, unlike the GP, does not attempt to learn any structure for the discrepancy (the mean of the ARMA process remains zero even once conditioned upon data). The ARMA process is a simple approach for introducing correlation into the residuals. The motivation is that if the mechanistic model is correct, the residuals should be uncorrelated, but for misspecified models, they are typically correlated. For further details of the method, please refer to supplementary information, Section S3.

For all methods, i.i.d. noise, GP($t$), GP($O, V$), and ARMA(2, 2), we infer the parameters using Eq. (2), where the priors are specified in Supplementary Section S4. The posteriors are approximated by using an Adaptive Covariance MCMC method in PINTS [29, 25]. The inferred (marginal) posterior distributions for Model A are shown in Figure 7, and they are used to generate the posterior predictive shown below. Supplementary Figure S12 shows the same plot for Model B. Note that the choice of the discrepancy model can shift the posterior distribution significantly, both in terms of its location and spread. In particular, the ARMA(2, 2) model gives a much wider posterior than the other discrepancy models.

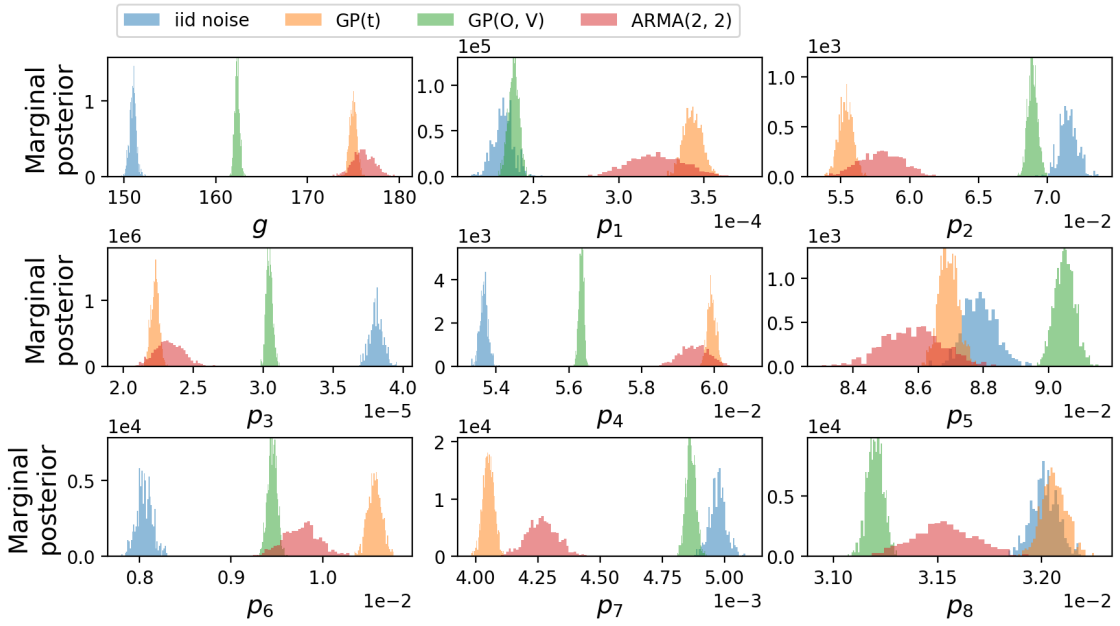


Figure 7: Model A inferred marginal posterior distributions for the conductance, $g$ in Eq. (6), and kinetic parameters $p_1, \ldots, p_8$ (a list of parameters referring to $A_{i,j}$ and $B_{i,j}$ in Eq. (8)) with different discrepancy models: i.i.d. noise (blue), GP($t$) (orange), GP($O, V$) (green), and ARMA(2, 2) (red).

Figure 8 shows the posterior predictives of the calibration for Model A using different discrepancy models (Supplementary Figure S13 for Model B). The top panel shows the sinusoidal calibration voltage protocol, and the panels underneath are calibrated models with i.i.d. noise (blue), GP($t$) (orange), GP($O, V$) (green), and ARMA(2, 2) (red). The calibration data are shown in grey in each panel. Visually, we can already see that the two GP models, GP($t$) (orange) and GP($O, V$) (green), are able to fit to the data with very high accuracy; we will see one of them is overfitting while the other is not. The ARMA(2, 2) model (red) is able to increase the width of the posterior, although its posterior mean prediction does not follow the data as closely as the two GP models.

Tables 2 and 3 show two quantifications of the goodness of fits, which is coloured so that yellow shows the best performing model and red shows the worst. Table 2 shows the RMSE values of the posterior mean predictions for all of the models, and Table 3 shows the marginal log-likelihoods (a proper scoring rule [44], which assesses the entire predictive distribution, not just the mean). The first row of the two tables shows the results for calibration (sine wave), and it is clear that the GP($t$) and GP($O, V$) models give the best RMSE values, while the ARMA(2, 2) and GP($O, V$) models give the best marginal log-likelihoods.

Figure 9 shows the prediction results for the staircase validation protocol for Model A (Supplementary Figure S14 for Model B) using different discrepancy models, with the same layout as Figure 8. Similar

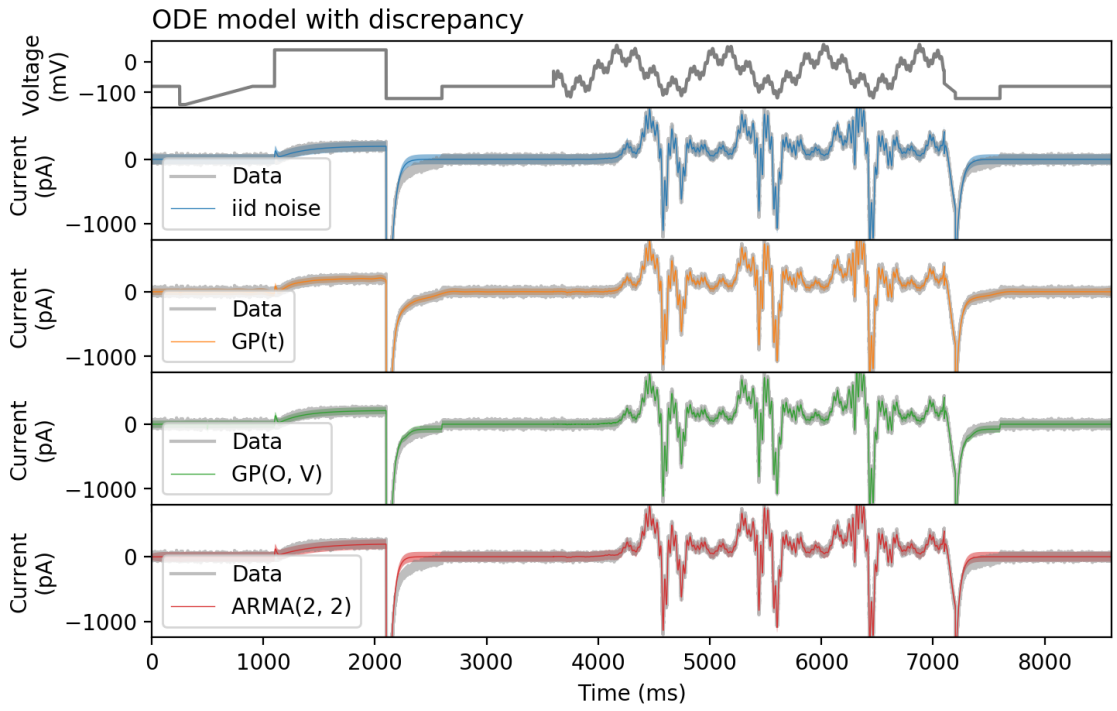


Figure 8: Model A fitted to the sinusoidal calibration protocol using the different discrepancy models: i.i.d. noise, GP($t$), GP($O, V$), and ARMA(2, 2). The plots show the mean (solid lines) and the 95% credible intervals (shaded) of the posterior predictive for each model.

figures for AP protocol predictions are shown in Supplementary Figures S5 (for Model A) and S15 (for Model B). GP($t$) does not have any information that we have changed from calibration protocol to prediction protocol, and it 'predicts' as if it were still under the sinusoidal protocol. Thus, there is some residual from the calibration shown in the GP($t$) (orange) prediction, e.g. see 'wobbly' current at $\sim 7000$ ms as pointed at by the blue arrow.

For Model A, it is interesting to see that none of the discrepancy models are able to predict the data better than the i.i.d. noise model according to the RMSE value of the posterior mean prediction in Table 2, and the next best is the GP($O, V$) model. However, it is also interesting to notice that the GP($O, V$) model is able to capture and predict very nicely the tail current after the two activation steps, as indicated by the red arrows on Figure 6 — a visible area of model mismatch in our calibration without model discrepancy.

For Model B, the GP($O, V$) discrepancy model gives the best overall predictions for both the staircase and the AP protocols, although when we examine the contributions of the mechanistic and discrepancy models, we see that an element of unidentifiability between them has arisen (Supplementary Section S7S7.2). In terms of the marginal log-likelihood, Table 2 (bottom) again highlights that the ARMA(2, 2) and GP($O, V$) models tend to be better than the i.i.d. noise and GP($t$) models.

Supplementary Figures S6, S7, and S8 show the model discrepancy for the sine wave protocol, AP protocol, and staircase protocol, respectively, for Model A; Supplementary Figures S16, S17, and S18 show the same plots for Model B. Supplementary Figures S8 and S18 in particular highlight that the GP($t$) model has, by design, learnt nothing of relevance about model discrepancy for extrapolation under an independent validation protocol (in which $V(t)$, and indeed the range of $t$, differs from that of the training protocol). Furthermore, the discrepancy model is based only on information extending to 8000 ms (the duration of the training protocol), after which the credible interval resorts to the width of the GP prior kernel. In contrast, GP($O, V$) learns independently of $t$ about the discrepancy under combinations of $(O, V)$ present in the training data (such as the activation step to 40 mV followed by a step to $-120$ mV), which is why it is able to better predict the tail current after the two activation steps. Finally, the ARMA(2, 2) model looks very similar to the i.i.d. noise model in terms of the 95% credible interval of the discrepancy term only, as it is defined to have zero mean. The ion channel (ODE)

| Model A | | Fitted with iid noise | | Fitted with GP(t) | | Fitted with GP(O, V) | | Fitted with ARMA(2, 2) | |
|---|---|---|---|---|---|---|---|---|---|
| | | ODE model & iid noise | ODE model only | ODE model & GP(t) | ODE model only | ODE model & GP(O, V) | ODE model only | ODE model & ARMA(2, 2) | ODE model only |
| Calibration | sinewave | 39.41 | 39.41 | 25.73 | 48.83 | 29.32 | 40.70 | 45.03 | 45.33 |
| Prediction | staircase | 68.10 | 68.10 | 109.38 | 106.08 | 76.61 | 80.47 | 104.90 | 102.89 |
| | ap | 59.57 | 59.57 | 90.38 | 80.67 | 61.25 | 61.80 | 85.72 | 83.33 |

| Model B | | Fitted with iid noise | | Fitted with GP(t) | | Fitted with GP(O, V) | | Fitted with ARMA(2, 2) | |
|---|---|---|---|---|---|---|---|---|---|
| | | ODE model & iid noise | ODE model only | ODE model & GP(t) | ODE model only | ODE model & GP(O, V) | ODE model only | ODE model & ARMA(2, 2) | ODE model only |
| Calibration | sinewave | 47.99 | 47.99 | 27.30 | 55.76 | 33.29 | 196.89 | 56.41 | 56.18 |
| Prediction | staircase | 191.19 | 191.19 | 489.66 | 489.32 | 116.29 | 183.44 | 346.93 | 341.82 |
| | ap | 141.19 | 141.19 | 133.56 | 123.84 | 91.65 | 235.15 | 139.29 | 136.86 |

Table 2: Models A (top) and Model B (bottom) RMSE with different discrepancy models: i.i.d. noise, GP($t$), GP($O, V$), and ARMA$(2, 2)$. Here 'ODE model-only' refers to the predictions using only the calibrated ODE model under different discrepancy models, see Supplementary Figures S9–S11 for Model A and Supplementary Figures S19–S21 for Model B.

| Model A | | iid noise | GP(t) | GP(O, V) | ARMA(2, 2) |
|---|---|---|---|---|---|
| Calibration | Sinewave | -41655.39 | -1913.39 | -13891.78 | 0.00 |
| Prediction | Staircase | -609823.19 | -339259.07 | -355212.12 | 0.00 |
| | AP | -73168.91 | -218470.48 | -108060.59 | 0.00 |

| Model B | | iid noise | GP(t) | GP(O, V) | ARMA(2, 2) |
|---|---|---|---|---|---|
| Calibration | Sinewave | -60114.48 | -6673.18 | -18955.64 | 0.00 |
| Prediction | Staircase | -1909304.26 | -658100.77 | -371446.15 | 0.00 |
| | AP | -428642.83 | -383410.92 | 0.00 | -24147.85 |

Table 3: Relative log-likelihood of fits and predictions for Models A (left) and Model B (right) with different discrepancy models: i.i.d. noise, GP($t$), GP($O, V$), and ARMA$(2, 2)$. Note that only differences in log-likelihoods within a row are meaningful, we therefore subtract the maximum log-likelihood for each dataset from the results giving the best model in each row a score of zero.

model-only predictions for the sine wave protocol, AP protocol, and staircase protocol are shown in Supplementary Figures S9, S10, and S11 for Model A and Supplementary Figures S19, S20, and S21 for Model B.

Comparing the RMSE and likelihood in Tables 2 & 3, it is interesting to see the differences in performance when applied to different models (Models A and B). Note that, for a given dataset, the RMSE and likelihood values in the tables are comparable across models. First we notice that with the i.i.d. noise model, Model A has a lower RMSE value than Model B for both the calibration and the two predictions. With Model A, none of the discrepancy models that we tried are able to outperform the simplest i.i.d. noise model when it comes to the mean predictions for the staircase and AP protocols; but GP($O, V$) has a very similar RMSE value as compare to the i.i.d. noise model while being able to capture some of the nonlinear dynamics that Model A misses as discussed above. However, with Model B, the GP($O, V$) model has the best RMSE value for the predictions, and the second best for the calibration where the best one is the GP($t$) model that overfits the data. The ARMA$(2, 2)$ model consistently gives the best likelihood value for Models A and B, as it gives a wider posterior distribution compared to other methods (Figure 7).

To conclude, in this example, we have used two different, incorrect model structures (Models A, B) to fit to a third model structure (Model C) generated synthetic data. We considered both ignoring discrepancy when calibration and incorporating discrepancy when calibration. Depending on the model, calibration with discrepancy could improve predictions notably as compared to calibration ignoring discrepancy (for the case of Model B), but not all (for Model A). Although our problem was a time-dependent (ODE) problem, constructing the discrepancy model as a pure time-series based function might not be useful in predicting unseen situations; the GP($O, V$) model performed the best compared with the other two time-series based models GP($t$) and ARMA$(2, 2)$.

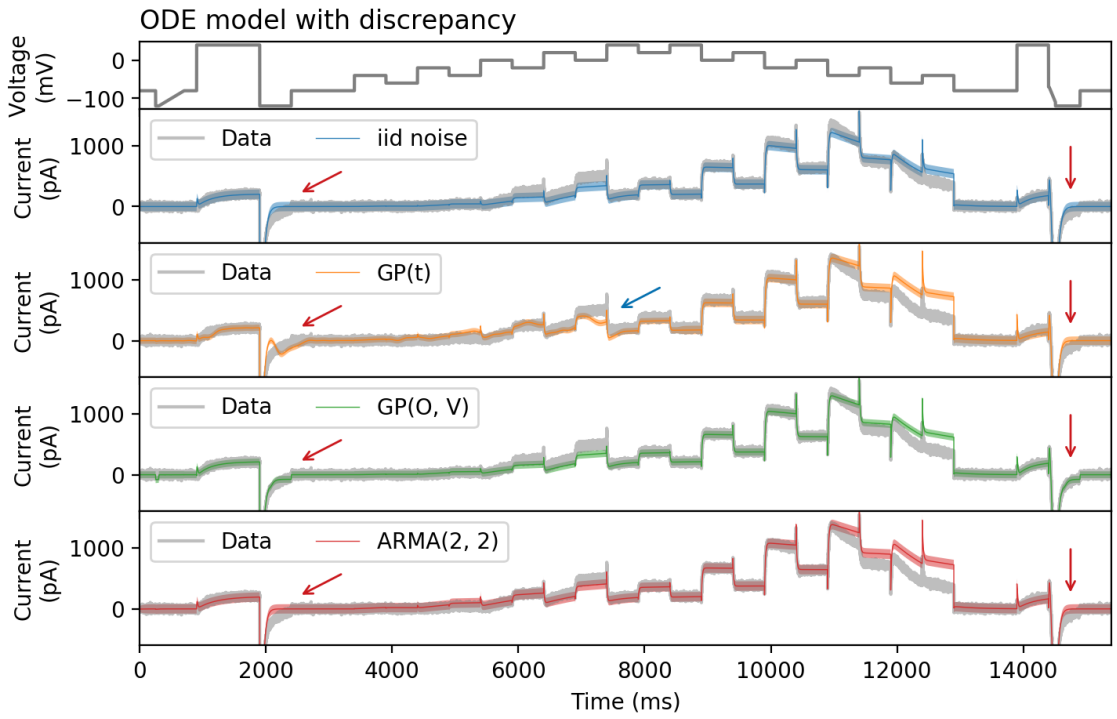


Figure 9: Models A prediction with different discrepancy models: i.i.d. noise, GP($t$), GP($O, V$), and ARMA(2, 2). The voltage clamp protocol for calibration is the staircase protocol [19]. We plot the posterior predictive with the mean (solid lines) and the bounds showing the 95% credible interval (shaded). The red arrows point to the tail current after the two activation steps which mark a visible area of model mismatch when calibrated without model discrepancy (i.i.d. noise, blue), and how the GP($O, V$) and ARMA(2, 2) models handle the mismatch differently. The blue arrow points to an obvious artefact at $\sim$ 7000 ms induced by the GP($t$) prediction which was trained on the sinusoidal protocol, and doesn't know anything about this staircase protocol.

# 4 Discussion

In this review and perspective piece we have drawn attention to an important and under-appreciated source of uncertainty in mechanistic models — that of uncertainty in the model structure or the equations themselves (model discrepancy). Focusing on cardiac electrophysiology models, we provided two examples of the consequences of ignoring discrepancy when calibrating models at the ion channel and action potential scales, highlighting how this could lead to wrongly-confident parameter posterior distributions and subsequently spurious predictions.

Statistically we can explicitly admit discrepancy exists, and include it in the model calibration process and predictions. We attempted to do this by modelling discrepancy using two proposals from the literature — Gaussian processes (GPs) trained on different inputs and an autoregressive-moving-average (ARMA) model. We saw how GPs could successfully describe discrepancy in the calibration experiment. A two-dimensional GP in voltage and time was used previously by Plumlee *et al.* [14, 15] where it could extrapolate to new voltages for a given single step voltage-clamp experiment. To be most useful in making new predictions for unseen situations, the discrepancy model needs to be a function of something other than time, otherwise features specific to the calibration experiment are projected into new situations. One promising example of such a discrepancy model was our two-dimensional GP as a function of the mechanistic model's open probability and voltage, although in the Model B case this led to ambiguity between the role of the ODE system and the role of the discrepancy (see Supplementary Section S7S7.2).

The modelling community would hope to study any discrepancy model that is shown to improve predictions, and use insights from this process to iteratively improve the mechanistic model. How we handle model discrepancy may depend on whether we are more interested in learning about what is missing in the model, or in making more reliable predictions: both related topics are worthy of more

investigation.

### 4.1 Recommendations

Very rarely do computational studies use more than one model to test the robustness of their predictions to the model form. We should bear in mind that all models are approximations and so when we are comparing to real data, all models have discrepancy. Here we have seen, using synthetic data that pretends we know a true data-generating mechanistic model, how fragile the calibration process can be for models with discrepancy and how this discrepancy manifests itself in predictions for new, unseen situations. Synthetic data studies, simulating data from different parameter sets and different model structures, allow the modeller to test how well the inverse problem can be solved and how robust predictions from the resulting models are. We strongly recommend performing such studies to learn more about your chosen model, and alternative models; as well as the effects of your model choice on parameter calibration and your subsequent predictions. To develop our field further, it will be important to document the model-fitting process, and to make datasets and infrastructure available to perform and reproduce these fits with different models [45].

### 4.2 Open questions and future work

The apparent similarity of the action potential models we looked at (summarised in Figure 1) is a challenge for model calibration. A number of papers have emphasised that more information can be gained to improve parameter identifiability with careful choice of experimental measurements, in particular by using membrane resistance [23, 27], or other protocols promoting more information-rich dynamics [24, 25] and some of these measurements may be more robust to discrepancy than others. In synthetic data, fitting the model used to generate the data will recover the same parameter set from any different protocol (where there is sufficient information to identify the parameters). But in the presence of discrepancy, fitting the same model to data from different protocols/experiments will result in different parameter sets, as the models make the best possible compromise (as shown schematically in Figure 4). This phenomenon may be an interesting way to approach and quantify model discrepancy.

If the difference between imperfect model predictions represented the difference between models and reality then this may also provide a way to estimate discrepancy. For instance, the largest difference between the ion channel Model A & B predictions in the staircase protocol was at the point in time that both of them showed largest discrepancy (Figure 6). Some form of Bayesian Model Averaging [46], using variance-between-models to represent discrepancy, may be instructive if the models are close enough to each other and reality, but can be misleading if the ensemble of models is not statistically exchangeable with the data generating process [47, 48] or if there is some systematic error (bias) due to experimental artefacts [49].

In time-structured problems, rather than adding a discrepancy to the final simulated trajectory, as we have done here, we can instead change the dynamics of the model directly. It may be easier to add a discrepancy term to the differential equations to address misspecification, than it is to correct their solution, but the downside is that this makes inference of the discrepancy computationally challenging. One such approach is to 'noise-up' the ODE by converting it to a stochastic differential equation [50, 51], i.e., replace $\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = f_\theta(\mathbf{x}, t)$ by $\mathrm{d}\mathbf{x} = f_\theta(\mathbf{x}, t)\mathrm{d}t + \Sigma^{\frac{1}{2}}\mathrm{d}W_t$ where $W_t$ is a Brownian motion with a variance matrix $\Sigma$. This turns the deterministic ODE into a stochastic model and can improve the UQ, but cannot capture any structure missing from the dynamics. We can go further and attempt to modify the underlying model equations, by changing the ODE system to

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = f_\theta(\mathbf{x}, t) + \delta(\mathbf{x}) \tag{10}$$

where again $\delta(\mathbf{x})$ is an empirical term to be learnt from the data. For example, this has been tried with a discretized version of the equations using a parametric model for $\delta$ [52], with GPs [53], nonlinear autoregressive exogenous (NARX) models [54], and deep neural networks [55]. Computation of posterior distributions for these models is generally challenging, but is being made easier by the development of automatic-differentiation software, which allows derivative information to be used in MCMC samplers, or in variational approaches to inference (e.g., [56, 57]).

Ultimately, modelling our way out of trouble, by expanding the model class, may prove impossible given the quantity of data available in many cases. Instead, we may want to modify our inferential approach to allow the best judgements possible about the parameters given the limitation of the model and data.

Approaches such as approximate Bayesian computation (ABC) [58] and history-matching [59, 60] change the focus from learning a statistical model within a Bayesian setting, to instead only requiring that the simulation gets within a certain distance of the data. This change, from a fully specified statistical model for $\delta$ to instead only giving an upper bound for $\delta$, is a conservative inferential approach where the aim is not to find the best parameter values, but instead rule out only obviously implausible values [61, 62].

For example, in the action potential model from Section 2, instead of taking a Bayesian approach with an i.i.d. Gaussian noise model, we can instead merely try to find parameter values that get us within some distance of the calibration data (see Supplement and Figure S2 for details). In the Supplement, we describe a simple approach, based on the methods presented in [63], where we find 1079 candidate parameter sets that give a reasonable match to the calibration data. When we use these parameters to predict the 2 Hz validation data, and the 75% $I_{\mathrm{Kr}}$ block CoU data, we get a wide range of predictions that incorporate the truth (Figure S3) — for a small subset of 70 out of 1079, we get good predictions and not the catastrophic prediction shown in Figure 2. By acknowledging the existence of model discrepancy, the use of wider error bounds (or higher-temperature likelihood functions) during the fitting process may avoid fitting parameters overly-precisely. However, we have no way of knowing which subset of remaining parameter space is more plausible (if any) without doing these further experiments; testing the model as close as possible to the desired context of use helps us spot such spurious behaviour.

# 5 Conclusions

In this paper we have seen how having an imperfect representation of a system in a mathematical model (discrepancy) can lead to spuriously certain parameter inference and overly-confident and wrong predictions. We have examined a range of methods that attempt to account for discrepancy in the fitting process using synthetic data studies. In some cases we can improve predictions using these methods, but different methods work better for different models in different situations, and in some cases the best predictions were still made by ignoring discrepancy. A large benefit of the calibration methods which include discrepancy is that they better represent uncertainty in predictions, although all the methods we trialled still failed to allow for a wide enough range of possible outputs in certain parts of the protocols. Methodological developments are needed to design reliable methods to deal with model discrepancy for use in safety-critical electrophysiology predictions.

## Data access

Code to reproduce the results in the tutorials is available at https://github.com/CardiacModelling/fickleheart-method-tutorials.

## Author contributions

CLL and SG wrote the code to perform the examples in the main paper. CLL, SG, DGW, GRM and RDW drafted the manuscript. All authors conceived and designed the study. All authors read and approved the manuscript.

## Competing interests

The authors declare that they have no competing interests.


## Funding

This work was supported by the Wellcome Trust [grant numbers 101222/Z/13/Z and 212203/Z/18/Z]; the Engineering & Physical Sciences Research Council [grant numbers EP/R014604/1, EP/P010741/1, EP/L016044/1, EP/R006768/1, EP/S014985/1, and EP/R003645/1]; the British Heart Foundation [grant numbers PG/15/59/31621, RE/13/4/30184, and SP/18/6/33805]; the Russian Foundation for Basic Research [grant number 18-29-13008]. CLL acknowledges support from the Clarendon Scholarship Fund; and the EPSRC, MRC and F. Hoffmann-La Roche Ltd. for studentship support. CDC and CH were supported by the BHF. MR acknowledges a BHF Turing Cardiovascular Data Science Award. AVP was partially supported by RF Government Act No. 211 of March 16, 2013, and RFBR. RWS was supported by the Brazilian Government via CAPES, CNPq, FAPEMIG, and UFJF, and by an Endeavour Research Leadership Award from the Australian Government Department of Education. KW would like

to acknowledge the support of the UK EPSRC. GMN was supported by CEFET-MG and CAPES. GRM & SG acknowledge support from the Wellcome Trust & Royal Society via a Sir Henry Dale Fellowship to GRM. GRM & DGW acknowledge support from the Wellcome Trust via a Wellcome Trust Senior Research Fellowship to GRM. The authors would like to thank the Isaac Newton Institute for Mathematical Sciences for support and hospitality during the 'Fickle Heart' programme.

## Acknowledgements

We would like to thank all the participants at the Isaac Newton Institute 'Fickle Heart' programme for helpful discussions which informed this manuscript.

# References

[1] D. Noble, A. Garny, and P. Noble, "How the Hodgkin-Huxley equations inspired the Cardiac Physiome Project," *The Journal of Physiology*, vol. 590, pp. 2613–2628, May 2012.

[2] J. Relan, P. Chinchapatnam, and M. Sermesant, "Coupled personalization of cardiac electrophysiology models for prediction of ischaemic ventricular tachycardia," *Interface Focus*, vol. 1, pp. 396–407, 2011.

[3] M. Sermesant, R. Chabiniok, P. Chinchapatnam, T. Mansi, F. Billet, P. Moireau, J. Peyrat, K. Wong, J. Relan, K. Rhode, M. Ginks, P. Lambiase, H. Delingette, M. Sorine, C. Rinaldi, D. Chapelle, R. Razavi, and N. Ayache, "Patient-specific electromechanical models of the heart for the prediction of pacing acute effects in CRT: A preliminary clinical validation," *Med Image Anal*, vol. 16, pp. 201–215, 2012.

[4] G. R. Mirams, M. R. Davies, Y. Cui, P. Kohl, and D. Noble, "Application of cardiac electrophysiology simulations to pro-arrhythmic safety testing," *British Journal of Pharmacology*, vol. 167, no. 5, pp. 932–945, 2012.

[5] S. A. Niederer, J. Lumens, , and N. A. Trayanova, "Computational models in cardiology," *Nat. Rev. Cardiol.*, vol. 16, pp. 100–111, 2018.

[6] Z. Li, B. J. Ridder, X. Han, W. W. Wu, J. Sheng, P. N. Tran, M. Wu, A. Randolph, R. H. Johnstone, G. R. Mirams, *et al.*, "Assessment of an in silico mechanistic model for proarrhythmia risk prediction under the ci pa initiative," *Clinical Pharmacology & Therapeutics*, vol. 105, no. 2, pp. 466–475, 2019.

[7] G. R. Mirams, P. Pathmanathan, R. A. Gray, P. Challenor, and R. H. Clayton, "Uncertainty and variability in computational and mathematical models of cardiac physiology," *J. Physiol.*, vol. 594, pp. 6833–6847, 2016.

[8] S. Niederer, E. Kerfoot, A. Benson, M. Bernabeu, O. Bernus, C. Bradley, E. Cherry, R. Clayton, F. Fenton, A. Garny, E. Heidenreich, S. Land, M. Maleckar, P. Pathmanathan, G. Plank, J. Rodríguez, I. Roy, F. Sachse, G. Seemann, O. Skavhaug, and N. Smith, "Verification of cardiac tissue electrophysiology simulators using an N-version benchmark," *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 369, pp. 4331–4351, 2011.

[9] S. Krishnamoorthi, L. E. Perotti, N. P. Borgstrom, O. A. Ajijola, A. Frid, A. V. Ponnaluri, J. N. Weiss, Z. Qu, W. S. Klug, D. B. Ennis, and A. Garfinkel, "Simulation methods and validation criteria for modeling cardiac ventricular electrophysiology," *PLoS ONE*, vol. 9, p. e114494, 2014.

[10] P. Pathmanathan and R. Gray, "Ensuring reliability of safety-critical clinical applications of computational cardiac models," *Front. Physiol.*, vol. 4, p. 358, 2013.

[11] P. Pathmanathan and R. Gray, "Verification of computational models of cardiac electro-physiology," *Int. J. Num. Methods Biomed. Eng.*, vol. 30, pp. 525–544, 2014.

[12] P. Pathmanathan and R. A. Gray, "Validation and trustworthiness of multiscale models of cardiac electrophysiology," *Front. Physiol.*, vol. 9, p. 106, 2018.

[13] P. Pathmanathan, J. M. Cordeiro, and R. A. Gray, "Comprehensive uncertainty quantification and sensitivity analysis for cardiac action potential models," *Front. Physiol.*, vol. 10, p. 721, 2019.

[14] M. Plumlee, V. R. Joseph, H. Yang, V. Roshan Joseph, and H. Yang, "Calibrating Functional Parameters in the Ion Channel Models of Cardiac Cells," *Journal of the American Statistical Association*, vol. 111, pp. 500–509, Apr. 2016.

[15] M. Plumlee, "Bayesian Calibration of Inexact Computer Models," *Journal of the American Statistical Association*, vol. 112, pp. 1274–1285, July 2017.

[16] S. A. Niederer, M. Fink, D. Noble, and N. P. Smith, "A meta-analysis of cardiac electrophysiology computational models," *Experimental Physiology*, vol. 94, p. 486, May 2009.

[17] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis, third edition*. Chapman and Hall, 2013.

[18] B. Lambert, *A student's guide to Bayesian statistics*. Sage, 2018.

[19] C. L. Lei, M. Clerx, D. J. Gavaghan, L. Polonchuk, G. R. Mirams, and K. Wang, "Rapid characterisation of hERG channel kinetics I: using an automated high-throughput system," *Biophysical Journal*, vol. 117, pp. 2438–2454, 2019.

[20] J. Brynjarsdóttir and A. O'Hagan, "Learning about physical parameters: the importance of model discrepancy," *Inverse Problems*, vol. 30, no. 11, p. 114007, 2014.

[21] K. H. Ten Tusscher, D. Noble, P.-J. Noble, and A. V. Panfilov, "A model for human ventricular tissue," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 286, no. 4, pp. H1573–H1589, 2004.

[22] M. Fink, D. Noble, L. Virag, A. Varro, and W. R. Giles, "Contributions of HERG K+ current to repolarization of the human ventricular action potential," *Progress in biophysics and molecular biology*, vol. 96, no. 1-3, pp. 357–376, 2008.

[23] J. Kaur, A. Nygren, and E. J. Vigmond, "Fitting membrane resistance along with action potential shape in cardiac myocytes improves convergence: application of a multi-objective parallel genetic algorithm.," *PloS one*, vol. 9, p. e107984, Jan. 2014.

[24] W. Groenendaal, F. A. Ortega, A. R. Kherlopian, A. C. Zygmunt, T. Krogh-Madsen, and D. J. Christini, "Cell-Specific Cardiac Electrophysiology Models.," *PLoS computational biology*, vol. 11, p. e1004242, Apr. 2015.

[25] R. H. Johnstone, E. E. Chang, R. Bardenet, T. P. de Boer, D. J. Gavaghan, P. Pathmanathan, R. H. Clayton, and G. R. Mirams, "Uncertainty and variability in models of the cardiac action potential: Can we build trustworthy models?," *Journal of Molecular and Cellular Cardiology*, vol. 96, pp. 49–62, July 2016.

[26] C. L. Lei, K. Wang, M. Clerx, R. H. Johnstone, M. P. Hortigon-Vinagre, V. Zamora, A. Allan, G. L. Smith, D. J. Gavaghan, G. R. Mirams, and L. Polonchuk, "Tailoring mathematical models to stem-cell derived cardiomyocyte lines can improve predictions of drug-induced changes to their electrophysiology," *Frontiers in Physiology*, vol. 8, 2017.

[27] E. Pouranbarani, R. Weber dos Santos, and A. Nygren, "A robust multi-objective optimization framework to capture both cellular and intercellular properties in cardiac cellular model tuning: Analyzing different regions of membrane resistance profile in parameter fitting," *PLOS ONE*, vol. 14, pp. 1–19, Nov. 2019.

[28] N. Hansen, *The CMA Evolution Strategy: A Comparing Review*, pp. 75–102. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.

[29] M. Clerx, M. Robinson, B. Lambert, C. L. Lei, S. Ghosh, G. R. Mirams, and D. J. Gavaghan, "Probabilistic inference on noisy time series (PINTS)," *Journal of Open Research Software*, vol. 7, no. 1, p. 23, 2019.

[30] M. Clerx, P. Collins, E. de Lange, and P. G. A. Volders, "Myokit: A simple interface to cardiac cellular electrophysiology," *Progress in Biophysics and Molecular Biology*, vol. 120, pp. 100–114, Jan. 2016.

[31] A. W. Van der Vaart, *Asymptotic statistics*, vol. 3. Cambridge university press, 2000.

[32] J. M. Bernardo and A. F. Smith, *Bayesian theory*, vol. 405. John Wiley & Sons, 2009.

[33] B. J. Kleijn, A. W. van der Vaart, *et al.*, "Misspecification in infinite-dimensional bayesian statistics," *The Annals of Statistics*, vol. 34, no. 2, pp. 837–877, 2006.

[34] P. De Blasi, S. G. Walker, *et al.*, "Bayesian asymptotics with misspecified models," *Statistica Sinica*, vol. 23, pp. 169–187, 2013.

[35] M. C. Kennedy and A. O'Hagan, "Bayesian calibration of computer models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 3, pp. 425–464, 2001.

[36] K. A. Beattie, A. P. Hill, R. Bardenet, Y. Cui, J. I. Vandenberg, D. J. Gavaghan, T. P. De Boer, and G. R. Mirams, "Sinusoidal voltage protocols for rapid characterisation of ion channel kinetics," *The Journal of physiology*, vol. 596, no. 10, pp. 1813–1828, 2018.

[37] C. S. Oehmen, W. R. Giles, and S. S. Demir, "Mathematical model of the rapidly activating delayed rectifier potassium current ikr in rabbit sinoatrial node," *Journal of cardiovascular electrophysiology*, vol. 13, no. 11, pp. 1131–1140, 2002.

[38] G. Y. Di Veroli, M. R. Davies, H. Zhang, N. Abi-Gerges, and M. R. Boyett, "High-throughput screening of drug-binding dynamics to herg improves early drug safety assessment," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 304, no. 1, pp. H104–H117, 2012.

[39] C. L. Lei, M. Clerx, K. A. Beattie, D. Melgari, J. C. Hancox, D. J. Gavaghan, L. Polonchuk, K. Wang, and G. R. Mirams, "Rapid characterisation of hERG channel kinetics II: temperature dependence," *Biophysical Journal*, vol. 117, pp. 2455–2470, 2019.

[40] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck, "Probabilistic programming in python using pymc3," *PeerJ Computer Science*, vol. 2, p. e55, 2016.

[41] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv e-prints*, vol. abs/1605.02688, May 2016.

[42] J. Durbin and S. J. Koopman, *Time series analysis by state space methods*. Oxford university press, 2012.

[43] S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with python," in *Proceedings of the 9th Python in Science Conference*, vol. 57, p. 61, Scipy, 2010.

[44] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007.

[45] A. C. Daly, M. Clerx, K. A. Beattie, J. Cooper, D. J. Gavaghan, and G. R. Mirams, "Reproducible model development in the cardiac electrophysiology Web Lab," *Progress in Biophysics and Molecular Biology*, vol. 139, pp. 3–14, Nov. 2018.

[46] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, "Bayesian Model Averaging: A Tutorial," *Statistical Science*, vol. 14, no. 4, pp. 382–401, 1999.

[47] R. E. Chandler, "Exploiting strength, discounting weakness: combining information from multiple climate simulators," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1991, p. 20120388, 2013.

[48] J. Rougier, M. Goldstein, and L. House, "Second-order exchangeability analysis for multimodel ensembles," *Journal of the American Statistical Association*, vol. 108, no. 503, pp. 852–863, 2013.

[49] C. L. Lei, M. Clerx, D. G. Whittaker, D. J. Gavaghan, T. P. de Boer, and G. R. Mirams, "Accounting for variability in ion current recordings using a mathematical model of artefacts in voltage-clamp experiments," *bioRxiv*, 2019.

[50] M. Crucifix and J. Rougier, "On the use of simple dynamical systems for climate predictions," *The European Physical Journal Special Topics*, vol. 174, no. 1, pp. 11–31, 2009.

[51] J. Carson, M. Crucifix, S. Preston, and R. D. Wilkinson, "Bayesian model selection for the glacial–interglacial cycle," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 67, no. 1, pp. 25–54, 2018.

[52] R. D. Wilkinson, M. Vrettas, D. Cornford, and J. E. Oakley, "Quantifying Simulator Discrepancy in Discrete-Time Dynamical Simulators," *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 16, pp. 554–570, Dec. 2011.

[53] R. Frigola, F. Lindsten, T. B. Schön, and C. E. Rasmussen, "Bayesian inference and learning in Gaussian process state-space models with particle MCMC," in *Advances in Neural Information Processing Systems*, pp. 3156–3164, 2013.

[54] K. Worden, W. Becker, T. Rogers, and E. Cross, "On the confidence bounds of gaussian process NARX models and their higher-order frequency response functions," *Mechanical Systems and Signal Processing*, vol. 104, pp. 188–223, 2018.

[55] T. Meeds, G. Roeder, P. Grant, A. Phillips, and N. Dalchau, "Efficient amortised bayesian inference for hierarchical and nonlinear dynamical systems," in *International Conference on Machine Learning*, pp. 4445–4455, 2019.

[56] R. M. Neal *et al.*, "MCMC using Hamiltonian dynamics," *Handbook of markov chain monte carlo*, vol. 2, no. 11, p. 2, 2011.

[57] T. Ryder, A. Golightly, A. S. McGough, and D. Prangle, "Black-box variational inference for stochastic differential equations," *arXiv preprint arXiv:1802.03335*, 2018.

[58] S. A. Sisson, Y. Fan, and M. Beaumont, *Handbook of Approximate Bayesian Computation*. Chapman and Hall/CRC Press, 2018.

[59] P. S. Craig, M. Goldstein, A. H. Seheult, and J. A. Smith, "Pressure matching for hydrocarbon reservoirs: a case study in the use of bayes linear strategies for large computer experiments," in *Case studies in Bayesian statistics*, pp. 37–93, Springer, 1997.

[60] P. Craig, M. Goldstein, J. Rougier, and A. Seheult, "Bayesian forecasting using large computer models," *J. Amer. Statist. Assoc*, vol. 96, pp. 717–729, 2001.

[61] R. D. Wilkinson, "Approximate bayesian computation (ABC) gives exact results under the assumption of model error," *Statistical applications in genetics and molecular biology*, vol. 12, no. 2, pp. 129–141, 2013.

[62] P. B. Holden, N. R. Edwards, J. Hensman, and R. D. Wilkinson, "ABC for climate: dealing with expensive simulators," *Handbook of Approximate Bayesian Computation*, pp. 569–95, 2018.

[63] G. M. Novaes, J. O. Campos, E. Alvarez-Lacalle, S. A. Muñoz, B. M. Rocha, and R. W. dos Santos, "Combining polynomial chaos expansions and genetic algorithm for the coupling of electrophysiological models," in *Computational Science – ICCS 2019, Lecture Notes in Computer Science*, vol. 11538, pp. 116–129, Springer, 2019.

# Considering discrepancy when calibrating a mechanistic electrophysiology model: Supplementary Material

Chon Lok Lei[1], Sanmitra Ghosh[2], Dominic G. Whittaker[3], Yasser Aboelkassem[4], Kylie A. Beattie[5], Chris D. Cantwell[6], Tammo Delhaas[7], Charles Houston[6], Gustavo Montes Novaes[8], Alexander V. Panfilov[9,10], Pras Pathmanathan[11], Marina Riabiz[12], Rodrigo Weber dos Santos[8], Keith Worden[13], Gary R. Mirams[3] and Richard D. Wilkinson[14]

[1] Computational Biology & Health Informatics, Dept. of Computer Science, University of Oxford, UK.
[2] MRC Biostatistics Unit, University of Cambridge, UK
[3] Centre for Mathematical Medicine & Biology, School of Mathematical Sciences, University of Nottingham, UK.
[4] Department of Bioengineering, University of California San Diego, USA.
[5] Systems Modeling and Translational Biology, GlaxoSmithKline R&D, Stevenage, UK.
[6] ElectroCardioMaths Programme, Centre for Cardiac Engineering, Imperial College London, UK.
[7] CARIM School for Cardiovascular Diseases, Maastricht University, the Netherlands.
[8] Graduate Program in Computational Modeling, Universidade Federal de Juiz de Fora, Brazil.
[9] Department of Physics and Astronomy, Ghent University, Belgium.
[10] Laboratory of Computational Biology and Medicine, Ural Federal University, Ekaterinburg, Russia.
[11] U.S. Food and Drug Administration, Center for Devices and Radiological Health, Office of Science and Engineering Laboratories, USA.
[12] Department of Biomedical Engineering King's College London and Alan Turing Institute, UK.
[13] Dynamics Research Group, Department of Mechanical Engineering, University of Sheffield, UK.
[14] School of Mathematics and Statistics, University of Sheffield, UK.

## Contents

**S1 Differences between Model T and Model F** 1

**S2 Modelling discrepancy using a Gaussian process** 6

**S3 Modelling residuals using an ARMA$(p, q)$ process** 9

**S4 Choice of priors for the ion channel example** 11

**S5 Computing and representing posterior predictive** 12

**S6 Supplementary results for the action potential example** 13

**S7 Supplementary results for the ion channel discrepancy example** 16
S7.1 Model A . . . 17
S7.1.1 Model A: Full model predictions . . . 17
S7.1.2 Model A: Discrepancy predictions . . . 18
S7.1.3 Model A: ODE model predictions . . . 21
S7.2 Model B . . . 24
S7.2.1 Model B: Full model predictions . . . 25
S7.2.2 Model B: Discrepancy predictions . . . 28
S7.2.3 Model B: ODE model predictions . . . 31

## S1 Differences between Model T and Model F

Here we show the equations of the currents in Model T [1] and Model F [2] that are different in kinetics. Below $g_X$ is the conductance (constant), $E_X$ is the reversal potential, $X_o$ is the extracellular concentration, $X_i$ is the intracellular concentration of ion $X$. $V$ is the membrane voltage, $F$ is Faraday constant, $R$ is the ideal gas constant.

## $I_{\mathbf{CaL}}$

### Model T

$$I_{CaL} = \frac{g_{CaL} \cdot d \cdot f \cdot fCa \cdot 4 \cdot V \cdot F^2}{R \cdot 310} \times \frac{\left(Ca_i \cdot e^{\frac{2 \cdot V \cdot F}{R \cdot 310}} - 0.341 \cdot Ca_o\right)}{\left(e^{\frac{2 \cdot V \cdot F}{R \cdot 310}} - 1\right)}$$

$$d_\infty = \frac{1}{\left(1 + e^{\frac{(-5-V)}{7.5}}\right)}$$

$$\alpha_d = \left(\frac{1.4}{\left(1 + e^{\frac{(-35-V)}{13}}\right)} + 0.25\right)$$

$$\beta_d = \frac{1.4}{\left(1 + e^{\frac{(V+5)}{5}}\right)}$$

$$\gamma_d = \frac{1}{\left(1 + e^{\frac{(50-V)}{20}}\right)}$$

$$\tau_d = \left(\alpha_d \cdot \beta_d + \gamma_d\right)$$

$$\frac{\mathrm{d}d}{\mathrm{d}t} = \frac{(d_\infty - d)}{\tau_d}$$

$$f_\infty = \frac{1}{\left(1 + e^{\frac{(V+20)}{7}}\right)}$$

$$\tau_f = \left(1125 \cdot e^{\frac{-(V+27)^2}{240}} + 80 + \frac{165}{\left(1 + e^{\frac{(25-V)}{10}}\right)}\right)$$

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \frac{(f_\infty - f)}{\tau_f}$$

$$\alpha_{fCa} = \frac{1}{\left(1 + \left(\frac{Ca_i}{0.000325}\right)^8\right)}$$

$$\beta_{fCa} = \frac{0.1}{\left(1 + e^{\frac{(Ca_i - 0.0005)}{0.0001}}\right)}$$

$$\gamma_{fCa} = \frac{0.2}{\left(1 + e^{\frac{(Ca_i - 0.00075)}{0.0008}}\right)}$$

$$fCa_\infty = \frac{\left(\alpha_{fCa} + \beta_{fCa} + \gamma_{fCa} + 0.23\right)}{1.46}$$

$$\tau_{fCa} = 2$$

$$d_{fCa} = \frac{(fCa_\infty - fCa)}{\tau_{fCa}}$$

$$\frac{\mathrm{d}fCa}{\mathrm{d}t} = \begin{cases} 0; & \text{if } (fCa_\infty > fCa) \text{ and } (V > -60), \\ d_{fCa} & \text{otherwise.} \end{cases}$$

### Model F

$$I_{CaL} = \frac{g_{CaL} \cdot d \cdot f \cdot f_2 \cdot fCass \cdot 4 \cdot (V - 15) \cdot F^2}{R \cdot 310} \times \frac{\left(0.25 \cdot Ca_{ss} \cdot e^{\frac{2 \cdot (V-15) \cdot F}{R \cdot 310}} - Ca_o\right)}{\left(e^{\frac{2 \cdot (V-15) \cdot F}{R \cdot 310}} - 1\right)}$$

$$d_\infty = \frac{1}{\left(1 + e^{\frac{(5-V)}{7.5}}\right)}$$

$\alpha_d =$ as per Model T,

$\beta_d =$ as per Model T,

$\gamma_d =$ as per Model T,

$\tau_d =$ as per Model T,

$\frac{\mathrm{d}d}{\mathrm{d}t} =$ as per Model T,

$f_\infty =$ as per Model T,

$$\tau_f = \frac{1}{4}\left(1102.5 \cdot e^{\frac{-(((V+27))^2)}{225}} + \frac{200}{\left(1 + e^{\frac{(13-V)}{10}}\right)} + \frac{180}{\left(1 + e^{\frac{(V+30)}{10}}\right)} + 20\right)$$

$\frac{\mathrm{d}f}{\mathrm{d}t} =$ as per Model T,

$$f2_\infty = \left(\frac{0.75}{\left(1 + e^{\frac{(V+35)}{7}}\right)} + 0.25\right)$$

$$\tau_{f2} = \frac{1}{2}\left(562 \cdot e^{\frac{-(V+27)^2}{240}} + \frac{31}{\left(1 + e^{\frac{(25-V)}{10}}\right)} + \frac{80}{\left(1 + e^{\frac{(V+30)}{10}}\right)}\right)$$

$$\frac{\mathrm{d}f_2}{\mathrm{d}t} = \frac{(f2_\infty - f_2)}{\tau_{f2}}$$

$$fCass_\infty = \left(\frac{0.4}{\left(1 + \left(\frac{Ca_{ss}}{0.05}\right)^2\right)} + 0.6\right)$$

$$\tau_{fCass} = \left(\frac{80}{\left(1 + \left(\frac{Ca_{ss}}{0.05}\right)^2\right)} + 2\right)$$

$$\frac{\mathrm{d}fCass}{\mathrm{d}t} = \frac{(fCass_\infty - fCass)}{\tau_{fCass}}$$

$I_{\mathbf{to}}$

**Model T**

$$I_{to} = g_{to} \cdot r \cdot s \cdot (V - E_K)$$

$$s_\infty = \frac{1}{\left(1 + e^{\frac{(V+28)}{5}}\right)}$$

$$\tau_s = 1000 \cdot e^{\left(\frac{-(V+67)^2}{1000}\right)} + 8$$

$$\frac{\mathrm{d}s}{\mathrm{d}t} = \frac{(s_\infty - s)}{\tau_s}$$

$$r_\infty = \frac{1}{\left(1 + e^{\frac{(20-V)}{6}}\right)}$$

$$\tau_r = \left(9.5 \cdot e^{\frac{-(V+40)^2}{1800}} + 0.8\right)$$

$$\frac{\mathrm{d}r}{\mathrm{d}t} = \frac{(r_\infty - r)}{\tau_r}$$

**Model F**

$I_{to}$ = as per Model T,

$$s_\infty = \frac{1}{\left(1 + e^{\frac{(V+20)}{5}}\right)}$$

$$\tau_s = \left(85 \cdot e^{\frac{-(V+45)^2}{320}} + \frac{5}{\left(1 + e^{\frac{(V-20)}{5}}\right)} + 3\right)$$

$$\frac{\mathrm{d}s}{\mathrm{d}t} = \text{as per Model T},$$

$r$ gate as per Model T.

## $I_{\mathbf{Kr}}$

### Model T

$$I_{Kr} = g_{Kr} \cdot \sqrt{\frac{K_o}{5.4}} \cdot X_{r1} \cdot X_{r2} \cdot (V - E_K)$$

$$xr1_\infty = \frac{1}{\left(1 + e^{\frac{(-26-V)}{7}}\right)}$$

$$\alpha_{xr1} = \frac{450}{\left(1 + e^{\frac{(-45-V)}{10}}\right)}$$

$$\beta_{xr1} = \frac{6}{\left(1 + e^{\frac{(V+30)}{11.5}}\right)}$$

$$\tau_{xr1} = \alpha_{xr1} \cdot \beta_{xr1}$$

$$\frac{\mathrm{d}X_{r1}}{\mathrm{d}t} = \frac{(xr1_\infty - X_{r1})}{\tau_{xr1}}$$

$$xr2_\infty = \frac{1}{\left(1 + e^{\frac{(V+88)}{24}}\right)}$$

$$\alpha_{xr2} = \frac{3}{\left(1 + e^{\frac{(-60-V)}{20}}\right)}$$

$$\beta_{xr2} = \frac{1.12}{\left(1 + e^{\frac{(V-60)}{20}}\right)}$$

$$\tau_{xr2} = \alpha_{xr2} \cdot \beta_{xr2}$$

$$\frac{\mathrm{d}X_{r2}}{\mathrm{d}t} = \frac{(xr2_\infty - X_{r2})}{\tau_{xr2}}$$

### Model F

$$I_{Kr} = g_{Kr} \cdot \left(\frac{310}{35} - \frac{55}{7}\right) \cdot \sqrt{\frac{K_o}{5.4}} \cdot Or4 \cdot (V - E_K)$$

$$\alpha_{xr1} = e^{(24.335+(0.0112 \cdot V - 25.914))}$$

$$\beta_{xr1} = e^{(13.688+(-(0.0603) \cdot V - 15.707))}$$

$$\alpha_{xr2} = e^{(22.746+(0 \cdot V - 25.914))}$$

$$\beta_{xr2} = e^{(13.193+(0 \cdot V - 15.707))}$$

$$\alpha_{xr3} = e^{(22.098+(0.0365 \cdot V - 25.914))}$$

$$\beta_{xr3} = e^{(7.313+(-(0.0399) \cdot V - 15.707))}$$

$$\alpha_{xr4} = e^{(30.016+(0.0223 \cdot V - 30.888))} \cdot \left(\frac{5.4}{K_o}\right)^{0.4}$$

$$\beta_{xr4} = e^{(30.061+(-(0.0312) \cdot V - 33.243))}$$

$$\frac{\mathrm{d}Cr1}{\mathrm{d}t} = (\beta_{xr1} \cdot Cr2 - \alpha_{xr1} \cdot Cr1)$$

$$\frac{\mathrm{d}Cr2}{\mathrm{d}t} = ((\alpha_{xr1} \cdot Cr1 + \beta_{xr2} \cdot Cr3) - (\alpha_{xr2} + \beta_{xr1}) \cdot Cr2)$$

$$\frac{\mathrm{d}Cr3}{\mathrm{d}t} = ((\alpha_{xr2} \cdot Cr2 + \beta_{xr3} \cdot Or4) - (\alpha_{xr3} + \beta_{xr2}) \cdot Cr3)$$

$$\frac{\mathrm{d}Or4}{\mathrm{d}t} = ((\alpha_{xr3} \cdot Cr3 + \beta_{xr4} \cdot Ir5) - (\alpha_{xr4} + \beta_{xr3}) \cdot Or4)$$

$$\frac{\mathrm{d}Ir5}{\mathrm{d}t} = (\alpha_{xr4} \cdot Or4 - \beta_{xr4} \cdot Ir5)$$

## $I_{\mathbf{K1}}$

### Model T

$$I_{K1} = g_{K1} \cdot xK1_\infty \cdot \sqrt{\frac{K_o}{5.4}} \cdot (V - E_K)$$
$$\alpha_{K1} = \frac{0.1}{\left(1 + e^{0.06 \cdot ((V-E_K)-200)}\right)}$$
$$\beta^a_{K1} = 3 \cdot e^{0.0002 \cdot ((V-E_K)+100)}$$
$$\beta^b_{K1} = e^{0.1 \cdot ((V-E_K)-10)}$$
$$\beta_{K1} = \frac{\beta^a_{K1} + \beta^b_{K1}}{\left(1 + e^{-(0.5) \cdot (V-E_K)}\right)}$$
$$xK1_\infty = \frac{\alpha_{K1}}{(\alpha_{K1} + \beta_{K1})}$$

### Model F

$$I_{K1} = g_{K1} \cdot \left(\frac{310}{35} - \frac{55}{7}\right) \cdot \sqrt{\frac{K_o}{5.4}} \cdot xK1_\infty \cdot (V - E_K)$$
$$Ki_{Mg} = 2.8 \cdot e^{\frac{-(V - \delta \cdot E_K)}{180}}$$
$$Kb_{Mg} = 0.45 \cdot e^{\frac{-(V - \delta \cdot E_K)}{20}}$$
$$Kd1_{SPM} = 0.7 - 3 \cdot e^{\frac{-\left((V - \delta \cdot E_K) + 8 \cdot Mg_{Buf}\right)}{4.8}}$$
$$Kd2_{SPM} = 40 - 3 \cdot e^{\frac{-(V - \delta \cdot E_K)}{9.1}}$$
$$X = \left(1 + \frac{Mg_{Buf}}{Kb_{Mg}}\right)$$
$$rec1 = \frac{X^2}{\left(\frac{SPM}{Kd1_{SPM}} + \frac{Mg_{Buf}}{Ki_{Mg}} + X^3\right)}$$
$$rec2 = \frac{1}{\left(1 + \frac{SPM}{Kd2_{SPM}}\right)}$$
$$xK1_\infty = (\phi \cdot rec1 + (1 - \phi) \cdot rec2)$$
$$Mg_{Buf} = 0.0356$$
$$SPM = 0.0014613$$
$$\phi = 0.8838$$
$$\delta = 1.0648$$

## $I_{\mathbf{Ks}}$

### Model T

$$I_{Ks} = g_{Ks} \cdot X_s^2 \cdot (V - E_{Ks})$$
$$xs_\infty = \frac{1}{\left(1 + e^{\frac{(-5-V)}{14}}\right)}$$
$$\alpha_{xs} = \frac{1100}{\sqrt{\left(1 + e^{\frac{(-10-V)}{6}}\right)}}$$
$$\beta_{xs} = \frac{1}{\left(1 + e^{\frac{(V-60)}{20}}\right)}$$
$$\tau_{xs} = \alpha_{xs} \cdot \beta_{xs}$$
$$\frac{\mathrm{d}X_s}{\mathrm{d}t} = \frac{(xs_\infty - X_s)}{\tau_{xs}}$$

### Model F

$$I_{Ks} = \text{as per Model T},$$
$$xs_\infty = \text{as per Model T},$$
$$\alpha_{xs} = \frac{1400}{\sqrt{\left(1 + e^{\frac{(5-V)}{6}}\right)}}$$
$$\beta_{xs} = \frac{1}{\left(1 + e^{\frac{(V-35)}{15}}\right)}$$
$$\tau_{xs} = (\alpha_{xs} \cdot \beta_{xs} + 80)$$
$$\frac{\mathrm{d}X_s}{\mathrm{d}t} = \text{as per Model T}.$$

## S2 Modelling discrepancy using a Gaussian process

In order to add a discrepancy term to our basic measurement model (see main text), we model the $i^{\text{th}}$ observation as:

$$(Y_C)_i = f_i(\boldsymbol{\theta}, u_C^i) + \delta_i(\boldsymbol{\phi}, \boldsymbol{v}_C^i) + \epsilon_i, \tag{S2.1}$$

where $\delta_i(\boldsymbol{\phi}, \boldsymbol{v}_C^i)$ is the model discrepancy term, a function with arguments $\boldsymbol{v}_C$ and parameters $\boldsymbol{\phi}$. Note that the inputs $\boldsymbol{v}_C$ can be independent from the inputs passed to the mechanistic model. We choose $v_C$ to be (1) time $t$, and (2) the open probability $O$ (i.e. $\mathcal{O}$ in Eq. (6)) and the voltage $V$.

Following [3] we place a zero mean Gaussian process prior on the discrepancy function given by

$$\delta(\boldsymbol{\phi}, \boldsymbol{v}_C) \sim \mathcal{GP}\big(0, \boldsymbol{\kappa}(\boldsymbol{v}_C, \boldsymbol{v}'_C; \boldsymbol{\phi})\big), \tag{S2.2}$$

where $\boldsymbol{\kappa}(\boldsymbol{v}_C, \boldsymbol{v}'_C; \boldsymbol{\phi})$ is the covariance function (also known as covariance *kernel*) parameterised by $\boldsymbol{\phi}$. One common choice for the covariance function is the squared exponential function given by

$$\boldsymbol{\kappa}(\boldsymbol{v}_C, \boldsymbol{v}'_C; \boldsymbol{\phi}) = \alpha^2 \exp\left(-\sum_{j=1}^{q} \frac{(v_{C_j} - v'_{C_j})^2}{2\rho_j^2}\right), \tag{S2.3}$$

where $q$ is the number of covariates, such as time or open probability as mentioned above, representing $\boldsymbol{v}_C$. The parameter $\rho_j$ quantifies the characteristic length-scale along the $j^{\text{th}}$ covariate and $\alpha$ denotes the marginal variance of the GP prior. Together they constitute the parameter vector $\boldsymbol{\phi} = [\alpha, \rho_1, \ldots, \rho_q]$.

Since our measurement noise is Gaussian with variance $\sigma^2$ we can analytically compute the discrepancy function to obtain the marginal likelihood of $N$ observations $\boldsymbol{Y}_C = (Y_C)_{i=1}^N$, conditioned on the parameters $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ of the mechanistic and discrepancy models respectively, as well as the calibration inputs $u_C$ and $\boldsymbol{v}_C$, given by

$$p(\boldsymbol{Y}_C | u_C, \boldsymbol{v}_C, \boldsymbol{\theta}, \boldsymbol{\phi}) \sim \mathcal{N}(\mathbf{f}_{\boldsymbol{\theta}, u_C}, \boldsymbol{\Sigma}_{NN} + \sigma^2 \boldsymbol{I}), \tag{S2.4}$$

where $\mathbf{f}_{\boldsymbol{\theta}, u_C} = [f_1(\boldsymbol{\theta}, u_C), \ldots, f_N(\boldsymbol{\theta}, u_C)]$ is a vector collecting the $N$ evaluations of the mechanistic model function, $\boldsymbol{\Sigma}_{NN}$ represents the covariance function ( Eq. S2.3) evaluated on all $N \times N$ pairs of the calibrations inputs $\boldsymbol{v}_C$:

$$\boldsymbol{\Sigma}_{NN} = \begin{pmatrix} \boldsymbol{\kappa}(\boldsymbol{v}_C^1, \boldsymbol{v}_C^1; \boldsymbol{\phi}) & \ldots & \boldsymbol{\kappa}(\boldsymbol{v}_C^1, \boldsymbol{v}_C^N; \boldsymbol{\phi}) \\ \vdots & \ddots & \vdots \\ \boldsymbol{\kappa}(\boldsymbol{v}_C^N, \boldsymbol{v}_C^1; \boldsymbol{\phi}) & \cdots & \boldsymbol{\kappa}(\boldsymbol{v}_C^N, \boldsymbol{v}_C^N; \boldsymbol{\phi}) \end{pmatrix}, \tag{S2.5}$$

and $\boldsymbol{I}$ is a $N \times N$ identity matrix.

### Inference of the model and GP parameters

We proceed by first placing suitable prior distributions, $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\phi})$, on the model and GP parameters and then obtain the posterior distribution using Bayes theorem as follows:

$$p(\boldsymbol{\theta}, \boldsymbol{\phi} | \boldsymbol{Y}_C, u_C, \boldsymbol{v}_C) \propto p(\boldsymbol{Y}_C | u_C, \boldsymbol{v}_C, \boldsymbol{\theta}, \boldsymbol{\phi}) p(\boldsymbol{\theta}) p(\boldsymbol{\phi}). \tag{S2.6}$$

Since this posterior distribution is analytically intractable due to the non-linear dependence on $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ we resort to Markov chain Monte Carlo (MCMC) in order to obtain samples from this distribution.

### Predictions

Having inferred the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ we may want to predict the output of the model in Eq. S2.1 for a new set of model inputs $u_V$ and $\boldsymbol{v}_V$. Note that these new model inputs are considered as validation inputs, denoted with subscript $V$. For the purpose of derivation here, we consider the number of validation points $M$ to be different than the number of measurement points $N$, although these numbers can be the same for specific choices of calibrations.

We denote the column vector for the corresponding $M$ predicted outputs as $\boldsymbol{Y}_V = (Y_V)_{i=1}^M$, and the model evaluations with the new inputs as $\mathbf{f}_{\boldsymbol{\theta}, u_V} = [f_1(\boldsymbol{\theta}, u_V), \ldots, f_M(\boldsymbol{\theta}, u_V)]^T$ Furthermore, we denote the collection of calibration inputs at the $N$ training (points corresponding to the measurements) as $\boldsymbol{I}_C = (u_C, \boldsymbol{v}_C)$, and at the prediction points as $\boldsymbol{I}_V = (u_V, \boldsymbol{v}_V)$

Note that for a fixed value of parameters, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ respectively, we can analytically obtain the predictive distribution of $\boldsymbol{Y}_V$ given by

$$p(\boldsymbol{Y}_V|\boldsymbol{I}_V, \boldsymbol{I}_C, \boldsymbol{Y}_C, \boldsymbol{\theta}, \boldsymbol{\phi}) = \mathcal{N}(\boldsymbol{\mu_V}, \boldsymbol{\sigma_V^2}), \tag{S2.7}$$

where the mean and variance is given by [4]

$$\begin{aligned} \boldsymbol{\mu_V} &= \mathbf{f}_{\boldsymbol{\theta},u_V} + \boldsymbol{\Sigma}_{MN}[\boldsymbol{\Sigma}_{NN} + \sigma^2\boldsymbol{I}]^{-1}(\boldsymbol{Y}_C - \mathbf{f}_{\boldsymbol{\theta},u_C}) \\ \boldsymbol{\sigma_V^2} &= \boldsymbol{\Sigma}_{MM} - \boldsymbol{\Sigma}_{MN}[\boldsymbol{\Sigma}_{NN} + \sigma^2\boldsymbol{I}]^{-1}\boldsymbol{\Sigma}_{NM}, \end{aligned} \tag{S2.8}$$

where $\boldsymbol{\Sigma}_{MN}$ and $\boldsymbol{\Sigma}_{NM}$ denotes the $M \times N$ and $N \times M$ matrices of covariance function evaluations between the training and prediction inputs given by

$$\boldsymbol{\Sigma}_{MN} = \begin{pmatrix} \boldsymbol{\kappa}(\boldsymbol{v}_V^1, \boldsymbol{v}_C^1; \boldsymbol{\phi}) & \dots & \boldsymbol{\kappa}(\boldsymbol{v}_V^1, \boldsymbol{v}_C^N; \boldsymbol{\phi}) \\ \vdots & \ddots & \vdots \\ \boldsymbol{\kappa}(\boldsymbol{v}_V^M, \boldsymbol{v}_C^1; \boldsymbol{\phi}) & \cdots & \boldsymbol{\kappa}(\boldsymbol{v}_V^M, \boldsymbol{v}_C^N; \boldsymbol{\phi}) \end{pmatrix}, \tag{S2.9}$$

$$\boldsymbol{\Sigma}_{NM} = \begin{pmatrix} \boldsymbol{\kappa}(\boldsymbol{v}_C^1, \boldsymbol{v}_V^1; \boldsymbol{\phi}) & \dots & \boldsymbol{\kappa}(\boldsymbol{v}_C^1, \boldsymbol{v}_V^M; \boldsymbol{\phi}) \\ \vdots & \ddots & \vdots \\ \boldsymbol{\kappa}(\boldsymbol{v}_C^N, \boldsymbol{v}_V^1; \boldsymbol{\phi}) & \cdots & \boldsymbol{\kappa}(\boldsymbol{v}_C^N, \boldsymbol{v}_V^M; \boldsymbol{\phi}) \end{pmatrix}, \tag{S2.10}$$

with inputs $\boldsymbol{v}_C$ and $\boldsymbol{v}_V$, respectively, and $\boldsymbol{\Sigma}_{MM}$ is the covariance evaluated at the prediction inputs $\boldsymbol{v}_V$ only:

$$\boldsymbol{\Sigma}_{MM} = \begin{pmatrix} \boldsymbol{\kappa}(\boldsymbol{v}_V^1, \boldsymbol{v}_V^1; \boldsymbol{\phi}) & \dots & \boldsymbol{\kappa}(\boldsymbol{v}_V^1, \boldsymbol{v}_V^M; \boldsymbol{\phi}) \\ \vdots & \ddots & \vdots \\ \boldsymbol{\kappa}(\boldsymbol{v}_V^M, \boldsymbol{v}_V^1; \boldsymbol{\phi}) & \cdots & \boldsymbol{\kappa}(\boldsymbol{v}_V^M, \boldsymbol{v}_V^M; \boldsymbol{\phi}) \end{pmatrix}. \tag{S2.11}$$

Finally, to obtain the marginal (i.e. integrating out the parameters) predictive distribution:

$$p(\boldsymbol{Y}_V|\boldsymbol{I}_V, \boldsymbol{I}_C, \boldsymbol{Y}_C) = \int \mathcal{N}(\boldsymbol{\mu_V}, \boldsymbol{\sigma_V^2}) p(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{Y}_C, u_C, \boldsymbol{v}_C)\, \mathrm{d}\,\boldsymbol{\theta}\, \mathrm{d}\,\boldsymbol{\phi}, \tag{S2.12}$$

we use Monte Carlo integration using the samples of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ obtained through MCMC.

## Sparse Gaussian Process

The above formulation of the discrepancy model suffers from a crucial computational bottleneck stemming from the need of inverting the covariance matrix $\boldsymbol{\Sigma}_{NN}$ while evaluating the marginal likelihood in Eq. S2.4, as well as drawing posterior predictions in Eq. S2.12 (in turn using Eq. S2.8). In all the calibration problems under consideration here, we have a large number of data points (time series measurements) where $N \geq 80000$. Thus, it becomes infeasible to apply Gaussian processes for modelling the discrepancy without tackling this excessive computational load related to repeated inversion of a large matrix.

In order to alleviate this computational bottleneck we use a sparse approximation of the true covariance function. Quiñonero-Candela et al. [5] provides an extensive review of such sparse approximations techniques. Following [5] we use a set of $P$ or inducing inputs (or pseudo-inputs) $x_C$ with associated latent function $\delta(\boldsymbol{\phi}, x_C)$ representing the discrepancy function corresponding to the inducing inputs. This inducing function is assigned a zero mean GP prior as follows:

$$\delta(\boldsymbol{\phi}, x_C) \sim \mathcal{GP}\big(0, \boldsymbol{\kappa}(x_C, x_C'; \boldsymbol{\phi})\big). \tag{S2.13}$$

Let us denote the vector of discrepancy function evaluations at all the training points as $\delta_{\boldsymbol{\phi},u_C} = [\delta_1(\boldsymbol{\phi}, u_C), \dots, \delta_N(\boldsymbol{\phi}, u_C)]^T$ and at inducing points as $\delta_{\boldsymbol{\phi},x_C} = [\delta_1(\boldsymbol{\phi}, x_C), \dots, \delta_P(\boldsymbol{\phi}, x_C)]^T$. We can then write the joint prior as a product of all the training and inducing points as $p(\delta_{\boldsymbol{\phi},u_C}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{NN})$ and $p(\delta_{\boldsymbol{\phi},x_C}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{PP})$ respectively, where $\boldsymbol{\Sigma}_{PP}$ denotes the covariance evaluated at all pairs of inducing inputs:

$$\boldsymbol{\Sigma}_{PP} = \begin{pmatrix} \boldsymbol{\kappa}(x_C^1, x_C^1; \boldsymbol{\phi}) & \dots & \boldsymbol{\kappa}(x_C^1, x_C^P; \boldsymbol{\phi}) \\ \vdots & \ddots & \vdots \\ \boldsymbol{\kappa}(x_C^P, x_C^1; \boldsymbol{\phi}) & \cdots & \boldsymbol{\kappa}(x_C^P, x_C^P; \boldsymbol{\phi}) \end{pmatrix}. \tag{S2.14}$$

We can then approximate the prior on the true discrepancy function $\delta_{\boldsymbol{\phi},u_C}$ marginalising the inducing discrepancies as:

$$\begin{aligned} p(\delta_{\boldsymbol{\phi},u_C}) \approx p(\delta_{\boldsymbol{\phi},u_C}|\delta_{\boldsymbol{\phi},x_C}) &= \int p(\delta_{\boldsymbol{\phi},u_C}|\delta_{\boldsymbol{\phi},x_C}) p(\delta_{\boldsymbol{\phi},x_C}) \,\mathrm{d}\, \delta_{\boldsymbol{\phi},x_C} \\ &= \mathcal{N}(\boldsymbol{\Sigma}_{NP}\boldsymbol{\Sigma}_{PP}^{-1}\delta_{\boldsymbol{\phi},x_C}, \boldsymbol{\Sigma}_{NN} - \boldsymbol{\Sigma}_{NP}\boldsymbol{\Sigma}_{PP}^{-1}\boldsymbol{\Sigma}_{PN}), \end{aligned} \tag{S2.15}$$

where $\boldsymbol{\Sigma}_{NP}$, $\boldsymbol{\Sigma}_{PN}$ denotes the covariance matrices containing the cross-covariances between the training and inducing inputs (evaluated in the same way as in Eqs. S2.9, S2.10). This sparse approximation was first introduced in [6] to scale the GP regression model. This approximation is widely known as the *fully independent training conditional* (FITC) approximation in machine learning parlance since the introduction of these inducing inputs and corresponding function values $\delta_{\boldsymbol{\phi},x_C}$ induces a conditional independence among all the elements of $\delta_{\boldsymbol{\phi},u_C}$ [5], that is we have

$$p(\delta_{\boldsymbol{\phi},u_C}|\delta_{\boldsymbol{\phi},x_C}) = \prod_{i=1}^{N} p(\delta_i(\boldsymbol{\phi}, \boldsymbol{v}_C^i)|\delta_{\boldsymbol{\phi},x_C}) = \mathcal{N}(\boldsymbol{\Sigma}_{NP}\boldsymbol{\Sigma}_{PP}^{-1}\delta_{\boldsymbol{\phi},x_C}, \boldsymbol{\Sigma}_{NN} - \boldsymbol{\Sigma}_{NP}\boldsymbol{\Sigma}_{PP}^{-1}\boldsymbol{\Sigma}_{PN}). \tag{S2.16}$$

Using this approximate prior $p(\delta_{\boldsymbol{\phi},u_C}|\delta_{\boldsymbol{\phi},x_C})$ to obtain the marginal likelihood and the prediction terms we essentially approximate the true covariance $\boldsymbol{\Sigma}_{NN}$ as [5]:

$$\boldsymbol{\Sigma}_{NN} \approx \hat{\boldsymbol{\Sigma}} = \boldsymbol{Q} + \mathrm{diag}(\boldsymbol{\Sigma}_{NN} - \boldsymbol{Q}), \tag{S2.17}$$

where $\mathrm{diag}(A)$ is a diagonal matrix whose elements match the diagonal of $A$ and the matrix $\mathbf{Q}$ is given by

$$\boldsymbol{Q} = \boldsymbol{\Sigma}_{NP}\boldsymbol{\Sigma}_{PP}^{-1}\boldsymbol{\Sigma}_{PN}. \tag{S2.18}$$

$\hat{\boldsymbol{\Sigma}}$ has the same diagonal elements as $\boldsymbol{\Sigma}_{NN}$ and the off-diagonal elements are the same as for $\boldsymbol{Q}$. Thus, inversion of $\hat{\boldsymbol{\Sigma}}$ scales as $\mathcal{O}(NP^2)$ as opposed to $\mathcal{O}(N^3)$ for the inversion of $\boldsymbol{\Sigma}_{NN}$.

## S3 Modelling residuals using an ARMA$(p, q)$ process

In the previous section we modelled the discrepancy as a function drawn from a GP prior. Alternatively, we can address the case of discrepancy using a correlated residual approach (see Section 3.6.2 in [7] for an introduction to this modelling approach). In this case we can model the residuals between the data $(Y_C)_i$ and the mechanistic model $f_i(\boldsymbol{\theta}, u_C^i)$ as an ARMA$(p, q)$ process as follows:

$$\begin{aligned}(Y_C)_i - f_i(\boldsymbol{\theta}, u_C^i) &= e_i \\ &= \varphi_1 e_{i-1} + \ldots + \varphi_p e_{i-p} + \nu_i + \zeta_1 \nu_{i-1} + \ldots + \zeta_q \nu_{i-1-q},\end{aligned} \tag{S3.19}$$

where

$$\nu_i \sim \mathcal{N}(0, \tau^2), \tag{S3.20}$$

and $\boldsymbol{\varphi} = [\varphi_1, \ldots, \varphi_p]^T$, $\boldsymbol{\zeta} = [\zeta_1, \ldots, \zeta_p]^T$ are the vectors representing the $p \geq 0$ autoregressive coefficients and $q \geq 0$ moving-average coefficients of the ARMA process.

The rationale behind this modelling approach comes from the fact that if the mechanistic model is able to explain the measurements adequately then the residuals are essentially uncorrelated measurement noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Note that we use a different symbol $\nu$, as opposed to $\epsilon$, to represent the noise term in order to highlight the difference in its interpretations. However, the existence of discrepancy between the model output and the observations points to the fact that the residuals, for each data sample, has unexplained structure that can be modelled using a pre-determined correlation structure, as expressed through an ARMA$(p, q)$ model.

### Inference

We first re-write the normally distributed error term $\nu_i$ as

$$\nu_i = (Y_C)_i - f_i(\boldsymbol{\theta}, u_C^i) - \sum_{j=1}^{p} \varphi_j \{(Y_C)_{i-j} - f_{i-j}(\boldsymbol{\theta}, u_C^{i-j})\} - \sum_{k=1}^{q} \zeta_k \nu_{i-k}, \tag{S3.21}$$

using which we can write the conditional likelihood of the observed data for $N$ measurements as [8]

$$p(\boldsymbol{Y}_C|\boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\zeta}, \tau) = (2\pi\tau^2)^{N/2} \exp\left( -\frac{1}{2\tau^2} \sum_{i=p+1}^{N} \nu_i^2 \right), \tag{S3.22}$$

where we have used $\nu_i$ for $i \geq p+1$ by assuming that $\nu_p = \nu_{p-1} = \ldots = \nu_{p+1-q} = 0$, its expected value. Note that to calculate the likelihood for all the $N$ measurements requires us to introduce extra parameter values, as latent variables, for the past history of the data as well as the error terms before measurement commences, that is for $[(Y_C)_0, (Y_C)_{-1}, \ldots, (Y_C)_{1-p}]$ and $[\nu_0, \nu_{-1}, \ldots, \nu_{1-q}]$. Alternatively, we can reformulate Eq. S3.19 in a state space form and use the Kalman filter algorithm to evaluate the unconditional full likelihood for all $i$. We refer the reader to [7] for the details of this approach. We point out here that the difference between these two approaches of calculating the likelihood is insignificant for long time series, which is the case for our calibration problems with $N \geq 80000$.

Having defined the likelihood we can again adopt the Bayesian framework to infer posterior distributions of the model parameters $\boldsymbol{\theta}$ and the set of ARMA parameters $\boldsymbol{\phi} = [\boldsymbol{\varphi}, \boldsymbol{\zeta}]$, by choosing suitable prior distributions $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\phi})$ respectively and using the Bayes theorem to obtain the posterior given by

$$p(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{Y}_C) \propto p(\boldsymbol{Y}_C|\boldsymbol{\theta}, \boldsymbol{\phi}) p(\boldsymbol{\theta}) p(\boldsymbol{\phi}). \tag{S3.23}$$

Note that we have considered the noise variance known since its maximum likelihood estimate is given by $\tau^2 = \frac{\sum_{i=p+1}^{N} \nu_i^2}{N-(2p+q+1)}$, which can be easily obtained once estimates of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are available. Similar to the GP based inference problem we again use MCMC to obtain the desired posterior distributions.

### Predictions

In a purely time series modelling context, where models such as the ARMA is extensively used, predictions are used to forecast ahead in time for short intervals. In the context of our calibration problem we are generally interested in predicting outputs for a new calibration $u_V$. However, considering the fact that

we want to to predict $M$ output values corresponding to the new validation inputs, we can simply recast our predictions as one-step-ahead forecasts.

We denote the $M$ predicted outputs as $\boldsymbol{Y}_V = (Y_V)_{m=1}^{M}$, while $\boldsymbol{Y}_p = (Y_C)_{N-(p-1)}^{N}$ and $\mathbf{f}_{\boldsymbol{\theta},u_C} = [f_N(\boldsymbol{\theta}, u_C), \ldots, f_{N-(p-1)}(\boldsymbol{\theta}, u_C)]$ are column vectors representing the last $p$ observations and model evaluations with the calibration $u_C$. We denote the vector of the last $q$ errors as $\boldsymbol{\nu}_C = [\nu_{N-(q-1)}, \ldots, \nu_N]^T$.

Note that in our formulation here, the $m^{\text{th}}$, $m = 1, \ldots, M$, prediction is to be considered as the $(N+1)^{\text{th}}$ prediction from the model in Eq. S3.19 with the following modification: we replace $f_{N+1}(\boldsymbol{\theta}, u_C^i)$ with $f_1(\boldsymbol{\theta}, u_V^1)$. Thus, for a particular value of the parameters we have

$$(Y_V)_m \sim \mathcal{N}(\mathbb{E}[(Y_V)_m|\boldsymbol{Y}_p, \boldsymbol{\nu}_C, \boldsymbol{\theta}, \boldsymbol{\phi}], \text{Var}[(Y_V)_m|\boldsymbol{Y}_p, \boldsymbol{\nu}_C, \boldsymbol{\theta}, \boldsymbol{\phi}]), \tag{S3.24}$$

where the mean and the variance of the one-step ahead prediction distribution is given by

$$\begin{aligned} \mathbb{E}[(Y_V)_m|\boldsymbol{Y}_p, \boldsymbol{\nu}_C, \boldsymbol{\theta}, \boldsymbol{\phi}] &= f_m(\boldsymbol{\theta}, u_V^m) + \boldsymbol{\varphi}^T(\boldsymbol{Y}_p - \mathbf{f}_{\boldsymbol{\theta},u_C}) + \boldsymbol{\zeta}^T\boldsymbol{\nu}_C, \\ \text{Var}[(Y_V)_m|\boldsymbol{Y}_p, \boldsymbol{\nu}_C, \boldsymbol{\theta}, \boldsymbol{\phi}] &= \text{Var}\left[(Y_V)_m - \mathbb{E}[(Y_V)_m|\boldsymbol{Y}_p, \boldsymbol{\nu}_C, \boldsymbol{\theta}, \boldsymbol{\phi}]\right] = \text{Var}[\nu_N] = \tau^2. \end{aligned} \tag{S3.25}$$

In order to quantify the uncertainty in the predictions we can integrate out the model and noise parameters [9]:

$$p((Y_V)_m|\boldsymbol{Y}_p, \boldsymbol{\nu}_C) = \int \mathcal{N}(\mathbb{E}[(Y_V)_m|\boldsymbol{Y}_p, \boldsymbol{\nu}_C, \boldsymbol{\theta}, \boldsymbol{\phi}]p(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{Y}_C)\,\mathrm{d}\,\boldsymbol{\theta}\,\mathrm{d}\,\boldsymbol{\phi}, \tag{S3.26}$$

where we use Monte Carlo integration as in Eq. S2.12.

In order to collect the full set of $M$ predictions $\boldsymbol{Y}_V$ we simply use the one-step-ahead forecasting distribution shown above in a recursive manner.

## S4 Choice of priors for the ion channel example

Here we specify the choice of priors for the ion channel example.

- For the ion channel ODE model parameters, we chose a uniform prior specified in Beattie *et al.* [10] and Lei *et al.* [11, 12].
- For the GP model, we have the unbiased noise parameter $\sigma$, the length-scale $\rho_i$, the marginal variance $\alpha$:
  - $\sigma$: Half-Normal prior with standard deviation of 25;
  - $\rho_i$: Inverse-Gamma prior with shape and scale being $(5, 5)$;
  - $\alpha$: Inverse-Gamma prior with shape and scale being $(5, 5)$.
- For the ARMA model, we have the autoregressive coefficients $\varphi_i$, and moving-average coefficients $\zeta_i$:
  - $\varphi_i$: Normal prior centred on the maximum likelihood estimates with standard deviation of 2.5;
  - $\zeta_i$: Normal prior centred on the maximum likelihood estimates with standard deviation of 2.5.

## S5 Computing and representing posterior predictive

To compute the posterior predictive, we follow Girolami [13] and write the posterior predictive in Eqs. S2.12 & S3.26 as

$$p(Y_V \mid Y_C) = \sum_k p(Y_V \mid \boldsymbol{\theta}_k, \boldsymbol{\phi}_k, Y_C) p(\boldsymbol{\theta}_k, \boldsymbol{\phi}_k \mid Y_C), \tag{S5.27}$$

where $\boldsymbol{\theta}_k, \boldsymbol{\phi}_k$ are the $k^{\text{th}}$ posterior sample of the parameters. We have checked the posterior predictive of the ODE models at a given time point is symmetric and similar to a Gaussian distribution, for the sake of simplicity, we therefore use summary statistics such as the predictive mean and credible intervals computed using variance to represent the posterior predictive in this paper. To obtain the predictive mean $\mathbb{E}[Y_V \mid Y_C]$ and variance $\text{Var}[Y_V \mid Y_C]$, we use

$$\mathbb{E}[Y_V \mid Y_C] = \sum_k \mathbb{E}[Y_V \mid \boldsymbol{\theta}_k, \boldsymbol{\phi}_k, Y_C] p(\boldsymbol{\theta}_k, \boldsymbol{\phi}_k \mid Y_C), \tag{S5.28}$$

$$\text{Var}[Y_V \mid Y_C] = \sum_k \left( \text{Var}[Y_V \mid \boldsymbol{\theta}_k, \boldsymbol{\phi}_k, Y_C] + \mathbb{E}[Y_V \mid \boldsymbol{\theta}_k, \boldsymbol{\phi}_k, Y_C]^2 \right) p(\boldsymbol{\theta}_k, \boldsymbol{\phi}_k \mid Y_C) - \mathbb{E}[Y_V \mid Y_C]^2. \tag{S5.29}$$

Finally, to show the 95% credible intervals of our predictions, we plot $\mathbb{E}[Y_V \mid Y_C] \pm 1.96\sigma_{Y_V}$ where $\sigma^2_{Y_V} = \text{Var}[Y_V \mid Y_C]$.

# S6 Supplementary results for the action potential example

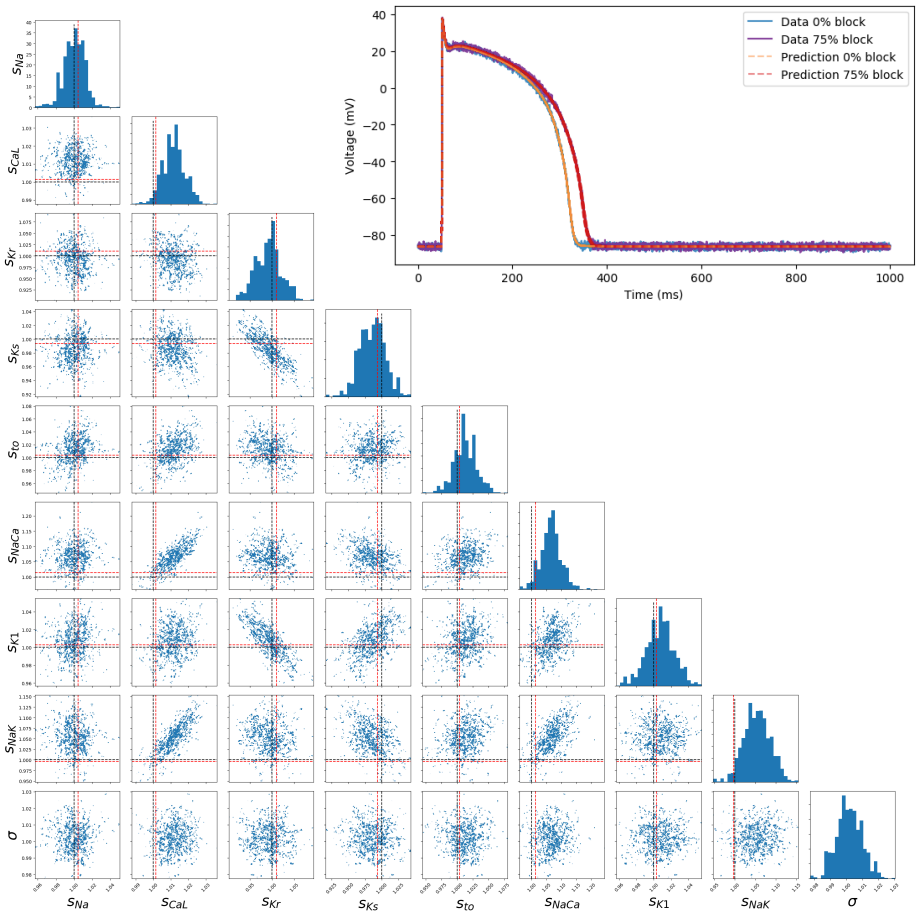


Figure S1: **Matrix plot and histograms**: Posterior distribution of Model T parameters. The dashed black lines indicate the true (data-generating) parameters; the dashed red lines are the result of the global optimisation routine. **Inset plot**: Posterior predictions for the 'context of use' (CoU), for the action potential model tutorial (in the scenario of no model discrepancy). The posterior predictions are model predictions simulated using the parameter samples from the posterior distribution; here, 200 samples/predictions are shown, Model T gives an almost-perfect prediction.

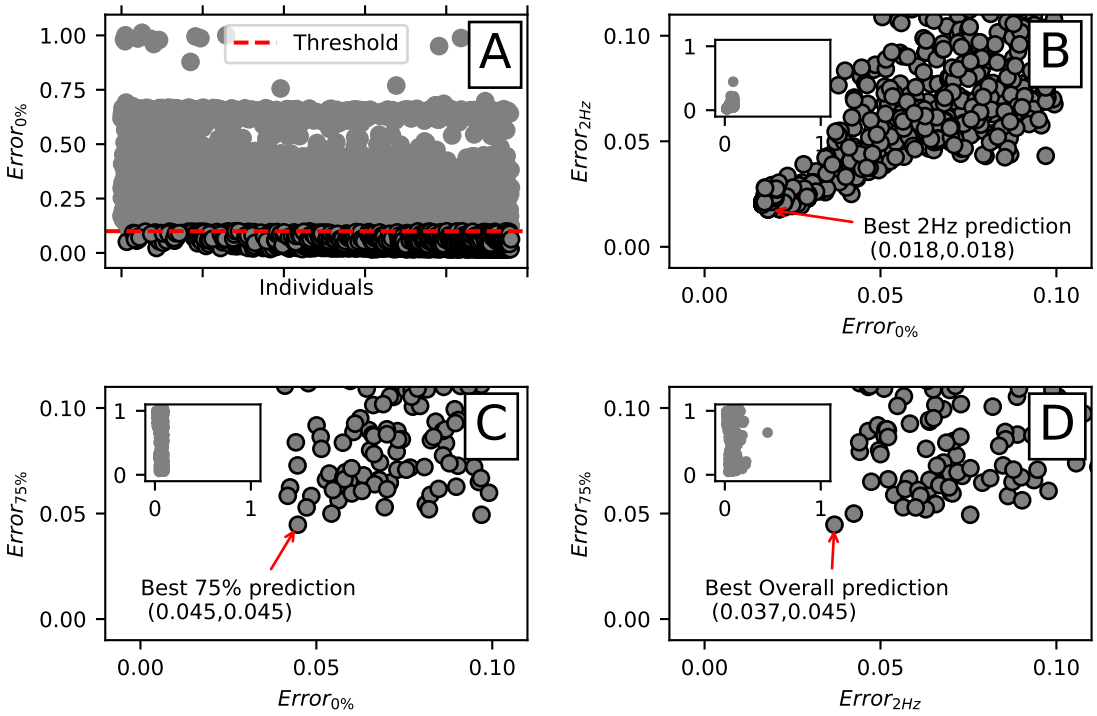


Figure S2: Results obtained via Genetic Algorithm and Differential Evolution to adjust the Fink Action Potential to ten Tusscher at control, i.e., 1 Hz, and 0% IKr block configuration. The Genetic Algorithm was used with a population size of 100 individuals and 10 generations. The Differential Evolution was used with 150 individuals and 15 generations. Both algorithms were implemented using the Python library Pygmo with the standard configurations. **A** From all the evaluations we selected the candidates (parameter sets) that satisfied $Error_{0\%} < 0.1$. Relative RMS errors are computed with $Error_X(i) = \frac{||T_X - aF_X(i)||_2}{||T_X||_2}$, where $T_X$ is the ten Tusscher model in scenario $X$, and $aF_X$ is the adjusted Fink model using the individual $i$ for scenario $X$. A total of 1079 candidates satisfied $Error_{0\%} < 0.1$, i.e., were below the displayed threshold. Using this metric, the best candidate had $Error_{0\%} = 1.6\%$. **B** Testing the performance of the 1079 candidates with respect to the 2Hz scenario a total of 990 candidates satisfy $Error_{0\%} < 0.1$ and $Error_{2Hz} < 0.1$. Using these two metrics, the best candidate had $Error_{0\%} = 1.8\%$ and $Error_{2Hz} = 1.8\%$. **C** Testing the performance of the 1079 candidates with respect to the 75% IKr block scenario only 80 candidates satisfy $Error_{0\%} < 0.1$ and $Error_{75\%} < 0.1$. Using these two metrics, the best candidate had $Error_{0\%} = 4.5\%$ and $Error_{75\%} = 4.5\%$. **D** Testing the performance of the 1079 candidates with respect to both 75% IKr block and 2Hz scenarios only 70 candidates satisfy $Error_{0\%} < 0.1$, $Error_{75\%} < 0.1$, and $Error_{2Hz} < 0.1$. Using the three metrics, the best candidate had $Error_{0\%} = 4.5\%$, $Error_{2Hz} = 3.7\%$, and $Error_{75\%} = 4.5\%$.

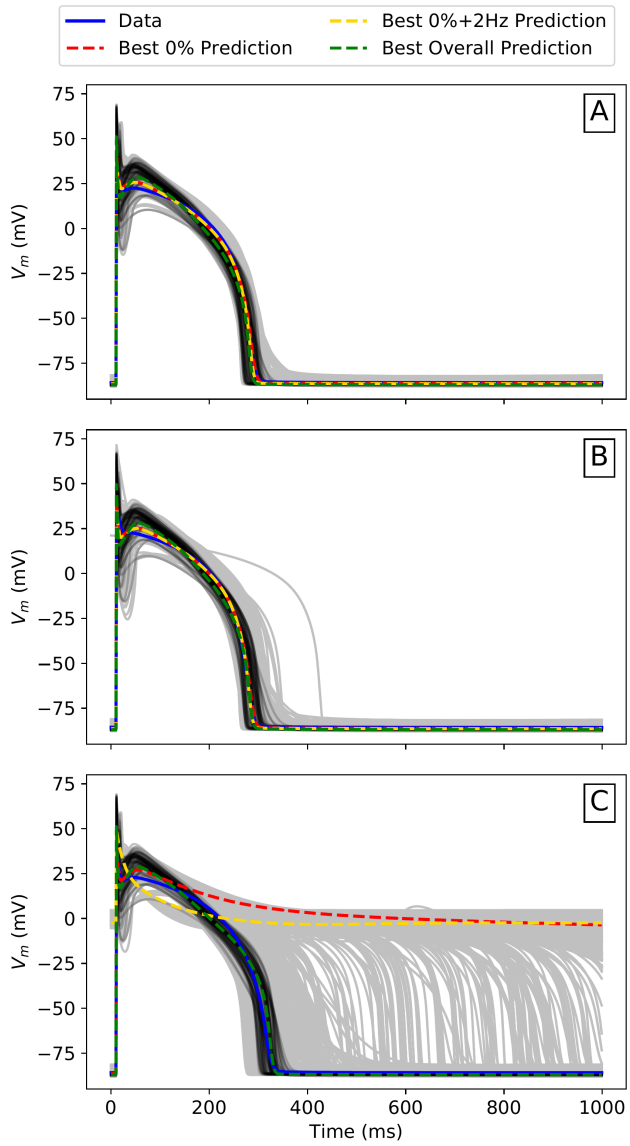


Figure S3: APs obtained by the Data (ten Tusscher model) and also by the fitting process for each scenario: **A** control, i.e., 1 Hz, and 0% IKr block; **B** 2 Hz; and **C** 75% IKr block. "Best 0% prediction" are the results obtained using the best candidate that satisfies $Error_{0\%} < 0.1$. "Best $0\% + 2Hz$ prediction" are results obtained using the best candidate that satisfies $Error_{0\%} < 0.1$ and $Error_{2Hz} < 0.1$. "Best overall prediction" are results obtained using the best candidate that satisfies $Error_{0\%} < 0.1$, $Error_{2Hz} < 0.1$, and $Error_{75\%} < 0.1$. The 1079 APs that satisfy $Error_{0\%} < 0.1$ are plotted with grey lines. The 70/1079 APs that satisfy $Error_{0\%} < 0.1$, $Error_{2Hz} < 0.1$, and $Error_{75\%} < 0.1$ are are plotted with black lines. Note that there is no way of knowing in advance what the best candidate parameter set will be without performing the experiment, so a distribution of possibilities should generally be shown.

## S7 Supplementary results for the ion channel discrepancy example

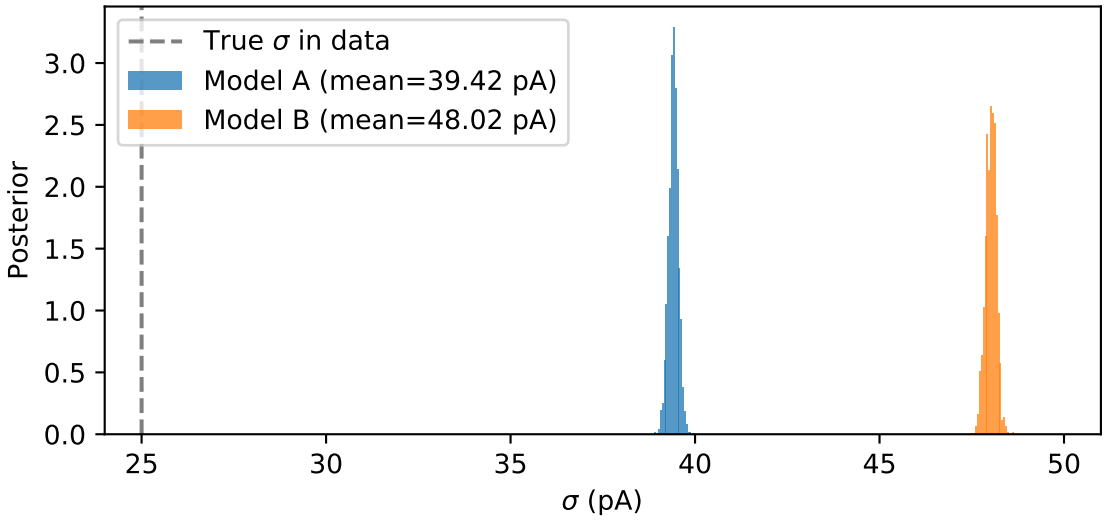


Figure S4: A comparison of the $\sigma$ values in the i.i.d. model for Model A and B, and $\sigma_{\text{true}}$ refers to the value used in generating the data with Model C. If we consider the inferred $\sigma$ value in Eq. 3 in the main text as $\sigma^2_{\text{inferred}} = \sigma^2_{\text{true}} + \sigma^2_{\text{discrepancy}}$, then we can see that both $\sigma_{\text{A}}, \sigma_{\text{B}} > \sigma_{\text{true}}$. Hence we have $\sigma^2_{\text{discrepancy}}$ term is non-zero for both models, which reflects the fact that there is discrepancy for both models. One may use the size of $\sigma_{\text{inferred}}$ to interpret the size of the model discrepancy here.

## S7.1 Model A

### S7.1.1 Model A: Full model predictions

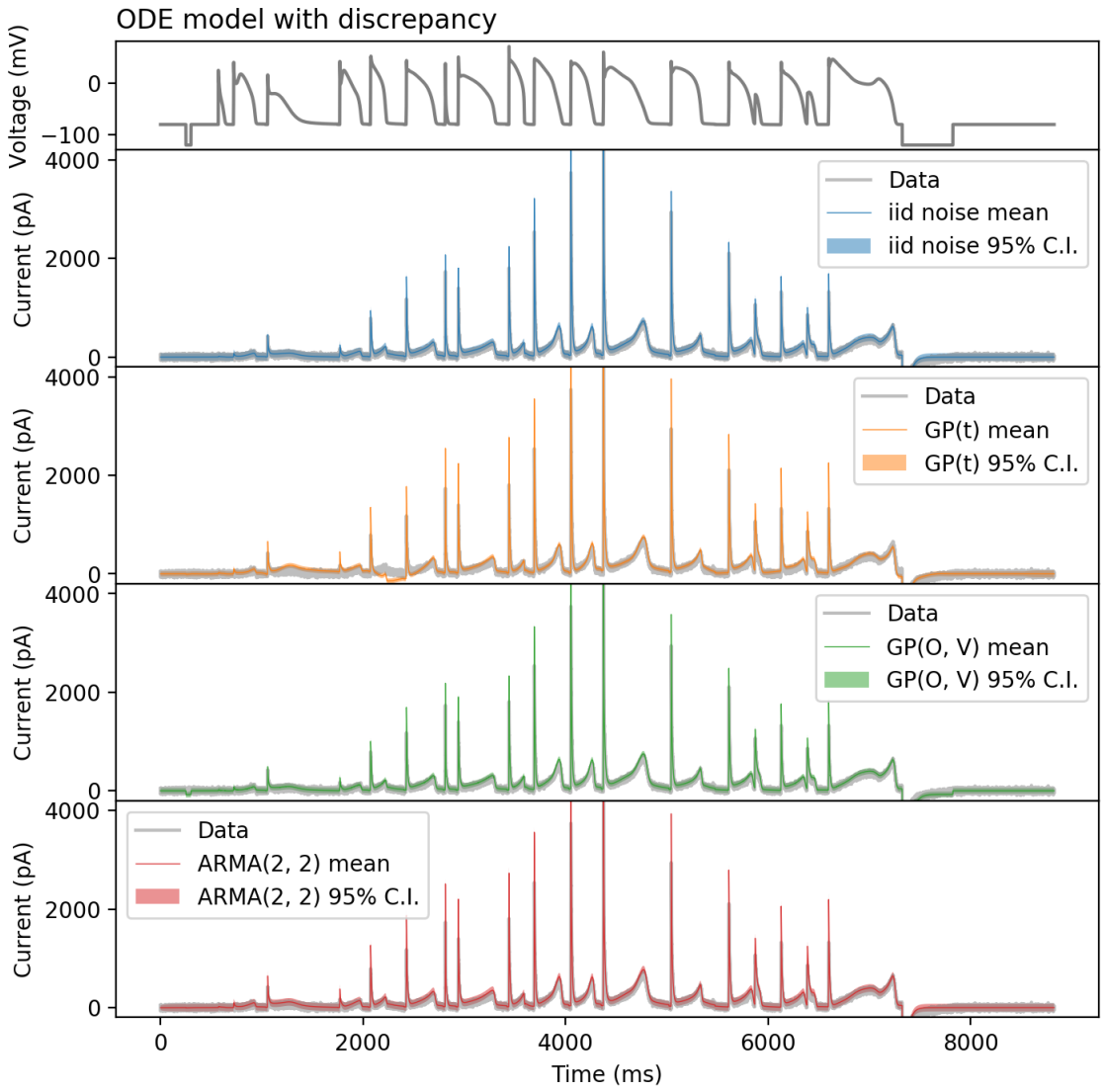


Figure S5: Model A prediction with different discrepancy models: no discrepancy (i.i.d. noise), GP($t$), GP($O, V$), and ARMA$(2, 2)$. The voltage clamp protocol for calibration is the action potential series protocol [10].

### S7.1.2 Model A: Discrepancy predictions

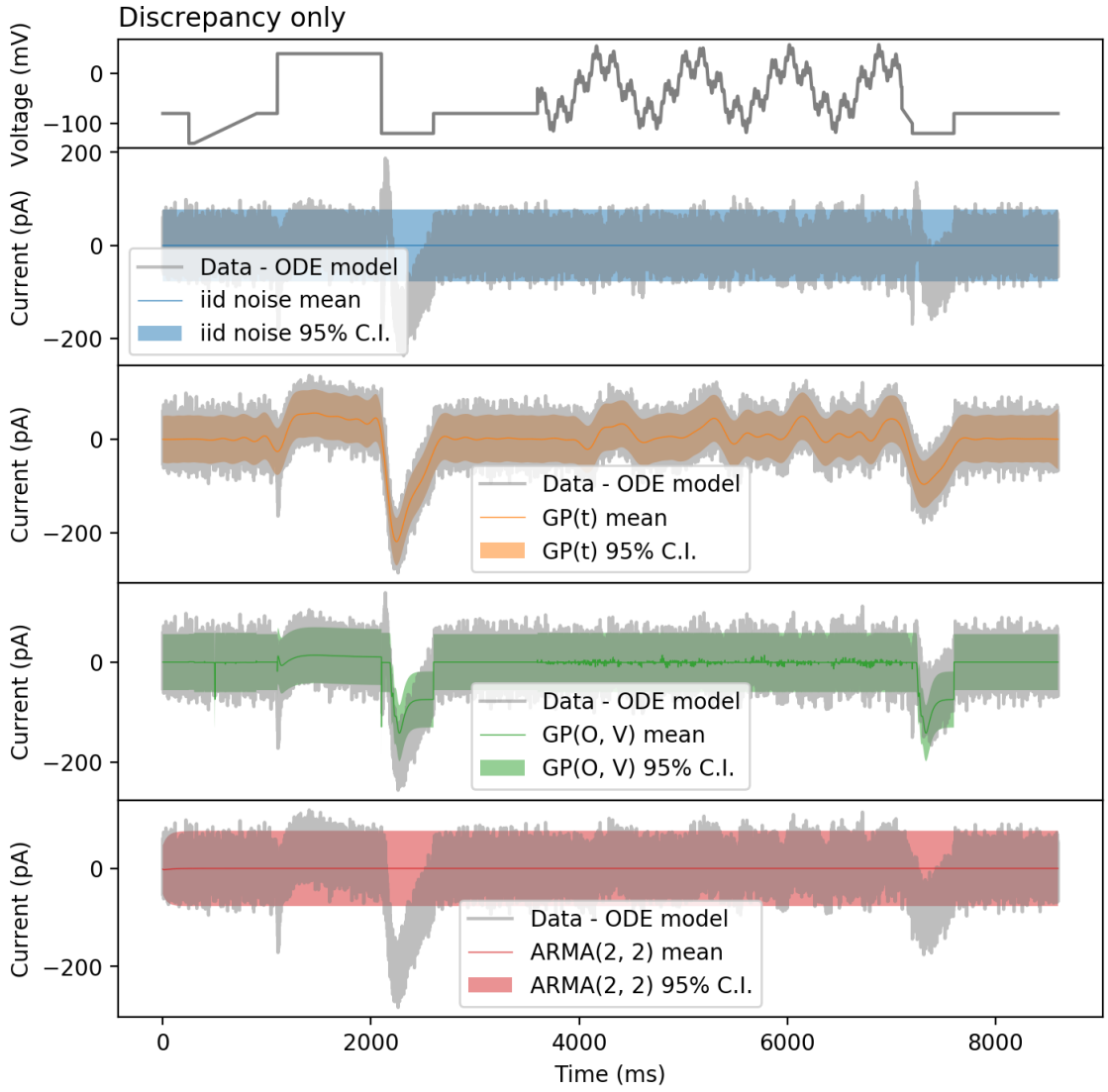


Figure S6: Model A fitting residuals of the MAP estimate accounted by different discrepancy models: no discrepancy (i.i.d. noise), GP($t$), GP($O, V$), and ARMA(2, 2). The voltage clamp protocol for calibration is the sinusoidal protocol [10].

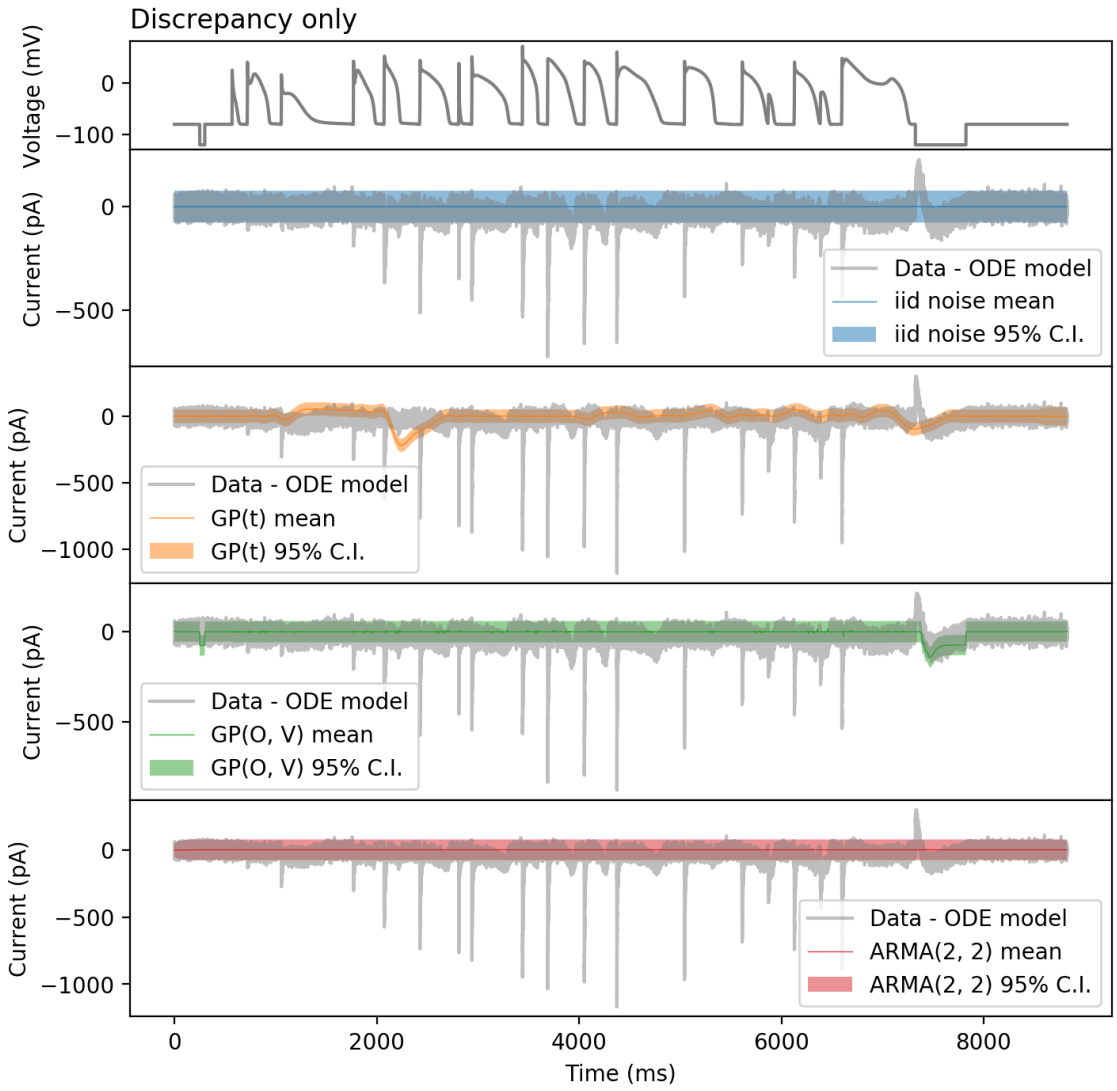


Figure S7: Model A prediction residuals of the MAP estimate accounted by different discrepancy models: no discrepancy (i.i.d. noise), GP($t$), GP($O, V$), and ARMA$(2, 2)$. The voltage clamp protocol for calibration is the action potential series protocol [10].

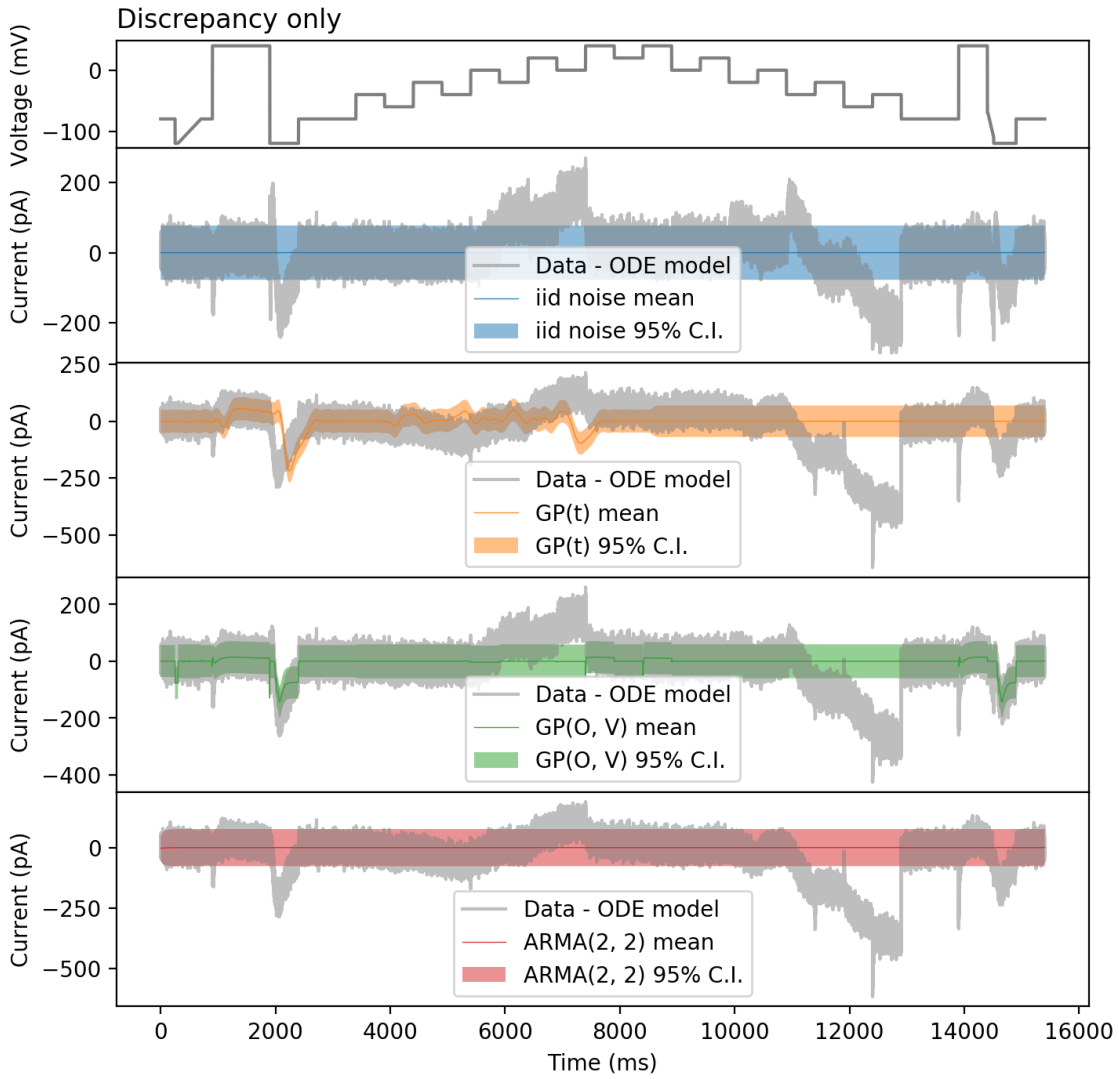


Figure S8: Model A prediction residuals of the MAP estimate accounted by different discrepancy models: no discrepancy (i.i.d. noise), GP($t$), GP($O, V$), and ARMA(2, 2). The voltage clamp protocol for calibration is the staircase protocol [11].

### S7.1.3 Model A: ODE model predictions

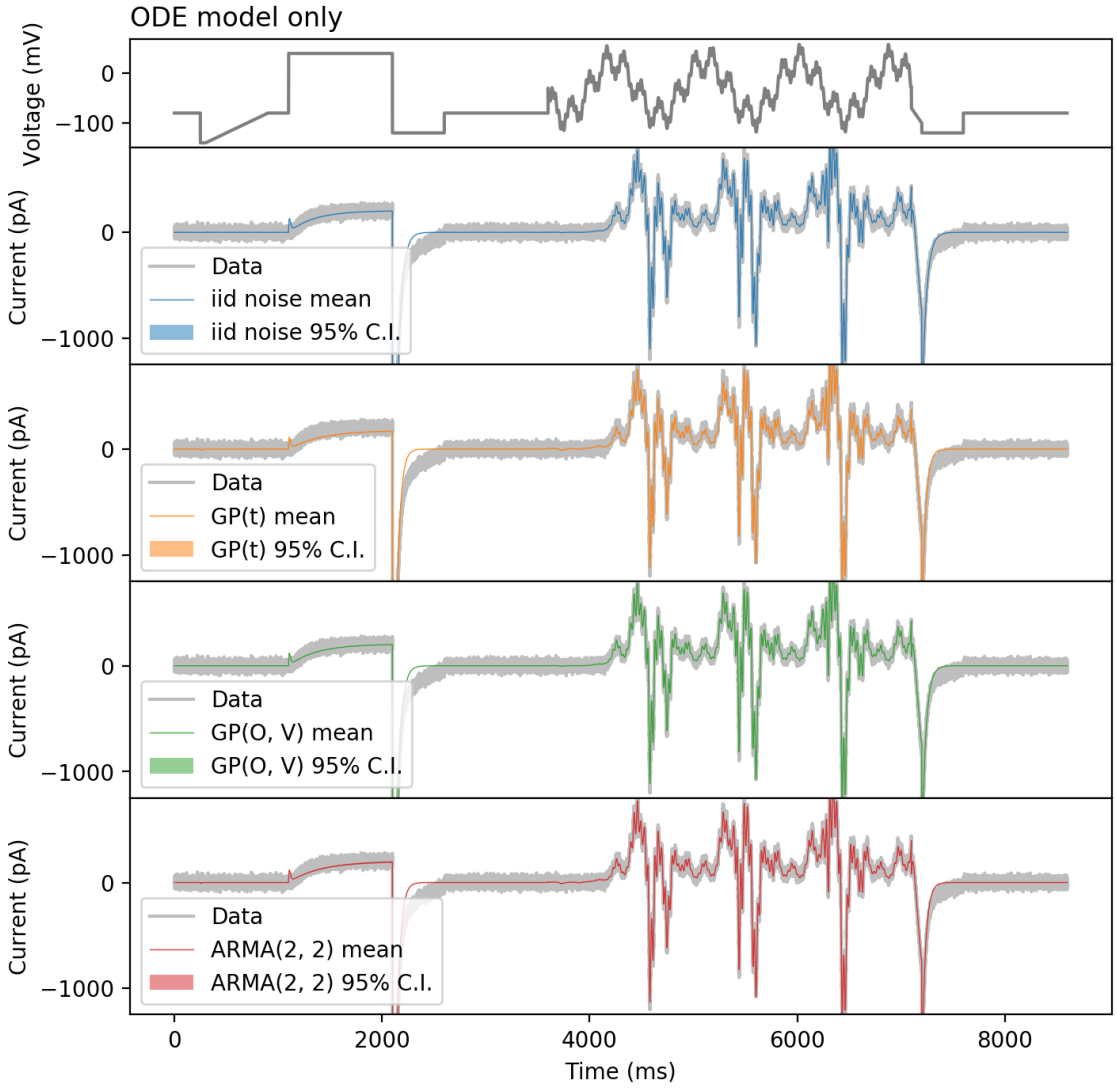


Figure S9: Fitting of the ODE model of Model A, using different discrepancy models: no discrepancy (i.i.d. noise), GP($t$), GP($O, V$), and ARMA(2, 2). The voltage clamp protocol for calibration is the sinusoidal protocol [10].

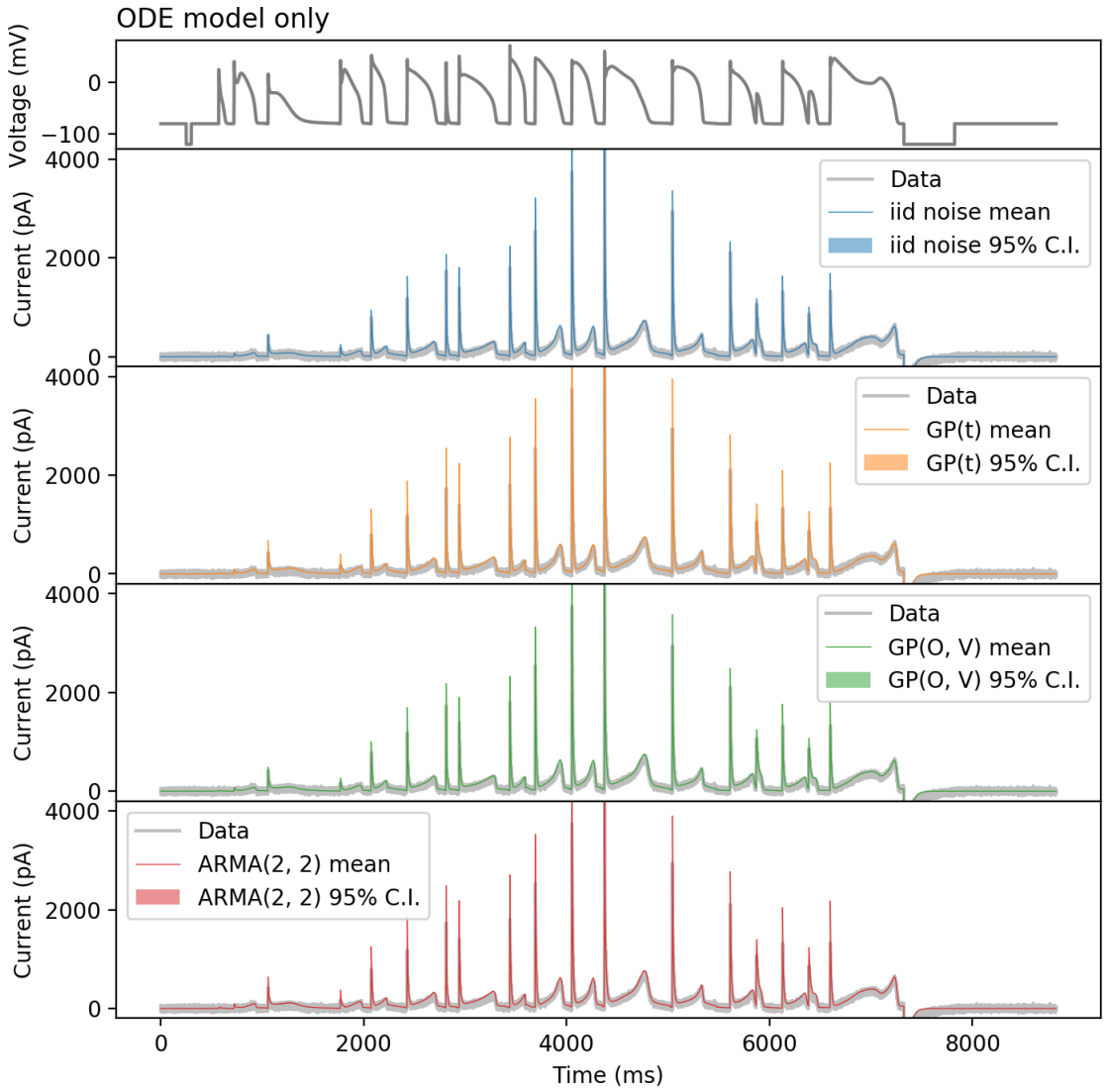


Figure S10: Predictions of the ODE model of Model A, using different discrepancy models: no discrepancy (i.i.d. noise), GP($t$), GP($O, V$), and ARMA$(2, 2)$. The voltage clamp protocol for calibration is the action potential series protocol [10].

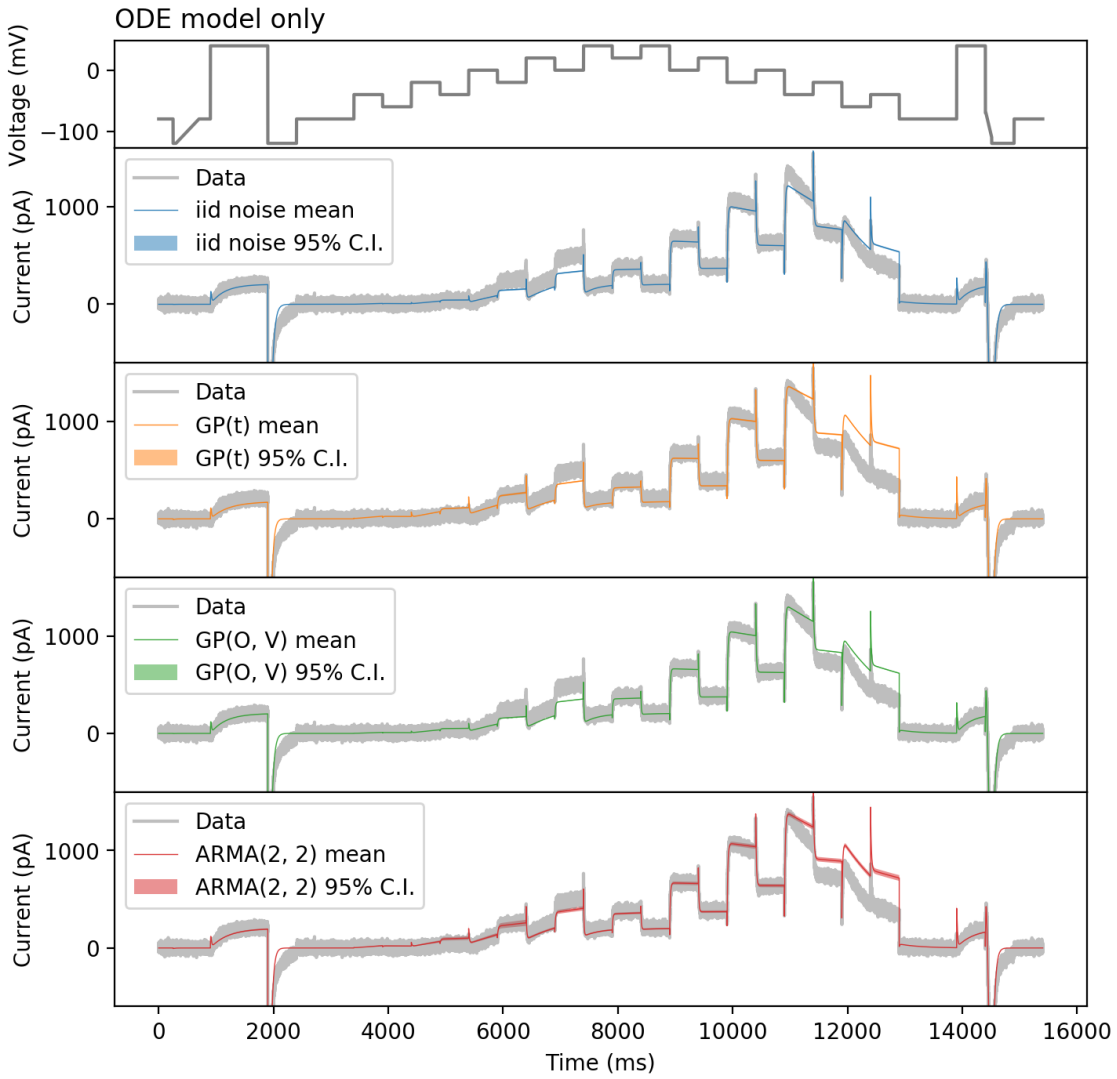


Figure S11: Predictions of the ODE model of Model A, using different discrepancy models: no discrepancy (i.i.d. noise), GP($t$), GP($O, V$), and ARMA$(2, 2)$. The voltage clamp protocol for calibration is the staircase protocol [11].

## S7.2 Model B

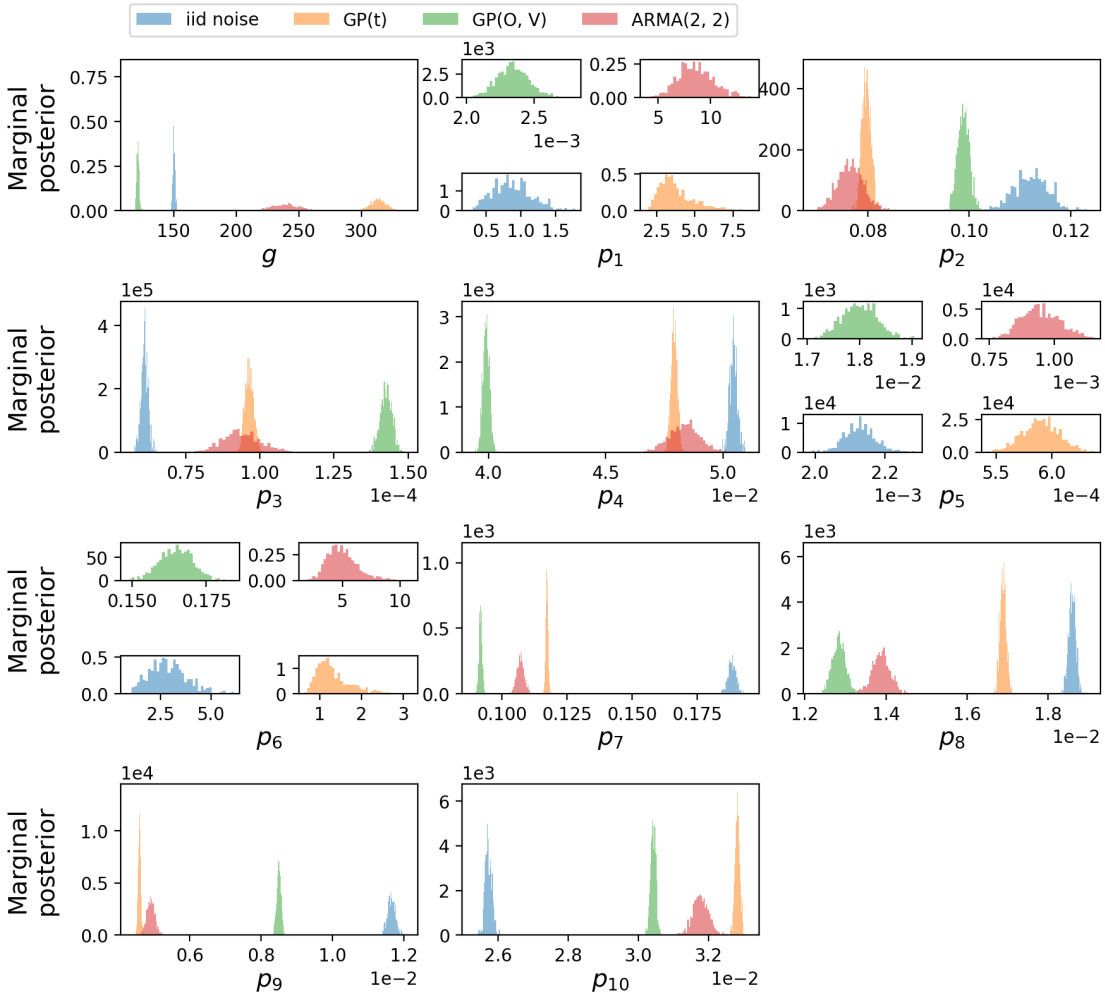


Figure S12: Model B inferred marginal posterior distributions for the conductance, $g$, and kinetic parameters $p_1, \ldots, p_{10}$ (a list of parameters referring to $A_{i,j}$ and $B_{i,j}$ in Eq. (8)) with different discrepancy models: i.i.d. noise (blue), GP$(t)$ (orange), GP$(O, V)$ (green), and ARMA$(2, 2)$ (red).

### S7.2.1 Model B: Full model predictions

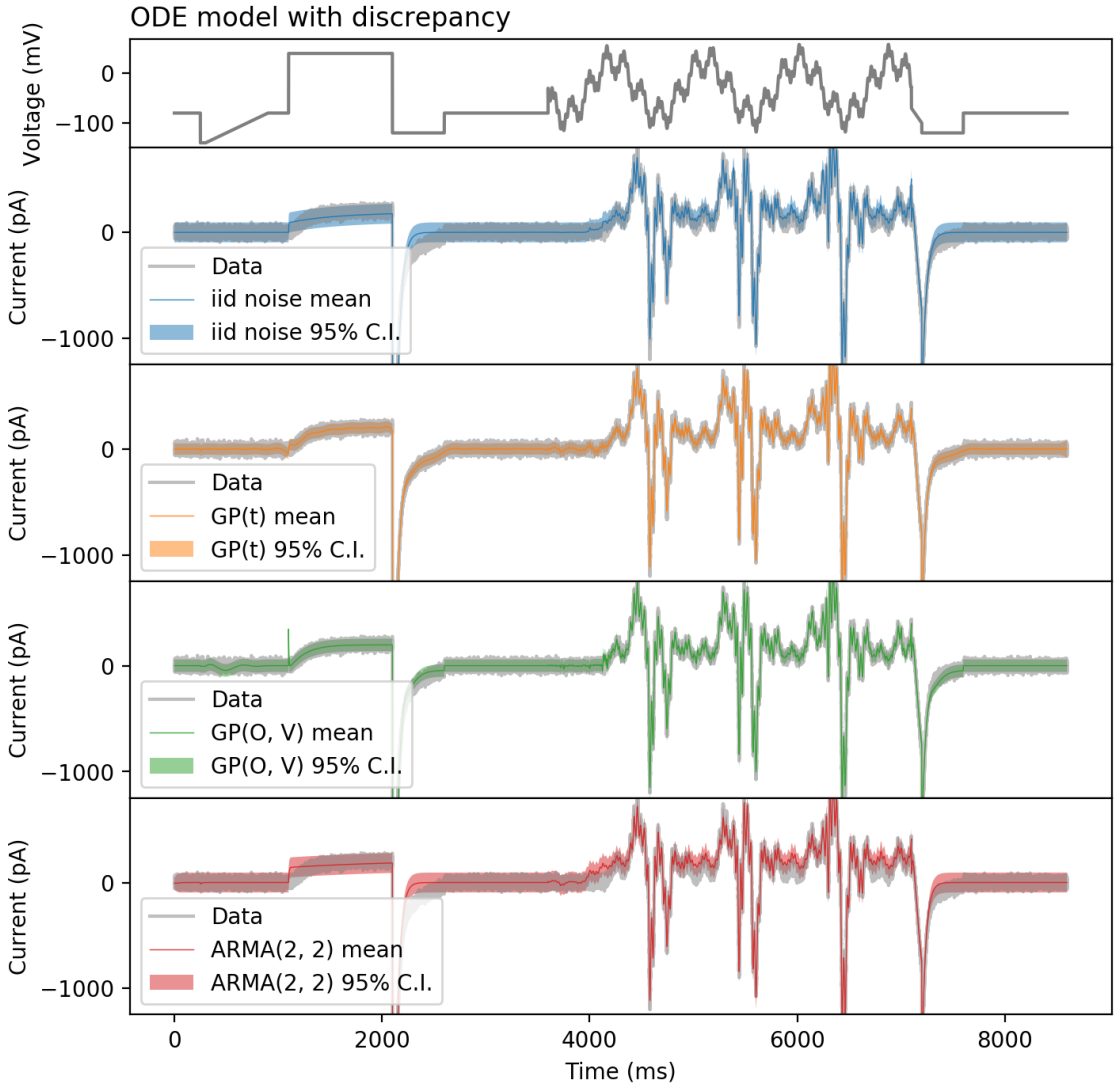


Figure S13: Model B fitting results with different discrepancy models: i.i.d. noise, GP($t$), GP($O, V$), and ARMA$(2, 2)$. The voltage clamp protocol for calibration is the sinusoidal protocol [10]. It shows the posterior predictive with the bounds showing the 95% credible interval.

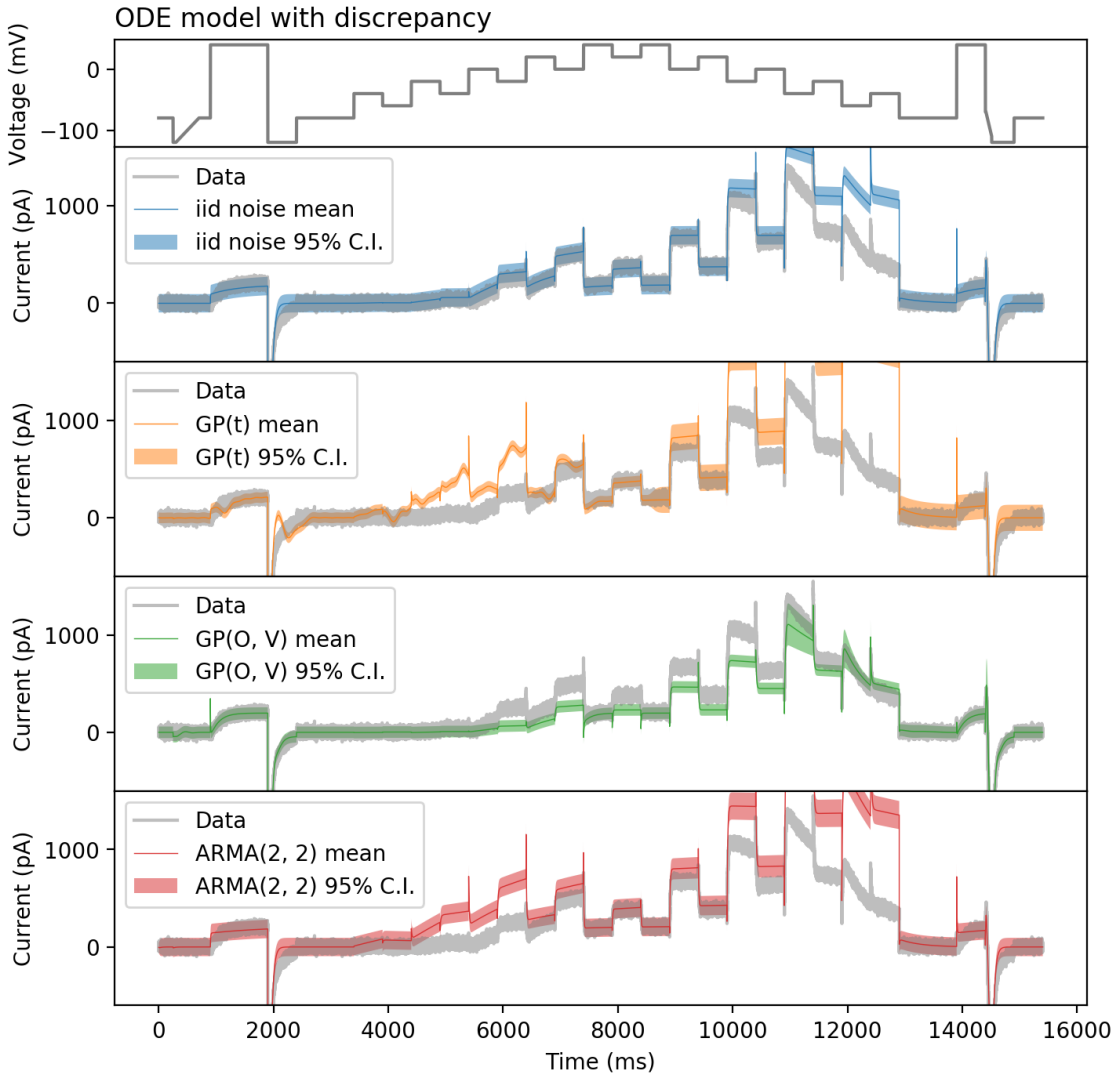


Figure S14: Model B prediction with different discrepancy models: i.i.d. noise, GP($t$), GP($O, V$), and ARMA$(2, 2)$. The voltage clamp protocol for calibration is the staircase protocol [11]. It shows the posterior predictive with the bounds showing the 95% credible interval.

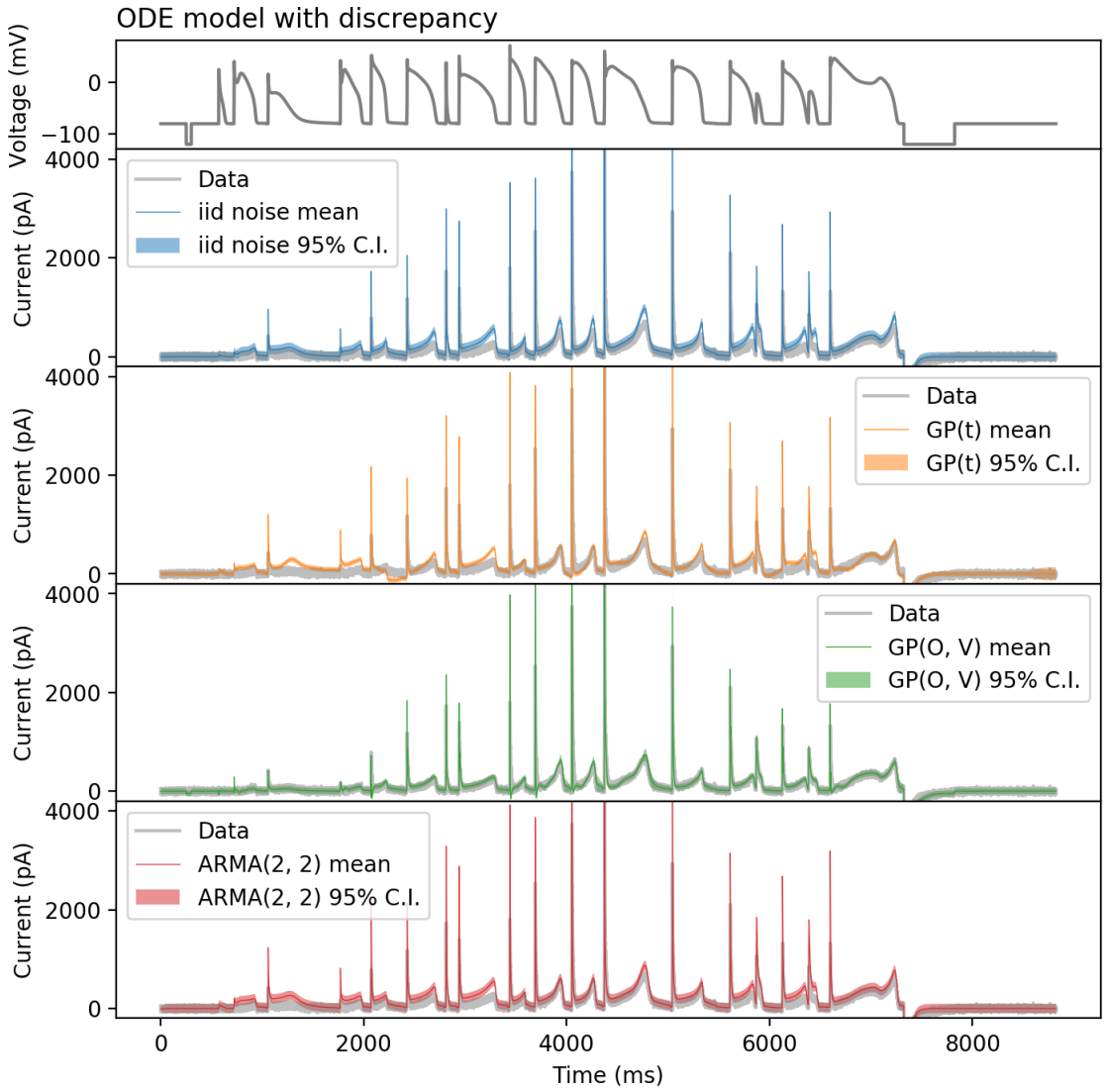


Figure S15: Model B prediction with different discrepancy models: no discrepancy (i.i.d. noise), GP($t$), GP($O, V$), and ARMA$(2, 2)$. The voltage clamp protocol for calibration is the action potential series protocol [10].

### S7.2.2 Model B: Discrepancy predictions

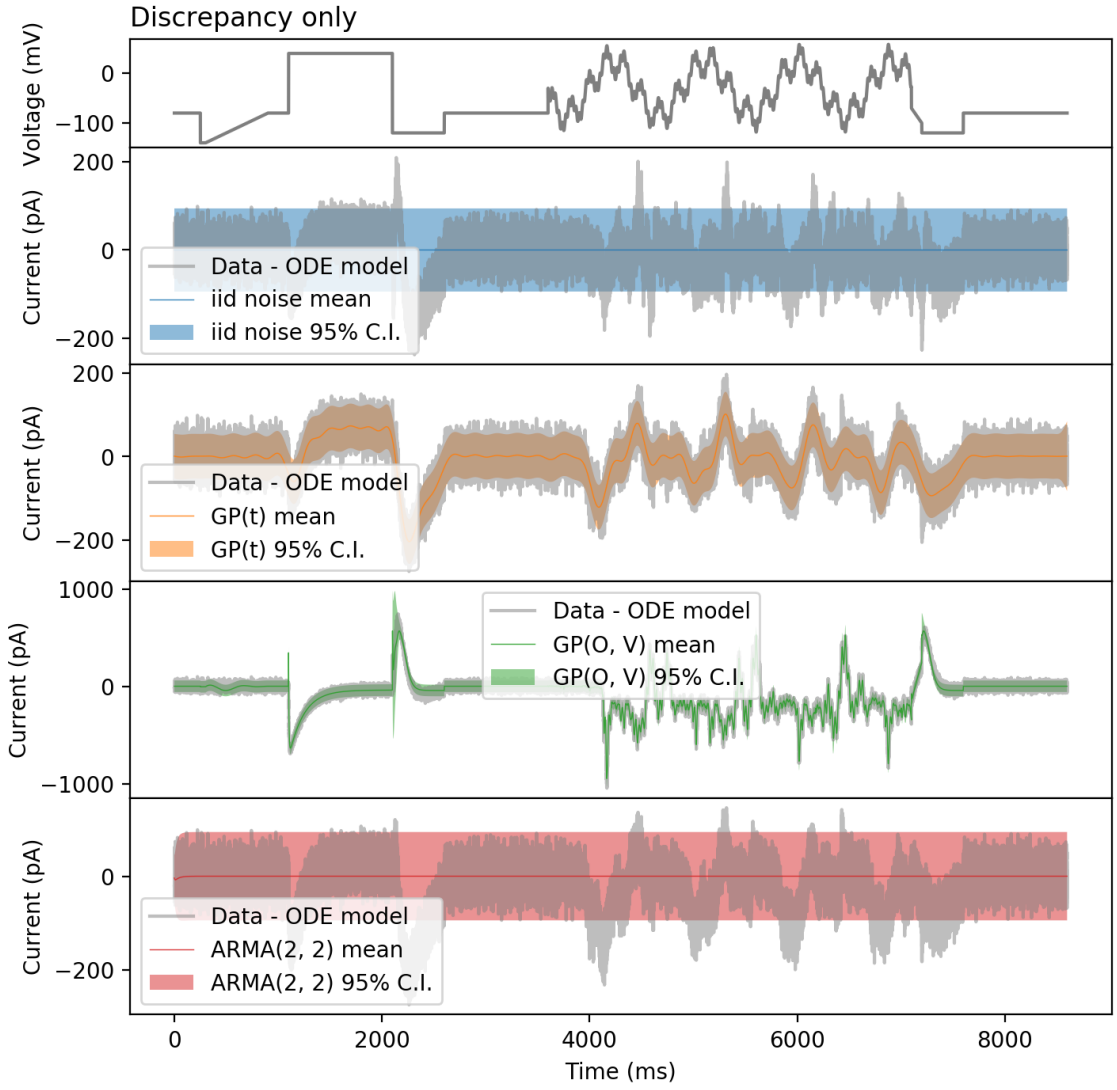


Figure S16: Model B fitting residuals of the MAP estimate accounted by different discrepancy models: no discrepancy (i.i.d. noise), GP($t$), GP($O, V$), and ARMA(2, 2). The voltage clamp protocol for calibration is the sinusoidal protocol [10].

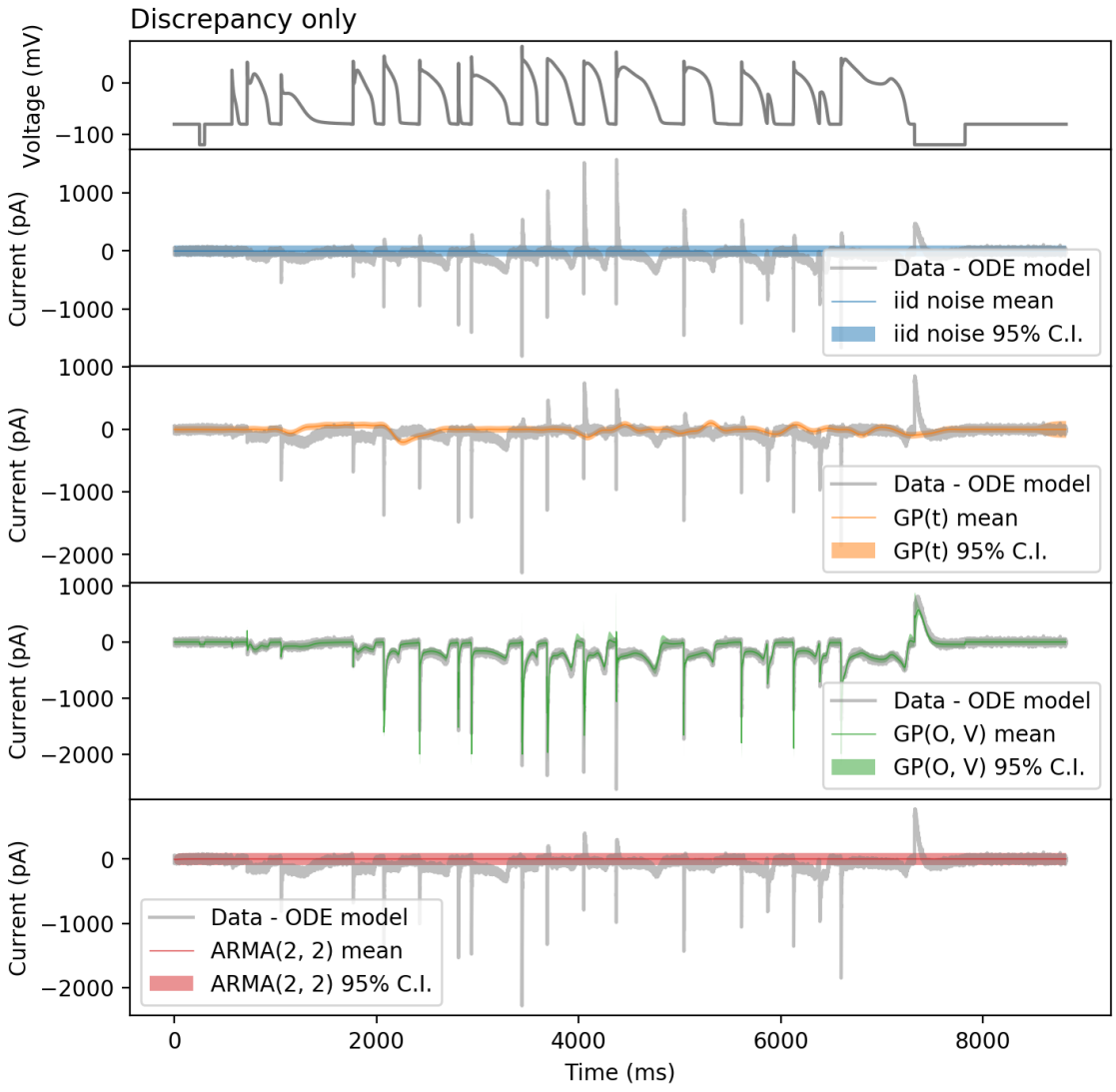


Figure S17: Model B prediction residuals of the MAP estimate accounted by different discrepancy models: no discrepancy (i.i.d. noise), GP($t$), GP($O, V$), and ARMA$(2, 2)$. The voltage clamp protocol for calibration is the action potential series protocol [10].

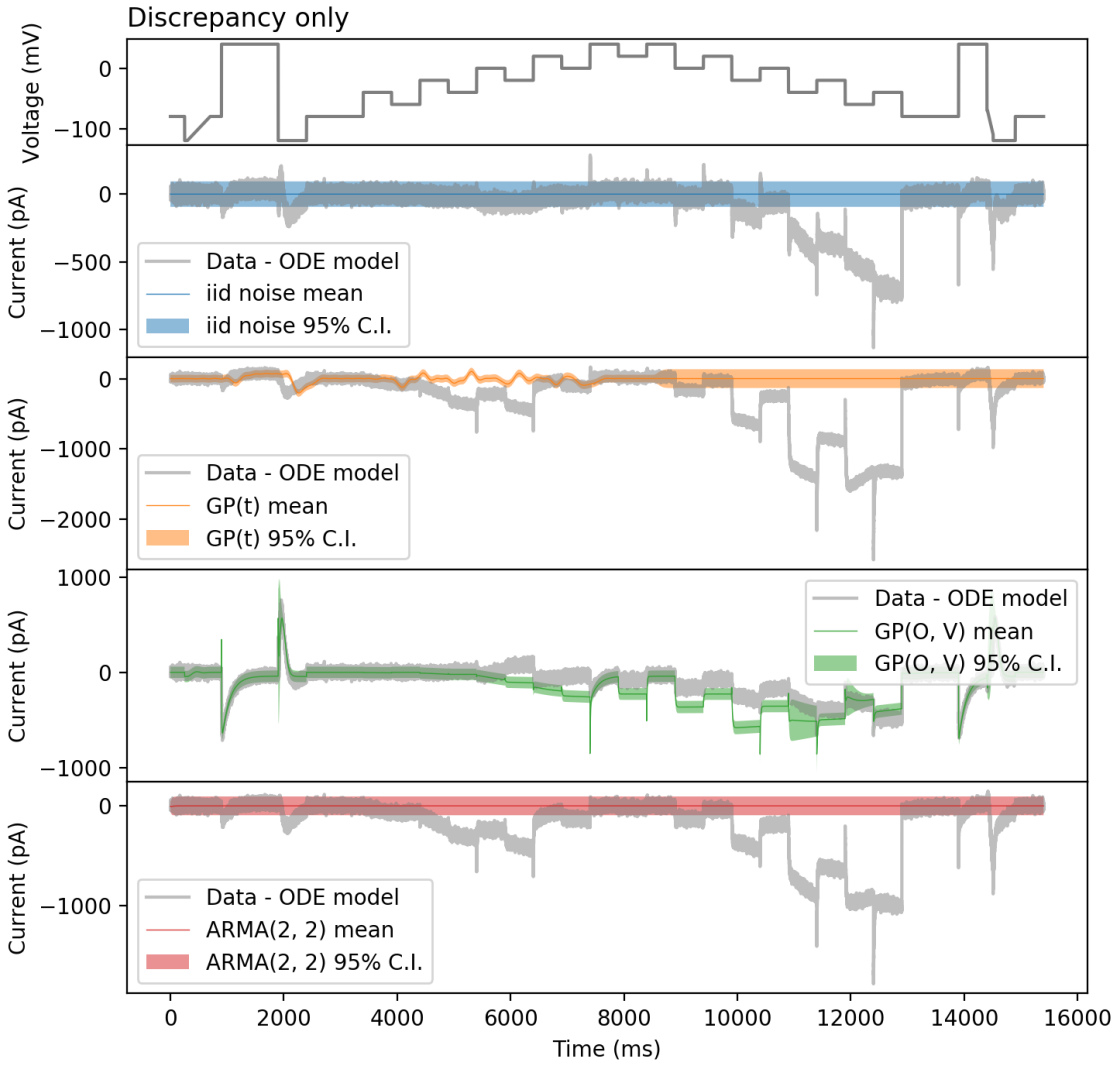


Figure S18: Model B prediction residuals of the MAP estimate accounted by different discrepancy models: no discrepancy (i.i.d. noise), GP($t$), GP($O, V$), and ARMA$(2, 2)$. The voltage clamp protocol for calibration is the staircase protocol [11].

### S7.2.3 Model B: ODE model predictions

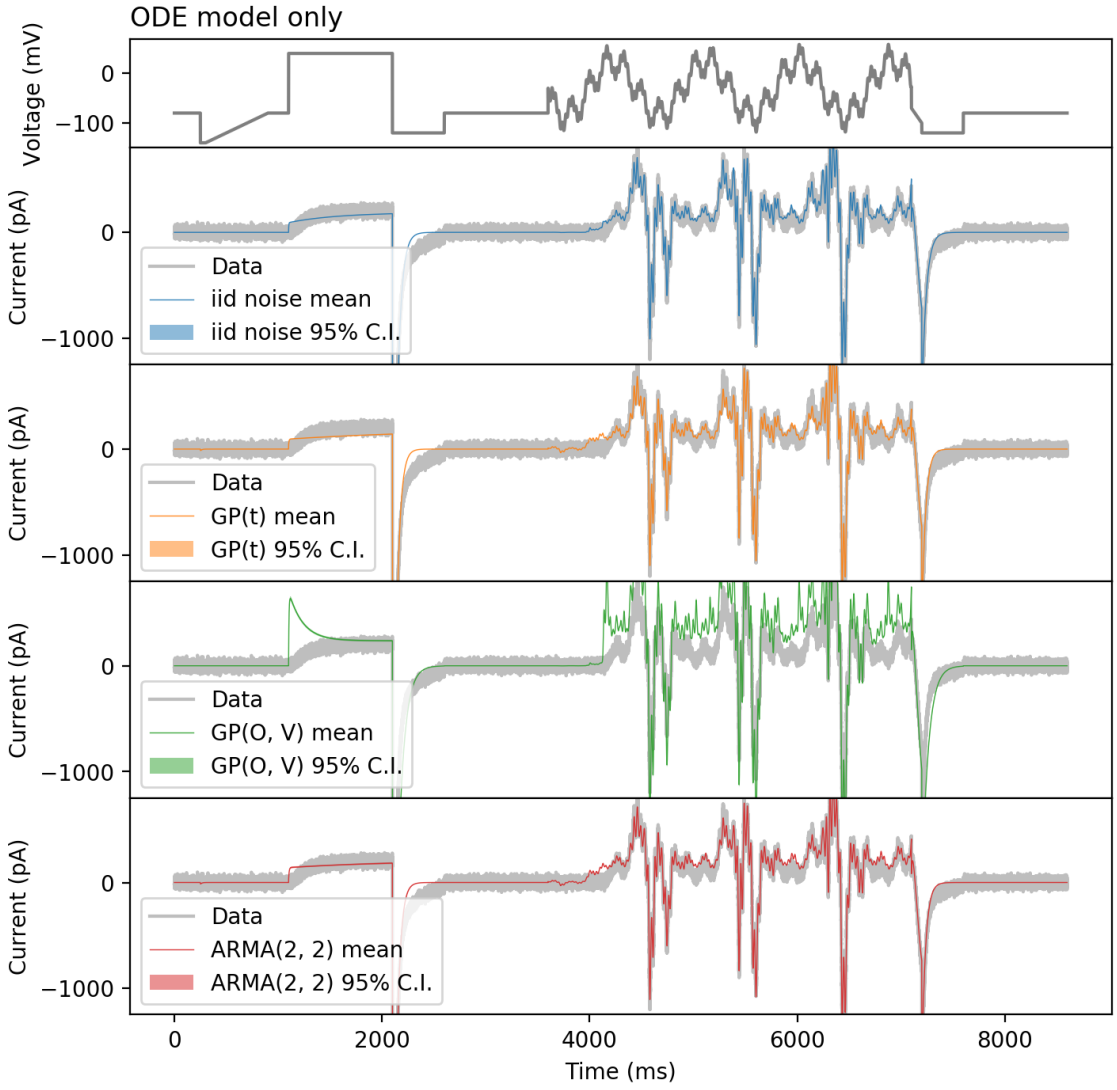


Figure S19: Fitting of the ODE model of Model B, using different discrepancy models: no discrepancy (i.i.d. noise), GP($t$), GP($O, V$), and ARMA(2, 2). The voltage clamp protocol for calibration is the sinusoidal protocol [10].

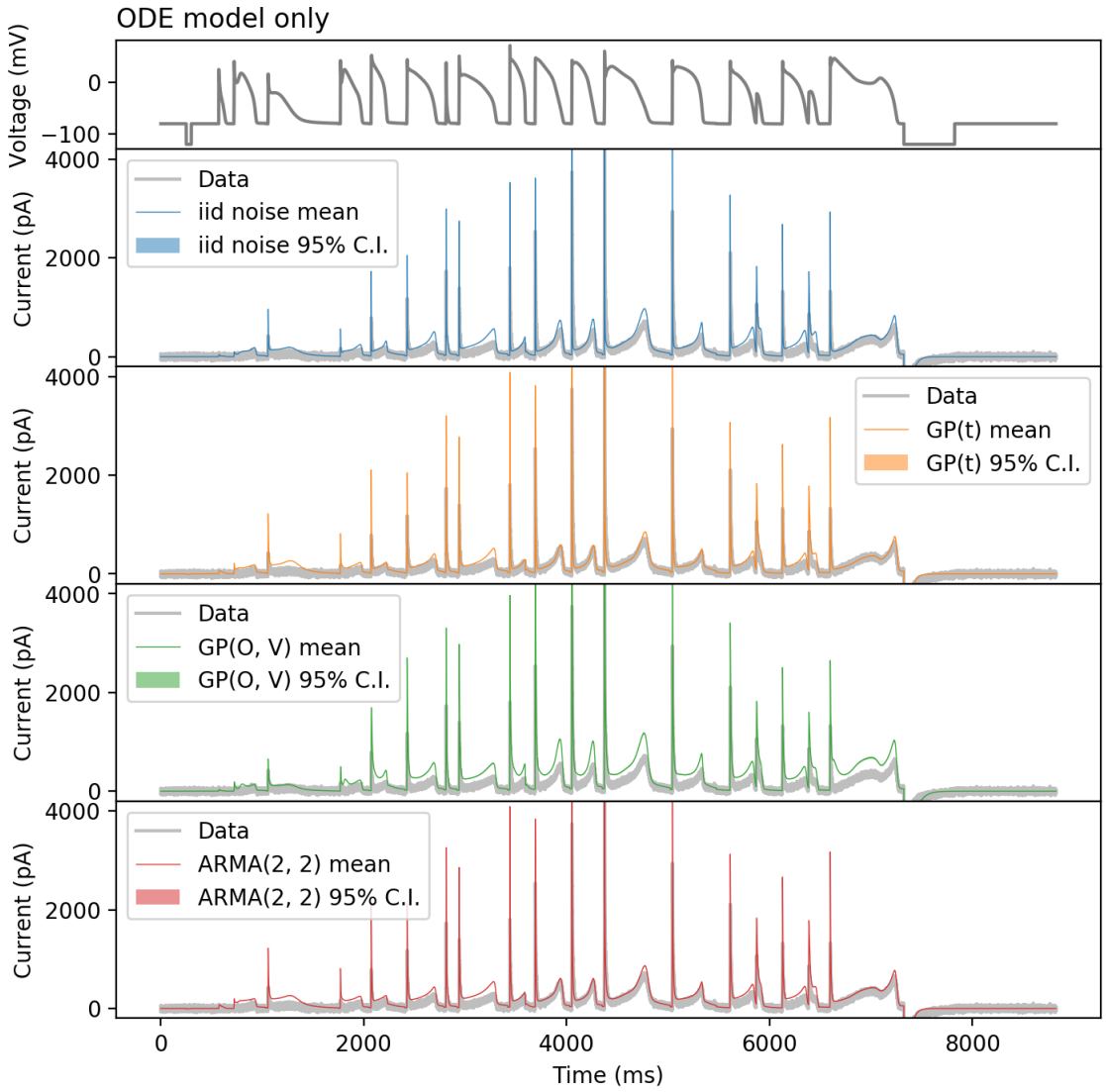


Figure S20: Predictions of the ODE model of Model B, using different discrepancy models: no discrepancy (i.i.d. noise), GP($t$), GP($O, V$), and ARMA$(2, 2)$. The voltage clamp protocol for calibration is the action potential series protocol [10].

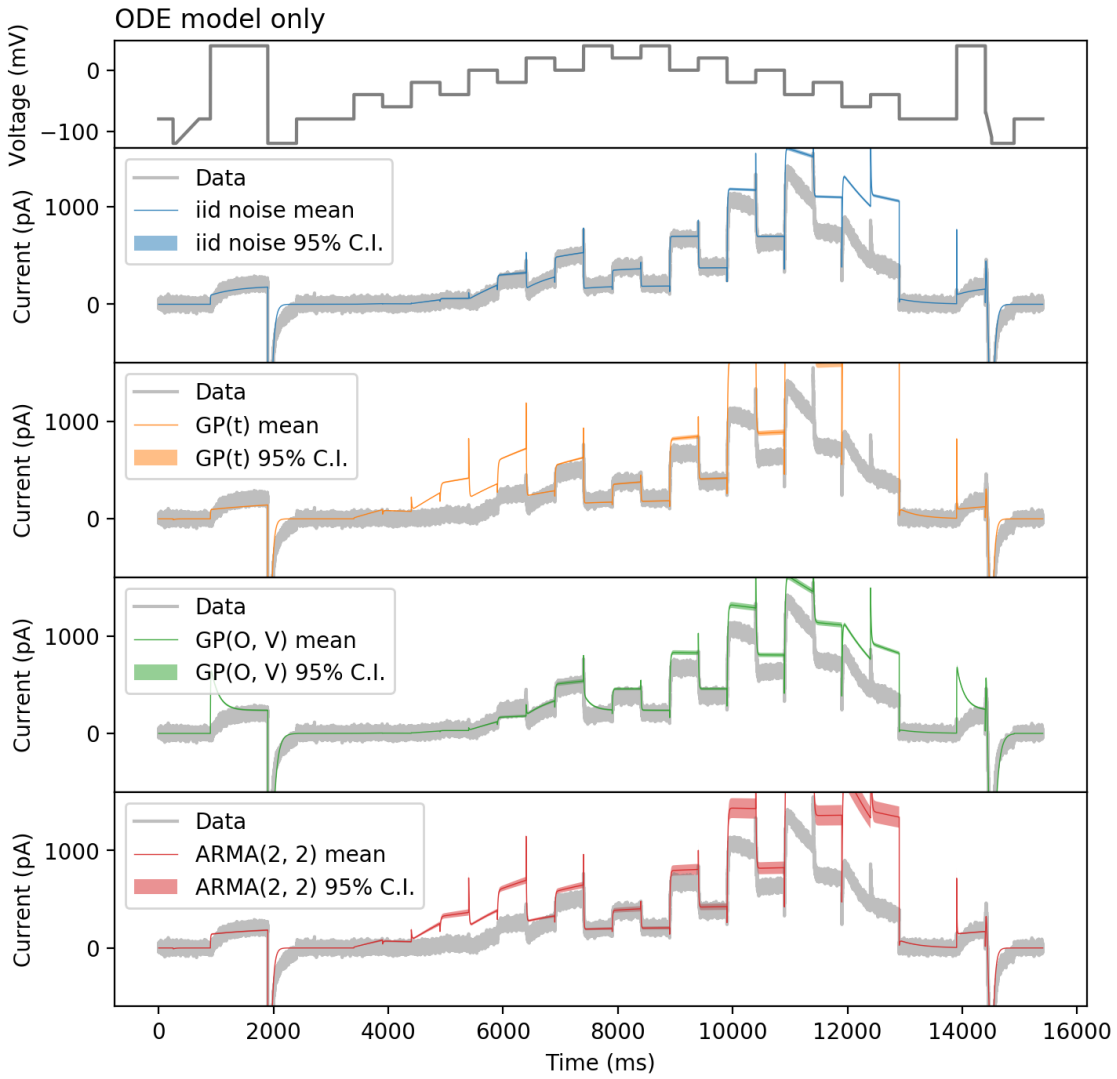


Figure S21: Predictions of the ODE model of Model B, using different discrepancy models: no discrepancy (i.i.d. noise), GP($t$), GP($O, V$), and ARMA$(2, 2)$. The voltage clamp protocol for calibration is the staircase protocol [11].

# References


[1] K. H. Ten Tusscher, D. Noble, P.-J. Noble, and A. V. Panfilov, "A model for human ventricular tissue," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 286, no. 4, pp. H1573–H1589, 2004.

[2] M. Fink, D. Noble, L. Virag, A. Varro, and W. R. Giles, "Contributions of HERG K+ current to repolarization of the human ventricular action potential," *Progress in biophysics and molecular biology*, vol. 96, no. 1-3, pp. 357–376, 2008.

[3] M. C. Kennedy and A. O'Hagan, "Bayesian calibration of computer models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 3, pp. 425–464, 2001.

[4] C. Rasmussen and C. Williams, *Gaussian processes for machine learning*. MIT Press, 2006.

[5] J. Quiñonero-Candela and C. E. Rasmussen, "A unifying view of sparse approximate gaussian process regression," *Journal of Machine Learning Research*, vol. 6, no. Dec, pp. 1939–1959, 2005.

[6] E. Snelson and Z. Ghahramani, "Sparse gaussian processes using pseudo-inputs," in *Advances in Neural Information Processing Systems*, pp. 1257–1264, 2006.

[7] J. Durbin and S. J. Koopman, *Time series analysis by state space methods*. Oxford university press, 2012.

[8] W. Wei, *Time Series Analysis: Univariate and Multivariate Methods, 2nd edition, 2006*. Pearson Addison Wesley, 2006.

[9] J. Marriott, N. Ravishanker, A. Gelfand, and J. Pai, "Bayesian analysis of arma processes: Complete sampling-based inference under exact likelihoods," *Bayesian analysis in statistics and econometrics*, pp. 243–256, 1996.

[10] K. A. Beattie, A. P. Hill, R. Bardenet, Y. Cui, J. I. Vandenberg, D. J. Gavaghan, T. P. De Boer, and G. R. Mirams, "Sinusoidal voltage protocols for rapid characterisation of ion channel kinetics," *The Journal of physiology*, vol. 596, no. 10, pp. 1813–1828, 2018.

[11] C. L. Lei, M. Clerx, D. J. Gavaghan, L. Polonchuk, G. R. Mirams, and K. Wang, "Rapid characterisation of hERG channel kinetics I: using an automated high-throughput system," *Biophysical Journal*, vol. 117, pp. 2438–2454, 2019.

[12] C. L. Lei, M. Clerx, K. A. Beattie, D. Melgari, J. C. Hancox, D. J. Gavaghan, L. Polonchuk, K. Wang, and G. R. Mirams, "Rapid characterisation of hERG channel kinetics II: temperature dependence," *Biophysical Journal*, vol. 117, pp. 2455–2470, 2019.

[13] M. Girolami, "Bayesian inference for differential equations," *Theoretical Computer Science*, vol. 408, no. 1, pp. 4–16, 2008.