Bulletin of Mathematical Biology (2024) 86:2
https://doi.org/10.1007/s11538-023-01224-6

Society for Mathematical Biology

ORIGINAL ARTICLE

Empirical Quantification of Predictive Uncertainty Due to Model Discrepancy by Training with an Ensemble of Experimental Designs: An Application to Ion Channel Kinetics

Joseph G. Shuttleworth[1] · Chon Lok Lei[2,3] · Dominic G. Whittaker[1,4] · Monique J. Windley[5,6] · Adam P. Hill[5,6] · Simon P. Preston[1] · Gary R. Mirams[1]

Received: 31 January 2023 / Accepted: 9 October 2023

Abstract

When using mathematical models to make quantitative predictions for clinical or industrial use, it is important that predictions come with a reliable estimate of their accuracy (uncertainty quantification). Because models of complex biological systems are always large simplifications, model discrepancy arises—models fail to perfectly recapitulate the true data generating process. This presents a particular challenge for making accurate predictions, and especially for accurately quantifying uncertainty in these predictions. Experimentalists and modellers must choose which experimental procedures (*protocols*) are used to produce data used to train models. We propose to characterise uncertainty owing to model discrepancy with an ensemble of parameter sets, each of which results from training to data from a different protocol. The variability in predictions from this ensemble provides an empirical estimate of predictive uncertainty owing to model discrepancy, even for unseen protocols. We use the example of electrophysiology experiments that investigate the properties of hERG

✉ Gary R. Mirams
gary.mirams@nottingham.ac.uk

1 Centre for Mathematical Medicine & Biology, School of Mathematical Sciences, University of Nottingham, University Park, Nottingham NG7 2RD, UK

2 Institute of Translational Medicine, Faculty of Health Sciences, University of Macau, Macau, China

3 Department of Biomedical Sciences, Faculty of Health Sciences, University of Macau, Macau, China

4 4 Systems Modeling & Translational Biology, Stevenage, GSK, UK

5 Computational Cardiology Laboratory, Victor Chang Cardiac Research Institute, Darlinghurst, NSW, Australia

6 School of Clinical Medicine, Faculty of Medicine and Health,, University of New South Wales, Sydney, NSW, Australia

Published online: 24 November 2023

potassium channels. Here, 'information-rich' protocols allow mathematical models to be trained using numerous short experiments performed on the same cell. In this case, we simulate data with one model and fit it with a different (discrepant) one. For any individual experimental protocol, parameter estimates vary little under repeated samples from the assumed additive independent Gaussian noise model. Yet parameter sets arising from the same model applied to different experiments conflict—highlighting model discrepancy. Our methods will help select more suitable ion channel models for future studies, and will be widely applicable to a range of biological modelling problems.

Keywords Mathematical model · Discrepancy · Misspecification · Experimental design · Ion channel · Uncertainty quantification

1 Introduction

Mathematical models are used in many areas of study to provide accurate quantitative predictions of biological phenomena. When models are used in safety-critical settings (such as drug safety or clinical decision-making), it is often important that our models produce accurate predictions over a range of scenarios, for example, for different drugs and patients. Perhaps more importantly, these models must allow a reliable quantification of confidence in their predictions. The field of *uncertainty quantification* (UQ) is dedicated to providing and communicating appropriate confidence in model predictions (Smith 2013). Exact models of biological phenomena are generally unavailable, and we resort to using approximate mathematical models instead. When our mathematical model does not fully recapitulate the data-generating process (DGP) of a real biological system, we call this *model discrepancy* or model misspecification. This discrepancy between the DGP and our models presents a particular challenge for UQ.

Often, models are trained using experimental data from a particular experimental design, and then used to make predictions under (perhaps drastically) different scenarios. We call the set of experimental designs under consideration the *design space* and denote it $\mathcal{D}$. We assume the existence of some DGP, which maps each element of $d \in \mathcal{D}$ to some random output. These elements are known as experimental designs, or, as is more common in electrophysiology, *protocols*, and each corresponds to some scenario that our model can be used to make predictions for. Namely, in Sect. 1.1, each protocol, $d \in \mathcal{D}$, is simply a set of observation times. By performing a set of experiments (each corresponding to a different protocol $d \in \mathcal{D}$) we can investigate (and quantify) the difference between the DGP and our models in different situations. When training our mathematical models using standard frequentist or Bayesian approaches, it is typically assumed that there is no model discrepancy; in other words, that the data arise from the model (for some unknown, true parameter set). This is a necessary condition for some desirable properties of the parameter estimators which provide some guarantees regarding the accuracy of parameter estimates when there is a large number of observations, as discussed in Sect. 2.1.3. However, when model discrepancy is not considered, we can find that the ability of a model to make accurate predictions is compromised. In particular, if we try to validate our model with a protocol dissimilar

to that used for training, there can be a noticeable difference between our predictions and the data—even when the model appears to fit the training data well. A simple illustration of this problem is introduced in the following section.

1.1 Motivating Example

In this section, we construct a simple example where we train a discrepant model with data generated from a DGP using multiple experimental designs. This example demonstrates that it is important to consider the protocol-dependence of parameter estimates and predictions when using discrepant models.

First, we construct a DGP formed of the sum of two exponential terms,

$$y^*(t) = \exp\{-t\} + \exp\left\{-\frac{t}{10}\right\}, \tag{1}$$

$$z^*(t) = y^*(t) + \varepsilon(t), \tag{2}$$

for some $t > 0$ where $\varepsilon(t)$ is an independent Gaussian random variable, each with zero mean and variance, $\sigma^2 = 10^{-4}$ for each $t > 0$. Here, $z^*(t)$ is a random variable representing an observation of the system at some time, t.

Next, we attempt to fit a model which takes the form of single exponential decay,

$$y(t;\boldsymbol{\theta}) = \theta_1 \exp\left\{-\frac{t_i}{\theta_2}\right\}, \tag{3}$$

$$z(t;\boldsymbol{\theta}) = y(t;\boldsymbol{\theta}) + \varepsilon(t_i), \tag{4}$$

to these data, denoting the column matrix $[\theta_1, \theta_2]^T$ by $\boldsymbol{\theta}$. We call this a discrepant model because there is no choice of $\boldsymbol{\theta}$ such that $y(t;\boldsymbol{\theta}) = y^*(t)$, for all $t > 0$.

To train our model, we choose a set of n observation times, $T = \{t_1, t_2, \ldots t_n\}$. We may then find the parameter set, $\hat{\boldsymbol{\theta}}(T)$, which minimises the sum-of-squares error between our discrepant model (Eq. 4) and each $z(\boldsymbol{\theta}; t_i)$, that is,

$$\hat{\boldsymbol{\theta}}(T) = \text{argmin}_{\boldsymbol{\theta}\in\Theta}\left\{\sum_{t_i\in T}\left(y(t_i;\boldsymbol{\theta}) - z^*(t_i)\right)^2\right\}, \tag{5}$$

where T is a set of observation times.

Then, we consider multiple experimental protocols which we may use to fit this model (Eq. (4)). In particular, we consider the following sets of observation times

$$T_1 = \{0,\ 0.01,\ 0.02,\ \ldots, 0.01\}, \tag{6}$$

$$T_2 = \{0,\ 0.1,\ 0.2,\ 0.3,\ \ldots,\ 1\}, \tag{7}$$

$$T_3 = \{0.2,\ 0.3,\ 0.4,\ 0.5,\ \ldots,\ 1.2\}, \tag{8}$$

$$T_4 = \{0.5,\ 0.55,\ 0.6,\ \ldots,\ 1\}, \text{ and} \tag{9}$$

$$T_{\text{all}} = T_1 \cup T_2 \cup T_3 \cup T_4. \tag{10}$$

We sample from the DGP 10 times by computing Eq. 2 for each observation time, t, and adding IID Gaussian noise. Then, for each sample of the DGP, we compute parameter estimates using each set of observation times (T_1, T_2, T_3, T_4 and T_{all}). This process is then repeated with a ten-fold increase in sampling rate, that is, with observation times,

$$T_1' = \{0,\ 0.001,\ 0.002,\ \ldots,\ 0.01\}, \tag{11}$$
$$T_2' = \{0,\ 0.01,\ 0.02,\ 0.03,\ \ldots,\ 1\}, \tag{12}$$
$$T_3' = \{0.2,\ 0.21,\ 0.22,\ 0.23,\ \ldots,\ 1.2\}, \tag{13}$$
$$T_4' = \{0.5,\ 0.505,\ 0.51,\ \ldots,\ 1\},\ \text{and} \tag{14}$$
$$T_{\text{all}}' = T_1 \cup T_2 \cup T_3 \cup T_4. \tag{15}$$

If we choose a Bayesian approach to the problem, we may specify a (relatively uninformative) uniform prior distribution on the model parameters, that is,

$$\theta_1 \sim U(0, 10), \tag{16}$$
$$\text{and}\ \ \theta_2 \sim U(0, 10). \tag{17}$$

The likelihood of our misspecified model is

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{z}^*) = \prod_{i=1}^{n} \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left\{ \frac{(z^*(t_i) - y(t_i; \boldsymbol{\theta}))^2}{2\sigma^2} \right\}, \tag{18}$$

where n is the number of observations for this protocol, and $\mathbf{z}^* = (z(t_i))_{i=1}^n$ is a vector of observations of the DGP. We may then explore the posterior distribution (Gelman et al. 2013) using Markov chain Monte Carlo. In particular we use the Haario-Bardenet adaptive-covariance Metropolis-Hastings (Johnstone et al. 2016) algorithm as implemented by PINTS (Clerx et al. 2019b). This method was run using four parallel chains each containing 25,000 iterations and a 'burn in' period of 5,000 iterations.

As shown in Fig. 1, for each sample of the DGP, we obtain a parameter estimate from each set of observation times, each with a different distribution. For instance, training using T_1 results in a model that approximates the DGP well on short timescales, and training using T_4 allows us to recapitulate the behaviour of the system over longer timescales, as can be seen in panel **a**. From how closely the discrepant model (Eq. 4) fits the data in the regions where observations are made (in Fig. 1, panels a and b), we can see that in either case, a single exponential seems to provide a reasonable approximation to the DGP. However, if we require an accurate model for both the slow and fast behaviour of the system, model discrepancy presents an issue, and this model (namely, Eq. 4) may be unsuitable. This is the case for T_2 as shown in Fig. 1a. This variability in behaviour is shown in Fig. 1, panels **g** and **h**, which show how the model's predictions for $0 \leq t \leq 2$ depend on what protocol was used to fit the model.

The Bayesian posteriors illustrated in Fig. 1, panels **e** and **f**, show that we are not able to avoid the problems caused by model discrepancy by simply adopting a Bayesian framework—we will obtain precise parameter estimates that are highly dependent on the chosen training protocol, nevertheless. This problem becomes more

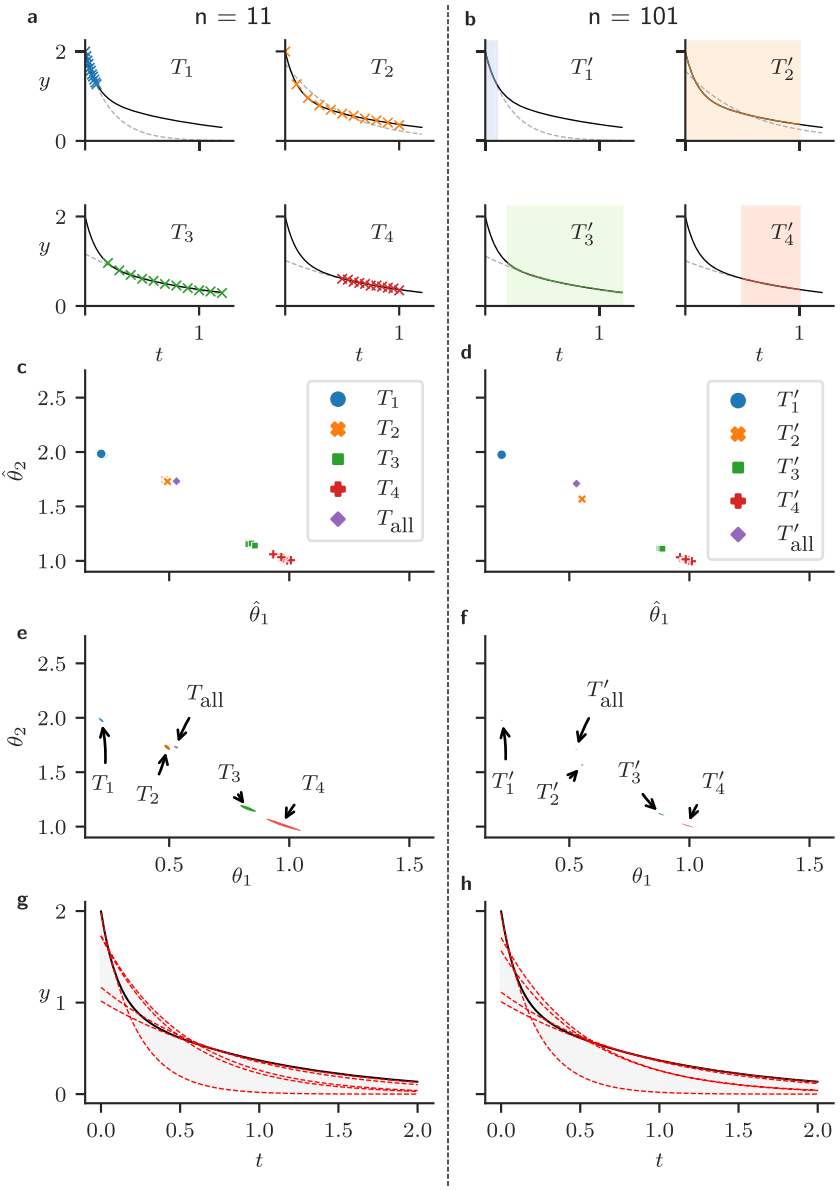

Fig. 1 (Color figure online) Under model discrepancy, parameter estimates depend on the design used for training. Panels to the left of the dotted line correspond to designs containing $n = 11$ observations at times $(T_1, \ldots, T_4$ as shown in panel (**a**). Panels on the right show designs with $n = 101$ observations, $(T'_1, \ldots, T'_4$ as shown in panel **b**). **a** and **b** representative datasets generated by the DGP shown with the solid black line (Eq. 1) with points indicating observations (sampled using Eq. 2) and the fitted discrepant model (Eq. 4), with calibrated $\boldsymbol{\theta}$) (grey dashed lines). **c** and **d** The parameter estimates for each design, each fitted to one of ten repeats of the DGP. **e** and **f** 99% Bayesian credible regions obtained using MCMC, a uniform prior and a single repeat of the DGP. **g** and **h** Predictions using the discrepant model fitted using a single repeat of each protocol (using the estimates shown in **e** and **f**), showing the true DGP (black), discrepant model predictions (red), and the difference between predictions (grey)

obvious when we increase the number of observations. In the examples detailed in this paper, we explore this 'high-data limit' where the variability in each parameter estimate (under repeated samples of the DGP) is minuscule compared to the difference between parameter estimates obtained from different protocols.

1.2 Ion Channel Modelling

Discrepancy has proven to be a difficult problem to address in modelling electrically-excitable cells (*electrophysiology* modelling, Lei et al. 2020b; Mirams et al. 2016). The same is true for many other mathematical models, such as rainfall-runoff models in hydrology (Beven 2006), models of the spread of infectious diseases in epidemiology (Guan et al. 2020; Creswell et al. 2023), and models used for the prediction of financial markets (Anderson et al. 2009).

The 'rapid delayed rectifier potassium current' (I_{Kr}), carried by the channel encoded primarily by hERG, plays an important role in the recovery of heart cells in from electrical stimulation. It allows the cell membrane to return to its 'resting potential' ahead of the next heartbeat. This current can be blocked by pharmaceutical drugs, disrupting this process and causing dangerous changes to heart rhythm. Mathematical models are now routinely used in drug safety assessment to test whether the expected dynamics of I_{Kr} under drug block are expected to cause such problems (Li et al. 2017). However, these models provide only an incomplete description of I_{Kr}, and do not, for example, account for the stochastic behaviour of individual ion channels (Mirams et al. 2016). For this reason, an understanding of model discrepancy and its implications is crucial in building accurate dynamical models of I_{Kr} which permit a realistic appraisal of their predictive uncertainty.

In Sect. 1.1, we presented a simple example, in which each protocol corresponds to a particular choice of observation times. However, there may other aspects of the design to be decided upon. For example, in electrophysiology, whole-cell patch-clamp experiments are performed by placing an electrode in the solution inside the cell membrane (the intracellular solution), and another in the solution outside the cell (the extracellular solution). Voltage-clamp experiments are a particular type of patch-clamp experiment in which a voltage signal is applied across the cell membrane, whilst the current flowing across the cell membrane is recorded. Here, the protocol consists of the chosen voltage for each time (treated as a forcing function in ODE-based models), together with a set of observation times for the resulting current (observed output).

Electrophysiologists have a lot of control, and therefore choice, regarding the protocol design; but little work has been done to explore how the choice of protocol used to gather training data affects the accuracy of subsequent predictions. We explore these protocol-dependent effects of model discrepancy in Sect. 3.

1.3 Previously Proposed Methods to Handle Discrepancy

One way of reducing over-confidence in inaccurate parameter estimates in the presence of model discrepancy may be to use approximate Bayesian computation (ABC) (Frazier et al. 2020). With ABC, a likelihood function is not explicitly specified;

instead, the model is repeatedly simulated for proposed values of the parameter sampled from a prior distribution. Each proposed value is accepted or rejected according to whether the simulated trajectory is "close" to the actual data, according to some chosen summary statistics. ABC compares the simulated with the real data using these summary statistics (rather than matching all aspects of the dynamics) and accepts approximate matches (subject to a chosen tolerance). It is suited to inference where there is substantial model discrepancy because this approach can decrease potential over-confidence in the inferred values of parameters. However, it is challenging to select suitable summary statistics, and the computational demands of ABC are much greater than those of the methods we propose.

Another approach was first introduced by Kennedy and O'Hagan (2001), who introduced Gaussian processes to the observables. This work has since been applied to electrophysiology models (Lei et al. 2020b). Elsewhere, Sung et al. introduced an approach to account for heteroscedastic errors using many repeats of the same experiment (Sung et al. 2020), although this seems to be less applicable to the hERG modelling problem (introduced in Sect. 2.2) because the number of repeats of each experiment (when training individual, cell-specific models) is limited. Alternatively, Lei and Mirams (2021) modelled the discrepancy using a neural network within the differential equations. However, these approaches reduce the interpretability of otherwise simple mechanistic models, and, when compared with models that simply ignore model discrepancy, could potentially result in worse predictions under protocols that are dissimilar to those used for training.

Instead, we use a diverse range of experiments to train our models and build a picture of how model discrepancy manifests under different training protocols. We are then able to judge the suitability of our models, and provide empirically-derived, *spread-of-prediction* intervals which provide a realistic level of predictive uncertainty due to model discrepancy. We demonstrate the utility of these methods under synthetically generated data by constructing two examples of model discrepancy.

2 Methods

We begin with a general overview of our proposed methods before providing two real-world examples of their applications. In Sect. 2.1.1, we outline some notation for a statistical model consisting of a dynamical system, an observation function, and some form of observational noise. This allows us to talk, in general terms, about model calibration and validation in Sect. 2.1.2. In particular, we describe a method for validating our models, in which we change the protocol used to train the model. This motivates our proposed methods for combining parameter estimates obtained from different protocols to empirically quantify model discrepancy for the prediction of unseen protocols.

2.1 Fitting Models Using Multiple Experimental Protocols

2.1.1 Partially Observable ODE Models

In this paper, we restrict attention to deterministic models of biological phenomena, in which a system of *ordinary differential equations* (ODEs) is used to describe the deterministic time-evolution of some finite number of states. Although, the method would generalise to other types of models straightforwardly. This behaviour may be dependent on the protocol, d, chosen for the experiment, and so, we express our ODE system as,

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}, t; \boldsymbol{\theta}_f, d), \tag{19}$$

where $\mathbf{x}$ is a column vector of length N describing the 'state' of the system, t is time, and the parameters specifying the dynamics of the system are denoted $\boldsymbol{\theta}_f$. Additionally, the system is subject to some initial conditions which may be dependent on $\boldsymbol{\theta}_f$. Owing to $\mathbf{x}$'s dependence on the protocol and model parameters, we use the notation,

$$\mathbf{x}(t; \boldsymbol{\theta}_f, d), \tag{20}$$

to denote the solution of Eq. 19 under protocol d and a specific choice of parameters, $\boldsymbol{\theta}_f$.

This ODE system is related to our noise-free observables via some *observation function* of the form,

$$h\left(\mathbf{x}, t; \boldsymbol{\theta}_h, d\right), \tag{21}$$

where $\mathbf{x}$ is the state of the ODE system (Eq. 19), t is the time that the observation occurs, d is the protocol, and some additional parameters $\boldsymbol{\theta}_h$, which are distinct from those in $\boldsymbol{\theta}_f$. Here, we make observations of the system, via this function, at a set of observation times, $\{t_i\}_{i=1}^{n_d}$ defined by the protocol, d.

For concision, we may stack $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_h$ into a single vector of model parameters,

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_f \\ \boldsymbol{\theta}_h \end{bmatrix}. \tag{22}$$

Then, we denote an observation at time t_i by

$$y_i(\boldsymbol{\theta}; d) = h\big(\mathbf{x}(t_i; \boldsymbol{\theta}_f, d), t_i; \boldsymbol{\theta}_h, d\big), \tag{23}$$

We denote the set of possible model parameters by Θ, such that $\boldsymbol{\theta} \in \Theta$. We call this collection of possible parameter sets the *parameter space*.

For each protocol, $d \in \mathcal{D}$, and vector of model parameters, $\boldsymbol{\theta}$, we may combine each of our observations into a vector,

$$\mathbf{y}(\boldsymbol{\theta}; d) = \begin{bmatrix} y_1(\boldsymbol{\theta}; d), \\ \vdots \\ y_{n_d}(\boldsymbol{\theta}; d) \end{bmatrix}. \tag{24}$$

Additionally, we assume some form of random observational error such that, for each protocol, d, each observation is a random variable,

$$z_i(d) = y_i(\boldsymbol{\theta}; d) + \varepsilon_i, \tag{25}$$

where each ε_i is the error in the i^{th} observation. Here each protocol, d, is performed exactly once so that we obtain one sample of each vector of observations (the vector $\mathbf{z}(d) = [z_1(d), \ldots, z_{n_d}(d)]$). In the examples presented in Sects. 3.1 and 3.2, we assume that our observations are subject to independent and identically distributed (IID) Gaussian errors, with zero mean 0 and standard deviation, σ.

2.1.2 Evaluation of Predictive Accuracy and Model Training

Given some parameter set $\boldsymbol{\theta}$, we may evaluate the accuracy of the resultant predictions under the application of some protocol $d \in \mathcal{D}$ by computing the root-mean-square error (RMSE) (Willmott et al. 1985) between these predictions, and our observations ($\mathbf{z}(d)$),

$$\text{RMSE}\big(\mathbf{y}(\boldsymbol{\theta}; d),\ \mathbf{z}(d)\big) = \sqrt{\frac{1}{n_d}\sum_{i=1}^{n_d}\big(y_i(\boldsymbol{\theta}; d) - z_i(d)\big)^2}, \tag{26}$$

where n_d is the number of observations in protocol d. We choose the RMSE as it permits comparison between protocols with different numbers of observations.

Similarly, we may train our models to data, $\mathbf{z}(d)$, obtained using some protocol, d, by finding the parameter set that minimises this quantity (Eq. 26). In this way, we define the parameter estimate obtained from protocol d as,

$$\hat{\boldsymbol{\theta}}_d = \text{argmin}_{\boldsymbol{\theta}\in\Theta}\left\{\text{RMSE}\big(\mathbf{y}(\boldsymbol{\theta}, d),\ \mathbf{z}(d)\big)\right\}, \tag{27}$$

which is a random variable (because it depends on our random data, $\mathbf{z}$). Since minimising the RMSE is equivalent to minimising the sum-of-squares error, this estimate is also the least-squares estimator (identical to Eq. 5). Moreover, under the assumption of Gaussian IID errors, Eq. 27 is exactly the *maximum likelihood estimator* because the natural logarithm of the likelihood can be written as,

$$\log\{\mathcal{L}(\boldsymbol{\theta}; \mathbf{z})\} = -\frac{n}{2}\log\left(2\pi\hat{\sigma}^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\big(y_i(\boldsymbol{\theta}; d) - z_i(d)\big)^2, \tag{28}$$

where $\hat{\sigma}$ is an estimate of σ. Equation 28 can be minimised by first finding the parameter set, $\hat{\boldsymbol{\theta}}$ which minimises the sum-of-squares error term, then finding the optimal σ. Whilst these estimates of $\boldsymbol{\theta}$ are identical whether or not σ is known, only examples with known (and not estimated) σ are explored in this paper.

Having obtained such a parameter estimate, we may validate our model, by computing predictions for some other protocol, $\tilde{d} \in \mathcal{D}$. To do this, we compute, $\mathbf{y}(\hat{\boldsymbol{\theta}}_d; \tilde{d})$. This is a simulation of the behaviour of the system (without noise) under protocol $\tilde{d}$ made using parameter estimates that were obtained by training the model to protocol d (as in Eq. 27). In this way, our parameter estimates, each obtained from different protocols, result in different out-of-sample predictions (predictions for the results for protocols other than the one used for training). Because we aim to train a model able to produce accurate predictions for all $d \in \mathcal{D}$, it is important to validate our model using multiple protocols.

By computing $\text{RMSE}\big(\mathbf{y}(\hat{\boldsymbol{\theta}}_d; \tilde{d}), \mathbf{z}(\tilde{d})\big)$ for each pair of training and validation protocols, d and $\tilde{d}$, we are able to perform model validation across multiple training and validation protocols. This allows us to ensure our models are robust with regard to the training protocol, and allow for the quantification of model discrepancy as demonstrated in Sect. 3.

2.1.3 Consequences of Model Error/Discrepancy

Ideally, we would have a model that is correctly specified, in the sense that the data arise from the model being fitted. In other words, our observations $\mathbf{z}(d)$ arise from Eq. 25 where $\boldsymbol{\theta}$ is some fixed, unknown value, $\boldsymbol{\theta}^* \in \Theta$. Then we may consider the distance between an estimate $\hat{\boldsymbol{\theta}}$ and the true value. When the model is correctly specified and given suitable regularity conditions on the model and the design, d, we can obtain arbitrarily accurate parameter estimates by increasing the number of observations, n. That is, more precisely, that $\hat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}^*$ as $n \to \infty$. This property is known as consistency (Seber and Wild 2005). These regularity conditions include that the model is structurally identifiable for the particular $d \in \mathcal{D}$ used for training, for example. That is, different values of the parameter, $\boldsymbol{\theta}$, result in different model output (Wieland et al. 2021). Other conditions ensure that $\mathbf{y}(\boldsymbol{\theta}; d)$ is suitably smooth as a function of $\boldsymbol{\theta}$ (Seber and Wild 2005).

However, when training discrepant models, we may find that our parameter estimates are heavily dependent on the training protocol, as demonstrated in Sect. 1.1. For unseen protocols, these discordant parameter sets may lead to a range of vastly different predictions, even if each parameter set provides a reasonable fit for its respective training protocol. In such a case, further data collection may reduce the variance of these parameter estimates, but fail to significantly improve the predictive accuracy of our models.

In Sect. 3, we explore two examples of synthetically constructed model discrepancy. In Sect. 3.1, we have that $\mathbf{f}$ and h (Eqs. 19 and 25) are exactly those functions used to generate the data, and the exact probability distribution of the observational errors is known. However, one parameter is fixed to an incorrect value. In other words, the true parameter set $\boldsymbol{\theta}^*$ lies outside the parameter space used in training the model. Under

the assumption of structural identifiability (and a compact parameter space), this is an example of model discrepancy because there is some limit to how well our model can recapitulate the DGP.

In Sect. 3.2 we explore another example of model discrepancy where our choice of $\mathbf{f}$ (and, in this case, the dimensions of $\boldsymbol{\theta}$ and $\mathbf{x}$) are misspecified by training a model which differs structurally from the one used in the DGP.

2.1.4 Ensemble Training and Prediction Interval

As outlined in Sect. 2.1.2, we can obtain parameter estimates from each protocol $d \in \mathcal{D}$ by finding the $\hat{\boldsymbol{\theta}} \in \Theta$ that minimises Eq. 27. We then obtain an *ensemble* of parameter estimates,

$$\left\{ \hat{\boldsymbol{\theta}}_d \ : \ d \in \mathcal{D}_{\text{train}} \right\}. \tag{29}$$

Then, for any validation protocol $\tilde{d}$, the set,

$$\left\{ \mathbf{y}(\hat{\boldsymbol{\theta}}_d; \tilde{d}) \ : \ d \in \mathcal{D}_{\text{train}} \right\}, \tag{30}$$

is an ensemble of predictions where $\mathcal{D}_{\text{train}} \subseteq \mathcal{D}$ is some set of training protocols. Each of these estimates may be used individually to make predictions. We may then use these ensembles of parameter estimates to attempt to quantify uncertainty in our prediction. We do this by considering the range of our predictions for each observation of interest. For the ith observation of our validation protocol, $\tilde{d}$, that is

$$\begin{aligned} \mathcal{B}^{(i)} &= \left[\mathcal{B}^{(i)}_{\text{lower}}, \mathcal{B}^{(i)}_{\text{upper}} \right] \\ &= \left[\min_{d \in \mathcal{D}_{\text{train}}} \left\{ y_i(\hat{\boldsymbol{\theta}}_d; \tilde{d}) \right\}, \max_{d \in \mathcal{D}_{\text{train}}} \left\{ y_i(\hat{\boldsymbol{\theta}}_d; \tilde{d}) \right\} \right], \end{aligned} \tag{31}$$

When all observations are considered at once, Eq. 31 comprises a band of predictions, giving some indication of uncertainty in the predictions. We demonstrate below that this band provides a useful indication of predictive error for unseen protocols, and provides a range of plausible predictions. We propose that a wide band of predictions for a given validation protocol suggests that there is model discrepancy and poor prediction accuracy for a particular context of use.

This interval (Eq. 31), cannot shrink as more protocols are added. If a large number of protocols are considered, percentiles of our ensemble of predictions may provide additional insight. However, in this paper, we only consider cases where there are a small number of protocols (five training protocols are used in each of the examples discussed in Sect. 3).

For the purposes of a point estimate, we use the midpoint of each interval,

$$\mathcal{B}^{(i)}_{\text{mid}} = \frac{\mathcal{B}^{(i)}_{\text{lower}} + \mathcal{B}^{(i)}_{\text{upper}}}{2}. \tag{32}$$

This is used to assess the predictive error of the ensemble in Fig. 8. There are other ways to gauge the central tendency of the set of predictions (Eq. 30). Such a change would have little effect on Sect. 3, but a median or weighted mean may be as (or more) suitable for other problems.

2.2 Application to an Ion Current Model

We now turn our attention to an applied problem in which dynamical systems are used to model cellular electrophysiology. We apply our methods to two special cases of model discrepancy using synthetically generated data.

Firstly, we introduce a common paradigm for modelling macroscopic currents in electrically excitable cells, so-called *Markov models* (Rudy and Silva 2006; Fink and Noble 2009). In this setting, the term 'Markov model' is often used to refer to systems of ODEs where the state variables describe the proportional occupancy of some small collection of 'states', and the model parameters affect transition rates between these states. These models are discussed in Sect. 2.2.1 and may be seen as a special case of the more general ODE model introduced in Sect. 2.1. Additionally, in Sect. 2.2.2, we briefly introduce some relevant electrophysiology and in Sect. 2.2.3, we provide a detailed overview of our computational methods.

2.2.1 Markov Models of I_{Kr}

Here, we use Markov models to describe the dynamics of I_{Kr}, especially in response to changes in the transmembrane potential. For any Markov model (as described above), the derivative function can be expressed in terms of a matrix, $\mathbf{A}$, which is dependent only on the transmembrane potential, V. Accordingly, where x_i denotes the proportion of channels in some state, i, Eq. 19 becomes,

$$\begin{aligned} \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} &= \mathbf{f}(\mathbf{x}, t; \boldsymbol{\theta}_f, d), \\ &= \mathbf{A}\left(V(t; d); \boldsymbol{\theta}_f\right)\mathbf{x}, \end{aligned} \tag{33}$$

where $V(t; d)$ is the specified transmembrane potential at the time t under protocol d. The elements of $\mathbf{A}(V; \boldsymbol{\theta}_f)$, that is, $\mathbf{A}_{i,j}(V; \boldsymbol{\theta}_f)$ describe the *transition rate* from state j to state i with transmembrane potential, V. Usually, the transition rates (elements of $\mathbf{A}$) are either constant or of the form $\theta_i e^{\pm\theta_j V(t;d)}$ with $\theta_i, \theta_j > 0$. Hence, each transition rate, k is either 0 for all $V \in \mathbb{R}$ or satisfies $k > 0$ for all $V \in \mathbb{R}$.

Before and after each protocol, cells are left to equilibrate with the voltage V set to the *holding potential*, $V_{\text{hold}} = -80\,\text{mV}$. Therefore, we require the initial conditions, for at time $t = 0$,

$$\mathbf{x}(0) = \mathbf{x}_\infty(V_{\text{hold}}), \tag{34}$$

where, $\mathbf{x}_\infty(V_{\text{hold}})$ is the unique steady-state solution for the linear system,

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{A}(V_{\text{hold}}; \boldsymbol{\theta}_f)\mathbf{x}, \tag{35}$$

subject to the constraint $\sum_{i=1}^{N} x_i(0) = 1$. Note also that $\mathbf{A}(V_{\text{hold}}; \boldsymbol{\theta}_f)$ may be singular, as is the case when the number of channels is conserved ($\sum_{i=1}^{N} x_i(t) = 1$ for all t). This is the case for both Markov models used in this paper. To find $\mathbf{x}_\infty$, we may follow Fink and Noble's method (see Supplementary Material of Fink and Noble 2009). A more technical discussion of the steady states of such ODE systems is found in Keizer (1972).

As is standard for models of I_{Kr} (Beattie 2015), we take our observation function to be

$$I_{\text{Kr}} = h(\mathbf{x}, t_i; \theta_h, d) = g \cdot [O](t; \boldsymbol{\theta}_f, d) \cdot (V(t; d) - E_{\text{Kr}}), \tag{36}$$

where $[O]$ denotes the proportion of channels in an 'open' conformation (one of the components of $\mathbf{x}$); and g is the sole parameter in $\boldsymbol{\theta}_h$, known as the *maximal conductance*; and E_{Kr} is the Nernst potential. E_{Kr} is found by calculating

$$E_{\text{Kr}} = \frac{RT}{F} \ln\left\{\frac{[K_{\text{out}}]}{[K_{\text{in}}]}\right\}, \tag{37}$$

where R is the gas constant, F is Faraday's constant, and T is the temperature and $[K_{\text{in}}]$ and $[K_{\text{out}}]$ are the intracellular and extracellular concentrations of K^+, respectively. Here, we choose the temperature to be room temperature ($T = 298\text{K}$), and our intracellular and extracellular concentrations to be 120 mM and 5 mM, respectively, which approximately correspond to physiological concentrations (Hille 2001). Hence, for all synthetic data generation, training, and validation in Sects. 3.1 and 3.2, we fix $E_{\text{Kr}} = -80.24\,\text{mV}$ (using Eq. 37).

From Eqs. 33–36, we can see that both the dynamics of the model and the observation function are dependent on the voltage, $V(t; d)$. This is a special case of Eqs. 23–25, in which $\mathbf{f}$ and h are time-dependent only via the voltage $V(t; d)$. In other words, at any fixed voltage (V), Eq. 33 is an autonomous system and h does not depend directly on t.

We assume that our observational errors are additive Gaussian IID random variables with zero mean and variance, σ^2.

The first model of I_{Kr} we consider is by Beattie et al. (2018). This is a four state Markov model with nine parameters (8 of which relate to the model's kinetics and form $\boldsymbol{\theta}_f$). We use the parameters that Beattie et al. obtained from one particular cell (cell #5) by training their model to data obtained from an application of the 'sinusoidal protocol' with a manual patch-clamp setup. The cells used were Chinese hamster ovary cells, which were made to heterologously over-express *hERG1a*. These experiments were performed at room temperature.

The second model is by Wang et al. (1997). This is a five-state model which has 15 parameters. These parameters were obtained by performing multiple voltage-clamp

protocols, all at room temperature, on multiple frog oocytes overexpressing *hERG*. These experiments are used to infer activation and deactivation time constants as well as steady-state current voltage relations, which are, in turn, used to produce estimates for the model parameters. Of the two parameter sets provided in Wang et al. (1997), we use the parameter set obtained by using the extracellular solution with a 2 mM concentration of potassium chloride, as this most closely replicates physiological conditions.

The systems of ODEs for the Beattie and Wang models, as well as the parameterisations of the transition rates, are presented in Appendix 5.1. The values of the model parameters, as used in Sects. 3.1 and 3.2 are given in Table 1.

2.2.2 Experimental Designs for Voltage Clamp Experiments

A large amount of data can be recorded in voltage-clamp electrophysiology experiments: the current can be recorded at a several-kHz sampling rate, for many minutes. In what follows, we take observations of the current at the same 10 kHz frequency for all protocols. Experimenters have a great deal of flexibility when it comes to specifying voltage-clamp protocol designs. We have published a number of studies on the benefits of 'information-rich' experimental designs for these protocols, focusing on short protocols which explore a wide range of voltages and timescales (Beattie et al. 2018; Lei et al. 2019a, b; Clerx et al. 2019a; Kemp et al. 2021). In a real patch-clamp experiment, the amount of data we can obtain from each cell is limited. Hence, it is not feasible to perform many long protocols in sequence on the same cell. For this reason, we use six short information-rich protocols, denoted d_0 to d_5, as shown in Fig. 3.

Here, we use simple designs consisting of a combination of sections where the voltage is held constant or 'ramps' where the voltage increases linearly with time for compatibility with automated high-throughput patch clamp machines which are restricted to protocols of this type. For the protocols included in this paper, short identical sequences including ramps are placed at the beginning and end of each protocol. In real experiments, these elements will allow for quality control, leak subtraction, and the estimation of the reversal potential (Lei et al. 2019a, b). The central portion, consisting of steps during which the voltage is held constant, is what varies between protocols.

Not all possible designs are suitable for training models. Sometimes we encounter protocols for which distant pairs of parameter sets yield approximately equal model output—i.e. the model output for a protocol is not sensitive to certain (possibly large) changes in the model parameters. Subsequently, when training the model to data generated from this protocol, many different parameter sets give similar fits that are almost equally plausible. This problem is loosely termed *numerical unidentifiability* (Fink and Noble 2009) and generally speaking is best avoided, unless the resulting uncertainty in the model parameters is known to be immaterial regarding any possible future context of use.

For both the Beattie and Wang models, numerical unidentifiability is a problem for design d_0 (data not shown, but this phenomenon is illustrated for a similar I_{Kr} model and protocol in Fig. 3 of Whittaker et al. (2020)). Yet d_0 mimics the transmembrane voltage of a heart cell in a range of scenarios, and so provides a good way to validate

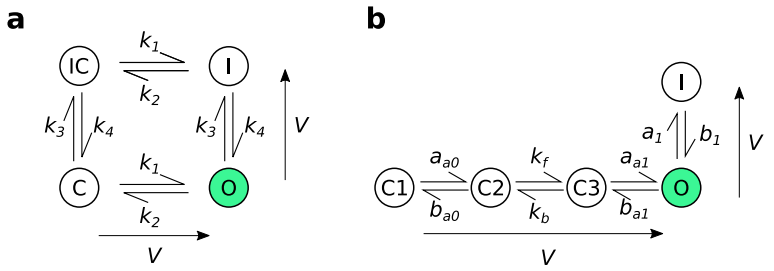

Fig. 2 (Color figure online) The structural differences between the two Markov model structures used in this paper for synthetic data generation and model training. **a** The four-state Beattie model used in both *Case I* and *Case II*. **b** The five-state Wang model used only for *Case II*. When a channel is in the open/conducting (*O*) state (green) current is able to flow. Whereas, when the model is in the other *closed* (C) or *inactivated* (I) states, no current can flow. The arrows adjacent to each model structure indicate the direction in which rates increase as the voltage increases

whether our models recapitulate well-studied, physiologically-relevant behaviour. In particular, the central portion of this voltage-protocol consists of a sequence of waveforms, each of which resembles the action potential of muscle cells found in the heart (ten Tusscher et al. 2004). So in this study we use d_0 as a validation protocol, but do not use it as a protocol for training models.

The remaining designs, d_1–d_5, were constructed using various criteria under constraints on voltage ranges and the duration of each step. The design d_1 was designed algorithmically by sampling from a probability distribution placed over possible parameter sets and maximising the difference in model outputs between all pairs of parameter sets sampled from this distribution; d_5 was the result of the same algorithm applied to the Wang model. In contrast, d_4 is a manual design we have used previously (Lei et al. 2019a) based on a simplification of a sinusoidal design (Beattie et al. 2018). The design, d_2 is a further manual refinement of d_4 to explore inactivation processes (rapid loss of current at high voltages) more thoroughly. Finally, d_3 is based on maximising the exploration of the model phase-space for the Beattie model, visiting as many combinations of binned model states and voltages as possible. The main thing to note for this study however, is that d_1–d_5 result in good numerical identifiability (Fink and Noble 2009) for both models—that is, when used in synthetic data studies that attempt to re-infer the underlying parameters, all five protocol designs yield very low-variance parameter estimates (as shown in Fig. 2 for $\lambda = 1$ for the Beattie model, and Fig. 5 for the Wang model). This is a useful property, because it allows us to disregard the (very small) effect of different random noise in the synthetic data on the spread of our predictions (Eq. 31).

2.2.3 Computational Methods

Numerical solution of ODEs

Any time we calculate $\mathbf{y}(\boldsymbol{\theta}; d)$, we must solve a system of ordinary differential equations. We use a version of the LSODA solver designed to work with the Numba package, and Python to allow for the generation of efficient just-in-time compiled code from symbolic expressions. We partitioned each protocol into sections where the

voltage is constant or changing linearly with respect to time because this sped up our computations. We set LSODA's absolute and relative solver tolerances to 10^{-8}. The fact that the total number of channels is conserved in our models, allows us to reduce the number of ODEs we need to solve from N to $N-1$ (Fink and Noble 2009).

Synthetic data generation

Having computed the state of the system at each observation time, $(\mathbf{x}(t_i, \boldsymbol{\theta}^*, d))_{i=1}^{n_d}$, it is simple to compute y_i by substituting $\mathbf{x}$ into our observation function (Eq. 25). Finally, to add noise, we obtain n_d independent samples using Eq. 25, using *NumPy*'s (Harris et al. 2020) interface to the PCG-64 pseudo-random number generator. Here, because we are using equally spaced observations with a 10 kHz sampling rate, $n_d = 10^4 \times t_{\text{dur}}$ where t_{dur} is the duration of the protocol in seconds.

Optimisation

Finding the least-squares estimates, or, equivalently, minimising Eq. 26 is (in general) a nonlinear optimisation problem for which there exist many numerical methods. We use CMA-ES (Hansen 2006) as implemented by the PINTS interface (Clerx et al. 2019b). CMA-ES is a global, stochastic optimiser that has been applied successfully to many similar problems.

We follow the optimisation advice described in Clerx et al. (2019a). That is, for parameters 'a' and 'b' in state transition rates of the form $k = a \exp(bV)$, the optimiser works with '$\log a$' and untransformed 'b' parameters. We enforce fairly lenient constraints on our parameter space, Θ, to prevent a proposed parameter set from forcing transitions to become so fast/slow that the ODE system becomes very stiff and computationally difficult to solve. In particular, we take a similar approach to Clerx et al. (2019a) we require that every parameter is positive, and, for ease of computation,

$$1.67 \times 10^{-5}\,\text{ms}^{-1} \leq k_{\max} \leq 10^{3}\,\text{ms}^{-1} \tag{38}$$

where $k_{\max}$ is the maximum transition rate, $k(V)$, for all

$$V \in [-120\,\text{mV}, +60\,\text{mV}],$$

which is the voltage range used in our protocols (Fig. 3).

Because CMA-ES is a stochastic algorithm, repeated runs can produce different output. To ensure that we have found the global minimum (Eq. 27), we repeat every optimisation numerous times (25 repeats for $\lambda = 1$ in Case I, 5 repeats for subsequent λ, and 25 repeats in Case II).

Moreover, in Sect. 3.2, when training the discrepant model, our initial guesses for the kinetic parameters were randomly sampled using

$$\log_{10}(p) \sim U(-7, -1), \tag{39}$$

whereas we set the maximal conductance initial guess (which only affects the observation function) to the value used for data generation (even though these data were generated using a different model structure). We then check that our parameter set satisfies Eq. 38, and resample if necessary before commencing the optimisation routine.

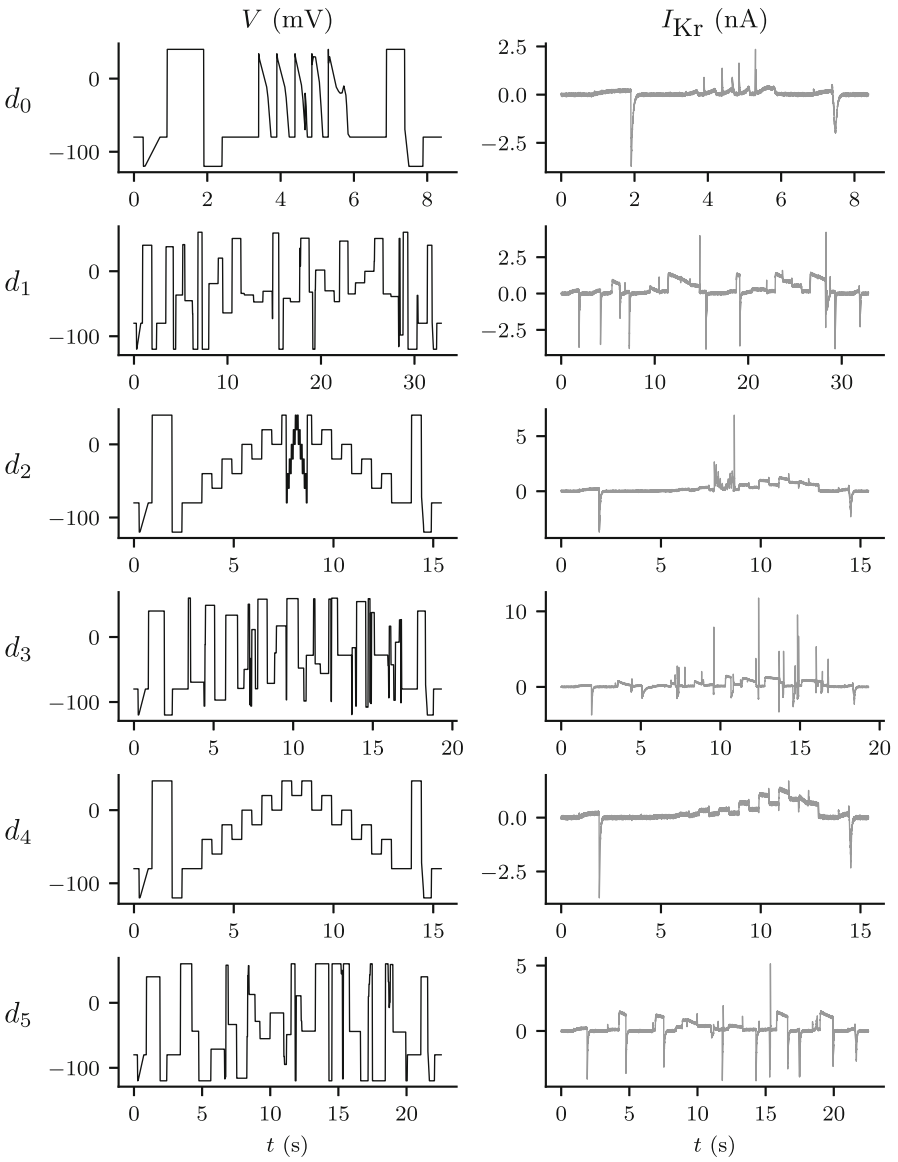

Fig. 3 Left: a range of different input voltage-clamp protocols (forcing functions) used in this study. Right: corresponding synthetic output data I_{Kr} simulated using the Beattie model with noise added as described in Sect. 2.2.3. Here, we generate and plot data observed at a 10 kHz sampling rate. Training protocols (all protocols except d_0) were tested for numerical identifiability (Fink and Noble 2009): inverse problems performed on synthetic data with repeatedly sampled random noise yielded parameter estimates with little variability

The examples presented in Sects. 3.1 and 3.2 require the solution of many optimisation problems. For speed, these tasks may be organised in such a way that multiple optimisation problems can be solved in parallel.

3 Results

In this section, we use synthetically generated data to explore two cases of model discrepancy in Markov models of I_{Kr}. In this first case, we gradually introduce discrepancy into a model with the correct structure by fixing one of its parameters to values away from the DGP parameter set. Then, in Sect. 3.2, we apply the same methods to another case where the model structure is incorrectly specified. In both cases, we take a literature model of I_{Kr} together with Gaussian IID noise to be the DGP.

3.1 Case I: Misspecified Maximal Conductance

In this case, we assume a correctly specified model, but assume increasingly erroneous values of one particular parameter and investigate how this impacts the protocol-dependence our parameter estimates and the predictive accuracy of our models. Also, we explore how the spread in our model predictions (Eq. 31) increases as the amount of discrepancy increases (in a particular manner).

To do this, we simulate data generation from each training protocol, as outlined, ten times using Gaussian IID noise with standard deviation (0.03nA). Specifically, we take the true DGP to be the Beattie model, as shown in Fig. 2. Then, we fix the maximal conductance (g) to a range of values, and infer the remaining model parameters from the synthetic data, generated using the true parameter set, $\boldsymbol{\theta}^*$. We assume that the standard deviation of the Gaussian noise is known because it can be well estimated from the initial portion of each protocol where the current is stationary.

When training our models, we use a restriction of the usual parameter space to fit the data by assuming some fixed value, λ, for the maximal conductance, g. In this way, we reformulate the optimisation problem slightly such that Eq. 27 becomes

$$\hat{\boldsymbol{\theta}}_\lambda(d) = \text{argmin}_{\boldsymbol{\theta}\in\Theta_\lambda} \left\{\text{RMSE}\left(\mathbf{y}(\boldsymbol{\theta}; d), \mathbf{z}(d)\right)\right\}, \tag{40}$$

where Θ_λ is the subset of parameter space where the maximal conductance is fixed to λg. For each repeat of each protocol, we solve this optimisation problem for each scaling factor, $\lambda \in \left\{\frac{1}{4}, \frac{1}{2}, 1, 2, 4\right\}$. These calculations are identical to those used in the computation of *profile likelihoods* under the assumption of additive IID Gaussian errors (Bates 1988).

Next, we fit these restricted parameter-space models to the same dataset and assess their predictive power under the application of a validation protocol. We do this for each possible pair of training and validation protocols.

To reduce the time required for computation we fit our discrepant models sequentially, starting at $\lambda = 1$ and increasing or decreasing λ, using previous parameter estimates as an initial guess. This is done so that, for example, the kinetic parameters

found by fixing $\lambda = 2$ are used as our initial guess when we fit the model with $\lambda = 4$, unless the original kinetic parameters (Table 1) provide a lesser RMSE than the results of the previous optimisation.

The spread in predictions for the validation protocol, d_0, for, $\lambda \in \{\frac{1}{4}, 1, 4\}$ is shown in Fig. 4. A more complete summary of these results is provided by Fig. 5. Here, when $\lambda = 1$ (the central row of Fig. 5), we can see that no matter what protocol is used to train the model, the distribution of parameter estimates (panel **a**) is centred around their true values, and the resultant predictions are all accurate (Fig. 5, panels **b** and **c**). In contrast, when the maximal conductance, g is set to an incorrect value our parameter estimates become biased, and overall, our predictions become much less accurate. This effect on predictive accuracy is also shown in Fig. 10.

Moreover, the inaccuracy in our parameter estimates and our predictions varies depending on the design used to train the model. This effect does not appear to be symmetrical, with $\lambda < 1$ seemingly resulting in more model discrepancy than $\lambda > 1$.

Further results are provided in Sect. 5.2. Figure 10 shows the error in our predictions of d_0 as λ varies, Table 2 examines the distribution of our parameter estimates for each protocol (under repeated samples of the DGP) for different values of λ and Table 3 shows the behaviour of our spread-of-predictions interval (Eq. 31) and midpoint prediction Eq. 32 for different values of λ.

3.2 Case II: Misspecified Dynamics

Next, we apply these methods to an example where we have misspecified the dynamics of the model (the function $\mathbf{f}$). We use two competing Markov models of hERG kinetics, the Beattie model (Beattie et al. 2018), and the Wang model (Wang et al. 1997). These models have differing structures and differing numbers of states, as shown in Fig. 2. We generate a synthetic dataset under the assumption of the Wang model with Gaussian IID noise (with standard deviation 0.03nA) and the original parameter set as given in Wang et al., for all the protocols shown in Fig. 3. As in Case I, we assume the standard deviation of this noise is known.

Then, we are able to fit both models to this dataset, obtaining an ensemble of parameter estimates and performing cross-validation as described in Sect. 2. In this way, we can assess the impact of choosing the wrong governing equations (the choice of $\mathbf{f}$ in Eq. 19), and its impact on the predictive accuracy of the model. We do this to investigate whether the techniques introduced in Sect. 2.1 are able to provide some useful quantification of model discrepancy when the dynamics of I_{Kr} are misspecified.

Our results, presented in Figs. 6 and 7, show how we expect a correctly specified model to behave in comparison to a discrepant model. We can see from the bottom row of Fig. 7, that when training using the correctly specified derivative matrix, we were able to accurately recover the true maximal conductance using each and every protocol. Moreover, similarly to *Case I*, no matter which protocol the correctly specified model was trained with, the resultant predictions were very accurate (as can be seen in Fig. 6).

However, when the discrepant model was used, there was significant protocol dependence in our parameter estimates, and our predictions were much less accurate overall, but perhaps accurate for many applications. Moreover, it seems that for

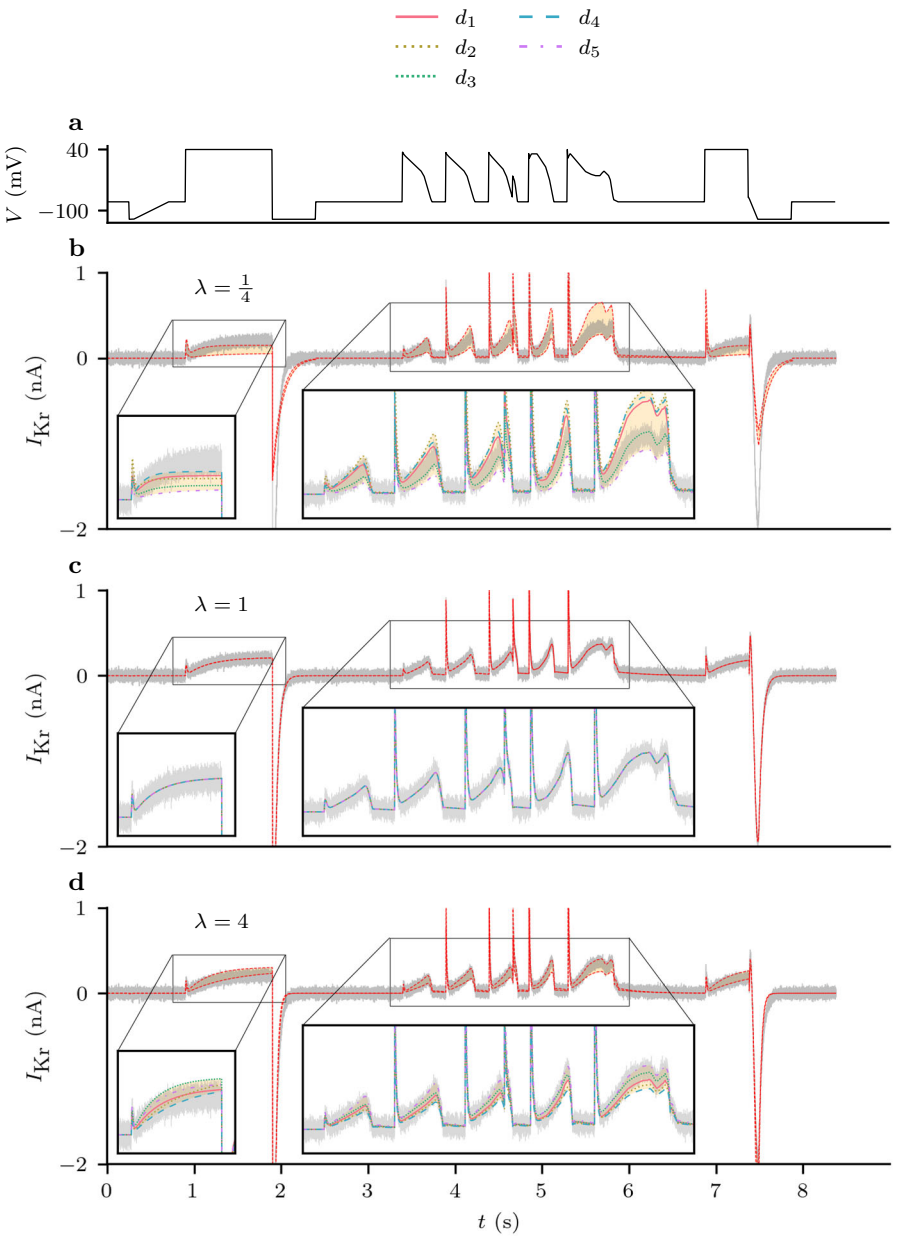

Fig. 4 (Color figure online) The set of predictions (Eq. 30) shown for parameter estimates obtained by training with different values of λ to synthetic data under $d_1, \ldots, d_5$ (using the Beattie model). The synthetically generated data used for model validation are shown in grey and and the spread of the predictions is highlighted in yellow. **a** The voltage trace for d_0. **b** The set of predictions with $\lambda = \frac{1}{4}$. **c** The set of predictions with $\lambda = 1$, that is, under the assumption of the correct maximal conductance (g). **d** The set of predictions with $\lambda = 4$. N.B. the 'angular' nature of the current is not a plotting artefact, but reflects the fact the voltage clamp (**a**) is constructed from a series of linear ramps for compatibility with automated voltage clamp machines

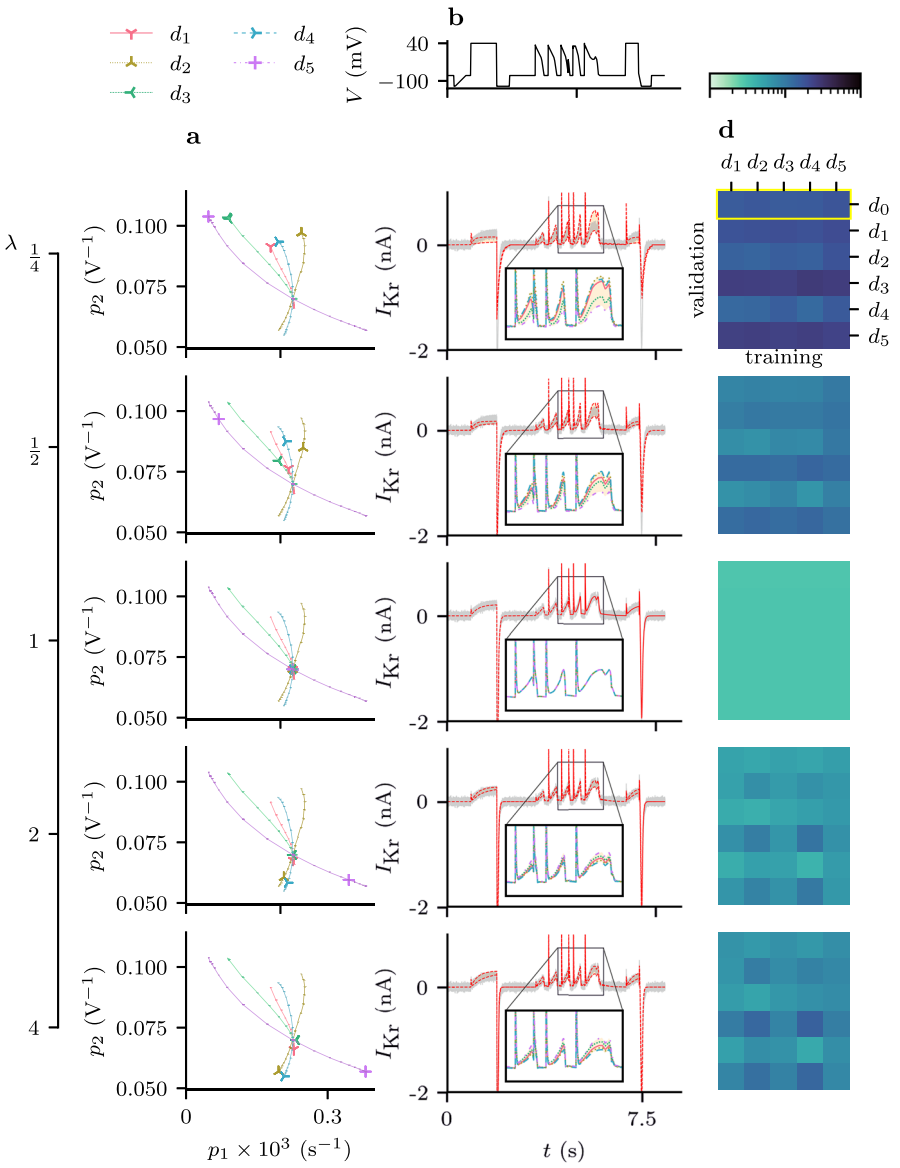

Fig. 5 (Color figure online) Discrepancy in parameter estimates and subsequent currents when a non-discrepant model is fitted to synthetic data, with all parameters free except the maximal conductance, g, which is scaled by some factor λ, ($g = \lambda g^*$, where g^* is the true value). **a** Estimates of θ_1 and θ_2 obtained by training with different protocols for 10 repeats of the DGP. The lines (linearly interpolated using 17 values for $\lambda \in [\frac{1}{4}, 4]$) show how the estimates from each protocol improve as $\lambda \to 1$. **b** d_0 voltage protocol. **c** The spread of predictions of I_{Kr} under the d_0 protocol using the parameter estimates in Column **a**. **d** A heatmap showing the predictive error obtained by training and validating for each pair of protocols. Here **c** corresponds with the top row of each heatmap, as indicated

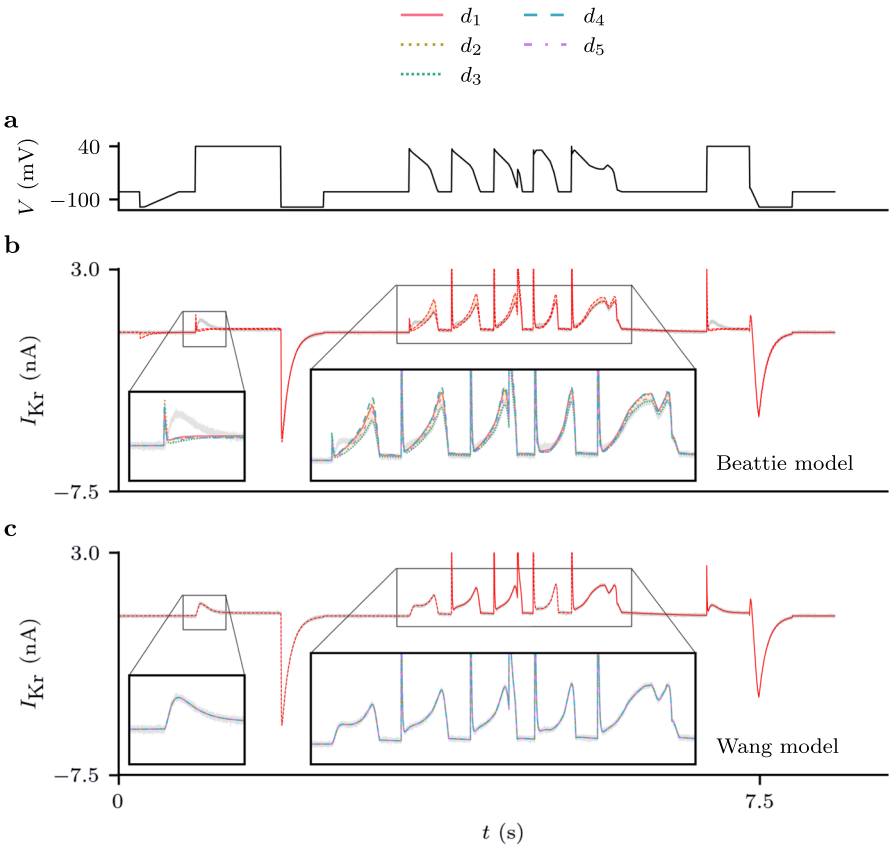

Fig. 6 (Color figure online) Case II: the set of predictions (Eq. 30) shown for parameter estimates obtained by training Beattie and Wang models with data synthetically generated using the Wang model. **a** The d_0 voltage-clamp protocol. **b** The set of predictions using the Beattie model. **c** The set of predictions with Wang model, that is, with under the assumption of the correct model structure

the majority of d_0, the spread in predictions across training protocols (Eq. 31) was smaller than those seen in Case I, but there are certain portions where the discrepant model and DGP are noticeably different (as highlighted in Fig. 6). This may be due to the structural differences between the Wang and Beattie model. In particular, in the Wang model, channels transitioning from the high-voltage inactive state (I), must transition through the conducting, open state (O) in order to reach low-voltage closed states ($C1$, $C2$, $C3$), causing a spike of current to be observed. Instead, channels in the Beattie model may transition through the inactive-and-closed state (IC) on their way between O and C, resulting in reduced current during steps from high voltage to low voltage.

Nevertheless, our methods provide a useful indication of this model discrepancy. Figure 8, examines the behaviour of our prediction interval (Eq. 31) in more detail. Importantly, we can see that our interval shows little uncertainty during sections of the protocol where there is little current (this is also seen in Fig. 7**a** and Fig. 6). This is

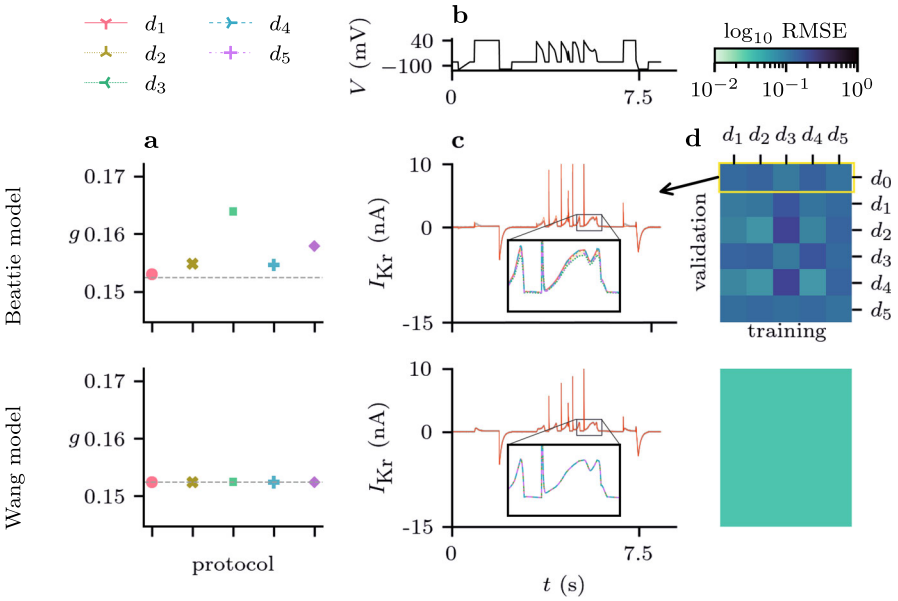

Fig. 7 (Color figure online) Model discrepancy between the Beattie model and synthetic data generated using the Wang model. **a** estimates of the maximal conductance obtained by training with different protocols for ten repeats of the DGP. There is a noticeable protocol dependence for estimates obtained using the (discrepant) Beattie model, but the true underlying parameter (dashed line) can be accurately determined from any protocol when using the (correct) Wang model. **b** d_0 voltage protocol. **c** the spread of predictions for I_{Kr} under the d_0 protocol for discrepant (Beattie) and correct (Wang) models which are shown in more detail in Fig. 6. **d** cross-validation heatmaps for both the Beattie and Wang models fitted to this suite of protocols, averaged over ten repeated samples of the DGP for each protocol

ideal, because no reasonable model would predict a sizeable current here. On the other hand, we see that our intervals show significant uncertainty around the spikes in current that occur at the start of each action-potential waveform. This is to be expected because it is known that these sections of the current time-series are particularly sensitive to differences in the 'rapid inactivation' process in these models (Clerx et al. 2019a).

Further results regarding Case II are provided in Sect. 5.3. Tables 4 and 5 summarise the behaviour of our parameter estimates for each choice of model.

4 Discussion

We have introduced an uncertainty quantification (UQ) approach to highlight when discrepancy is affecting model predictions. We demonstrated the use of this technique by providing insight into the effects of model discrepancy on a model of I_{Kr} in electrically excitable cells. Here, we saw that under synthetically constructed examples of model discrepancy, there was great variability between the parameter estimates obtained using different experimental designs. This variability is a consequence of the different compromises that a discrepant model has to make to fit different regimes of a

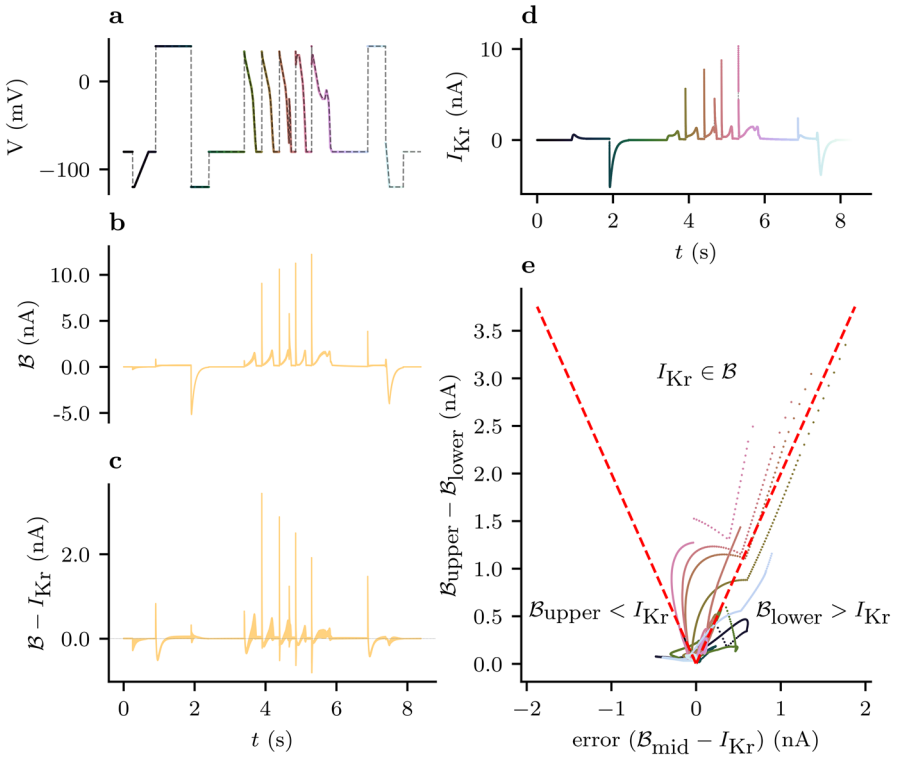

Fig. 8 (Color figure online) The spread in predictions obtained from different protocols provides a useful indicator of model discrepancy for *Case II*. **a** The validation voltage-clamp protocol, d_0, with colours corresponding to panels **d** and **e**. **b** The spread-of-predictions interval (Eq. 31) for d_0 using the Beattie model trained with $d_1, \ldots, d_5$. **c** The true DGP subtracted from the spread-in-predictions interval. **d** The true DGP with the colour of each observation corresponding to panels **a** and **e**. **e** a scatter plot of the midpoint prediction (Eq. 32) and the width of the predictive interval (Eq. 31) for every observation in d_0. Here, the red, dashed lines show the true value of I_{Kr} lies on the extremes of the range of predictions. Accordingly, points above these lines show the observations for which the DGP lies inside this range, and the points below the line correspond to observations for which the true DGP lies outside this range. The colours of these points correspond to those in panels **a** and **d**

true DGP's behaviour. Consequently, these parameter estimates produced a wide range of behaviour during validation, despite each individual parameter estimate having little variability under repeated samples of the DGP.

The variability in the model predictions stemming from this ensemble of parameter estimates is, therefore, an empirical way of characterising the predictive uncertainty due to model discrepancy. Usefully, our spread-of-prediction intervals (Eq. 31) correctly indicated little uncertainty when the ion channel model was exhibiting simple dynamics decaying towards a steady state, but more uncertainty during more complex dynamics, which was indeed when the largest discrepancies occurred. For many observations under our validation protocol, the true, underlying DGP lay inside this interval, indicating that Eq. 31 may provide a useful indication of predictive error under unseen protocols. We expect that the presented methods may be of use for problems where

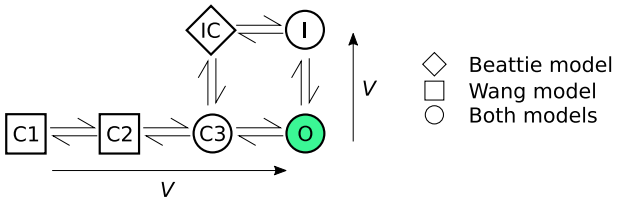

Fig. 9 (Color figure online) The Beattie and Wang models may be seen as special cases of this more complicated model. The state labelled 'C3' is called 'C' in the Beattie model and 'C3' in the Wang model. The arrows outside the Markov state diagram indicate the direction in which rates increase with increasing voltage

the variability in parameter estimates (from repetitions of each individual protocol) is smaller than the variability between parameter estimates obtained from different protocols—because there is little noise, and lots of observations for example. In such cases, the variability in the extremes of our ensembles ($\mathcal{B}_{\text{upper}}$ and $\mathcal{B}_{\text{lower}}$) is immaterial compared to the width of the interval ($\mathcal{B}_{\text{upper}} - \mathcal{B}_{\text{lower}}$).

At first, Case I may seem like an artificial example—in practice, the maximal conductance is taken to be a model parameter and fitted along with the rest of the model. But Case I and Case II are similar: any two Markov models may be regarded as two special cases of a more general model with some transition rates pinned to 0 (as shown in Fig. 9 for the models used in this paper, with some transition rate). Like in Case I, this means that different model structures can be seen as restrictions of this larger model's parameter space. Misspecified model structures can then be identified with subsets of parameter space which do not contain the true, data-generating parameter set (provided this larger model is structurally identifiable).

This means there is a setting in which Case II (misspecified governing equations) is an example of the type of discrepancy explored in Case I, where a "true" parameter value exists in the more general model, but is excluded in the parameter space being optimised over when training the model. This may prove a valuable perspective for modelling ion channel kinetics, where there are many candidate models (Mangold et al. 2021), and each model may be seen as corresponding to some subset of a general model with a shared higher-dimensional parameter space. Model selection problems have been framed in this way previously (Akaike 1998; Chen et al. 2017).

4.1 Limitations

Whilst the spread of predictions under some unseen protocol may provide some feasible range of predictions, we can see from Fig. 8, that our observables (the DGP without noise) often lie outside this range. This shown in Fig. 6. Here, certain structural differences between the model and DGP may mean that the truth lies outside. For this reason, Eq. 31 is best interpreted as a heuristic indication of predictive uncertainty stemming from model discrepancy, rather than providing any guarantees about the output of the DGP.

Using more training protocols in the training set may increase the coverage of the DGP by our interval. Whilst the number of protocols that can be performed on a single

biological cell is limited by time constraints (Beattie et al. 2018; Lei et al. 2019b), the utilisation of more protocols is likely preferable.

Besides the types of discrepancies considered in Sect. 3, there are other ways that the DGP can differ from the fitted models. For example, the DGP may not be accurately described by an ODE system, especially when ion channel numbers are small and the stochasticity of individual channels opening and closing is apparent. In this circumstance, the models can be cast in terms of stochastic differential equations (SDEs), as in Goldwyn et al. (2011), and we can again consider an ensemble of parameter estimates (Eq. 29) and an ensemble of model predictions (Eq. 30. The assumption of IID Gaussian errors for the observation noise model could also be inaccurate: auto-correlated noise processes (e.g. as explored in Creswell et al. 2020; Lambert et al. 2022), or even experimental artefacts may be present, but all of these could be included in the modelled DGP (Lei et al. 2020a) and it remains to be seen how well our method would perform in these cases.

4.2 Future Directions

We were able to quantify model discrepancy by considering the predictive error of our models across a range of training and validation protocols. This provides a way of quantifying model discrepancy that can be compared across models, and could be used to select the most suitable model from a group of candidate models. For a given context of use, we suggest that the spread of predictions can be used to gauge the trustworthiness of a model's predictions. Even in a model which produces a plausible fit to each individual training protocol, a wide spread of predictions may indicate that a model is ill-suited to a particular predictive task, and should prompt careful reconsideration of the model and the experiments used for its training. In this way, the disagreement between model predictions of d_0 in Sect. 3.1 shows that the $\lambda = \frac{1}{4}$ may not be suitable, owing to the relative width of this band of predictions, even in the absence of validation data, or knowledge of a more suitable model.

Our approach may provide insight into improved experimental design criteria (Lei et al. 2022). Optimal experimental design approaches that assume knowledge of a correctly specified model may not be the most suitable in the presence of known discrepancy. By adjusting these approaches to account for the uncertainty in choice of model, we may be able to use these ideas to design experiments which allow for more robust model validation. One method would be to fit to data from a collection of training protocols, and to find a new protocol, for which the spread in ensemble predictions (Eq. 31) is maximised.

In this paper, we have applied our methodology to mathematical models of electrophysiology. In both the cases we considered, we saw that model discrepancy could lead to inaccurate predictive models, and due to the use of information-rich training protocols, the variability of our parameter estimates, under repeated samples of the DGP, was negligible. We propose that our methodology could be applied to similar problems, where there is high-frequency time-series data is available as well as mathematical model which are relatively accurate. There are many such biochemical reaction networks for which model selection remains a challenge, and there are numer-

ous approaches to finding suitable mathematical models Klimovskaia et al. (2016). We propose the methodology outlined in this paper may be used to quantify the discrepancy in such models.

In these examples, we saw little variability due to noise in our parameter inference, and therefore in our ensemble prediction's spread-of-prediction intervals and/or midpoint predictions, as shown in Sect. 5.2 and Sect. 5.3. However, in other cases where there are fewer observations and more observational noise (for example), it may be more suitable to consider an analogous distribution-based approach where we consider Bayesian posteriors of our parameters instead of point estimates (such as the maximum likelihood estimator we used in this paper).

4.3 Concluding Remarks

The spread of predictions of our ensembles, based on training to data from multiple experimental designs, provides a good indication of possible predictive error due to model discrepancy. Ultimately, whilst our ensemble approach is no substitute for a correctly specified model, it is a useful tool for quantifying model discrepancy, predicting the size and direction of its effects, and may guide further experimental design and model selection approaches.

Acknowledgements This work was supported by the Wellcome Trust (grant no. 212203/Z/18/Z); the Science and Technology Development Fund, Macao SAR (FDCT) [reference no. 0048/2022/A]; the EPSRC [grant no. EP/R014604/1]; and the Australian Research Council [grant no. DP190101758]. GRM acknowledges support from the Wellcome Trust via a Wellcome Trust Senior Research Fellowship to GRM. CLL acknowledges support from the FDCT and support from the University of Macau via a UM Macao Fellowship. We acknowledge Victor Chang Cardiac Research Institute Innovation Centre, funded by the NSW Government. The authors would like to thank the Isaac Newton Institute for Mathematical Sciences for support and hospitality during the programme The Fickle Heart when some work on this paper was undertaken. This research was funded in whole, or in part, by the Wellcome Trust [212203/Z/18/Z]. For the purpose of open access, the authors have applied a CC-BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Data Availability Open source code for all the simulation studies and plots in this paper can be found at https://github.com/CardiacModelling/empirical_quantification_of_model_discrepancy. A permanently archived version is available on Zenodo https://doi.org/10.5281/zenodo.8409925.

5 Appendices

5.1 Appendix A: Parameterisation of Markov Models

5.1.1 Beattie Model

In full, the system of ODEs is,

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \begin{bmatrix} -k_1 - k_3 & 0 & k_4 & k_2 \\ 0 & -k_2 - k_4 & k_1 & k_3 \\ k_3 & k_2 & -k_1 - k_4 & 0 \\ k_1 & k_4 & 0 & -k_2 - k_3 \end{bmatrix} \mathbf{x}, \tag{41}$$

where

$$k_1 = p_1 e^{p_2 V}, \tag{42}$$
$$k_2 = p_3 e^{-p_4 V}, \tag{43}$$
$$k_3 = p_5 e^{p_6 V} \quad , \tag{44}$$
$$k_4 = p_7 e^{-p_8 V}. \tag{45}$$

Hence, the corresponding parameter set is,

$$\boldsymbol{\theta} = \left[g,\, p_1,\, \ldots,\, p_8\right]^T, \tag{46}$$

and,

$$\mathbf{x} = \left[C,\; I,\; IC,\; O\right]^T. \tag{47}$$

5.1.2 Wang Model

We may write this model's governing system of ODEs as

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \begin{bmatrix} -\alpha_{a0} & \beta_{a0} & 0 & 0 & 0 \\ \alpha_{a0} & -\beta_{a0} - k_f & k_b & 0 & 0 \\ 0 & k_f & -k_b - \alpha_{a1} & \beta_{a1} & 0 \\ 0 & 0 & \alpha_{a1} & -\beta_{a1} - \alpha_1 & \beta_1 \\ 0 & 0 & 0 & \alpha_1 & -\beta_1 \end{bmatrix} \mathbf{x}, \tag{48}$$

where

$$\alpha_1 = q_1 e^{q_2 V}, \tag{49}$$
$$\alpha_{a0} = q_3 e^{q_4 V}, \tag{50}$$
$$\alpha_{a1} = q_5 e^{q_6 V}, \tag{51}$$

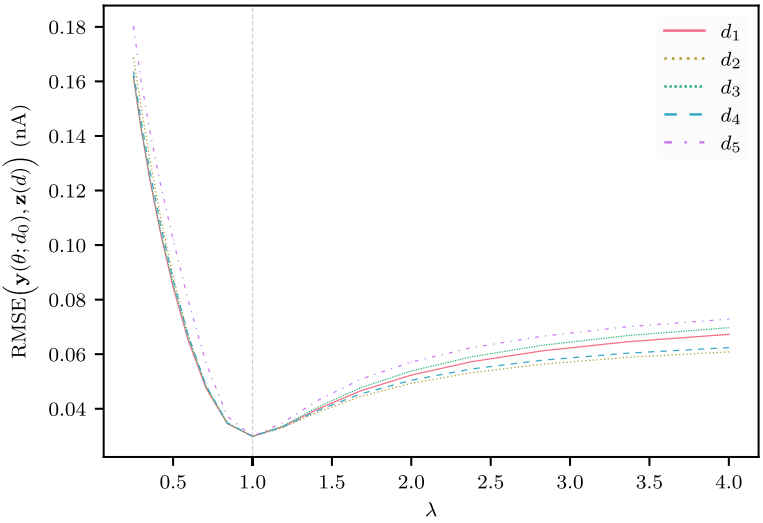

Fig. 10 (Color figure online) Case I: predictive accuracy (under our validation protocol, d_0) decreases as $\lambda \to 1$. For 17 values of λ ($\frac{1}{4} \leq \lambda \leq 4$), the predictive error (averaged over repeats) is shown for each training protocol (d_1–d_5)

$$\beta_{a1} = q_7 e^{-q_8 V}, \tag{52}$$

$$\beta_1 = q_9 e^{-q_{10} V}, \tag{53}$$

$$\beta_{a0} = q_{11} e^{-q_{12} V}. \tag{54}$$

The corresponding parameter set is,

$$\boldsymbol{\theta} = \left[g,\ k_b,\ k_f,\ q_1,\ \ldots,\ q_{12}\right]^T, \tag{55}$$

and

$$\mathbf{x} = \left[C_1,\ C_2,\ C_3,\ O,\ I\right]^T. \tag{56}$$

The default parameter values for both models are presented in Table 1.

5.2 Appendix B: Further Case I Results

The predictive accuracy (under the validation protocol, d_0) of the model used in Sect. 3.2, trained using each protocol, for a range of values of λ is shown in Fig. 10. Here, we see that as there is more model discrepancy (when λ moves away from 1) our predictions become less accurate.

The parameter estimates obtained in Sects. 3.1 and 3.2 are summarised in Tables 2, 4 and 5, respectively. Here, we can see that when the model is misspecified, the small standard deviation in our estimates (across fits to different samples of our DGP) is small

Table 1 The default parameter sets we use for the Wang et al. (1997) and Beattie et al. (2018) models

Parameter	Value	Units
Wang model		
g	1.52×10^{-1}	μS
k_b	3.68×10^{-2}	ms^{-1}
k_f	2.38×10^{-2}	ms^{-1}
q_1	9.08×10^{-2}	ms^{-1}
q_2	2.34×10^{-2}	mV^{-1}
q_3	2.23×10^{-2}	ms^{-1}
q_4	1.18×10^{-2}	mV^{-1}
q_5	1.37×10^{-2}	ms^{-1}
q_6	3.82×10^{-3}	mV^{-1}
q_7	6.89×10^{-5}	ms^{-1}
q_8	4.18×10^{-2}	mV^{-1}
q_9	6.50×10^{-3}	ms^{-1}
q_{10}	3.27×10^{-2}	mV^{-1}
q_{11}	4.70×10^{-2}	ms^{-1}
q_{12}	6.31×10^{-2}	mV^{-1}
Beattie model		
g	1.52×10^{-1}	μS
p_1	2.26×10^{-4}	ms^{-1}
p_2	6.99×10^{-2}	mV^{-1}
p_3	3.45×10^{-5}	ms^{-1}
p_4	5.46×10^{-2}	mV^{-1}
p_5	8.73×10^{-2}	ms^{-1}
p_6	8.91×10^{-3}	mV^{-1}
p_7	5.15×10^{-3}	ms^{-1}
p_8	3.16×10^{-2}	mV^{-1}

The same maximal conductance (g) is used for both models

compared to the differences between estimates obtained from different protocols—the choice of training protocol is less important when there is no model discrepancy.

Table 2 details the distribution of each parameter estimate (under repeated samples of the DGP) for each protocol as λ varies (as described in Sect. 3.1. Whereas, Table 3 shows how our spread-of-prediction intervals change under different values of λ. Here, we can see that each parameter estimate, as well as $\mathcal{B}$ itself, show little variability under repeated samples of the DGP.

5.3 Appendix C: Further Case II Results

Tables 4 and 5 summarise the parameter estimates obtained in Sect. 3.2 using the Beattie model and Wang model, respectively. Here, the Wang model was chosen as

Table 2 The mean and standard deviation (across different synthetic datasets) of estimates (Eq. 27) used in Case I, where the maximal conductance, is misspecified by scaling it with λ

λ		d_1	d_2	d_3	d_4	d_5
0.25	p_1	1.8E−4±2E−7	2.4E−4±7E−7	9.0E−5±1E−6	1.9E−4±4E−7	4.7E−5±8E−8
	p_2	9.1E−2±5E−5	9.7E−2±9E−5	1.0E−1±2E−4	9.3E−2±8E−5	1.0E−1±4E−5
	p_3	3.0E−5±2E−8	1.9E−5±3E−8	1.2E−5±4E−8	1.7E−5±4E−8	9.3E−6±2E−8
	p_4	4.8E−2±6E−6	5.2E−2±2E−5	5.5E−2±3E−5	5.3E−2±2E−5	5.8E−2±2E−5
	p_5	5.0E−2±5E−5	5.5E−2±6E−5	5.0E−2±5E−5	7.9E−2±3E−4	5.1E−2±8E−5
	p_6	1.5E−2±2E−5	1.0E−2±4E−5	1.0E−2±2E−5	1.8E−2±5E−5	1.3E−2±3E−5
	p_7	1.5E−2±2E−5	1.3E−2±2E−5	9.8E−3±2E−5	2.3E−2±7E−5	1.0E−2±5E−5
	p_8	4.4E−2±2E−5	4.6E−2±5E−5	5.1E−2±4E−5	3.6E−2±4E−5	5.8E−2±1E−4
0.50	p_1	2.2E−4±2E−7	2.5E−4±6E−7	2.0E−4±6E−7	2.1E−4±4E−7	6.9E−5±2E−7
	p_2	7.6E−2±3E−5	8.5E−2±8E−5	8.0E−2±6E−5	8.8E−2±7E−5	9.7E−2±6E−5
	p_3	3.7E−5±2E−8	3.4E−5±5E−8	3.6E−5±6E−8	3.6E−5±6E−8	2.5E−5±3E−8
	p_4	5.0E−2±4E−6	5.1E−2±1E−5	5.0E−2±2E−5	5.1E−2±1E−5	5.3E−2±9E−6
	p_5	7.5E−2±7E−5	7.5E−2±6E−5	6.8E−2±6E−5	8.4E−2±2E−4	7.7E−2±7E−5
	p_6	9.4E−3±1E−5	9.0E−3±3E−5	9.7E−3±2E−5	1.1E−2±3E−5	9.9E−3±1E−5
	p_7	9.3E−3±1E−5	8.2E−3±7E−6	6.8E−3±8E−6	1.0E−2±3E−5	8.5E−3±1E−5
	p_8	3.7E−2±1E−5	3.8E−2±2E−5	4.0E−2±1E−5	3.5E−2±3E−5	3.8E−2±2E−5
1.00	p_1	2.3E−4±2E−7	2.3E−4±4E−7	2.3E−4±6E−7	2.3E−4±3E−7	2.3E−4±6E−7
	p_2	7.0E−2±3E−5	7.0E−2±7E−5	7.0E−2±5E−5	7.0E−2±5E−5	7.0E−2±6E−5
	p_3	3.4E−5±1E−8	3.4E−5±4E−8	3.4E−5±5E−8	3.4E−5±5E−8	3.4E−5±2E−8
	p_4	5.5E−2±5E−6	5.5E−2±1E−5	5.5E−2±1E−5	5.5E−2±1E−5	5.5E−2±5E−6
	p_5	8.7E−2±7E−5	8.7E−2±8E−5	8.7E−2±6E−5	8.7E−2±3E−4	8.7E−2±6E−5
	p_6	8.9E−3±9E−6	8.9E−3±2E−5	8.9E−3±1E−5	8.9E−3±3E−5	8.9E−3±1E−5
	p_7	5.2E−3±6E−6	5.2E−3±4E−6	5.2E−3±4E−6	5.1E−3±1E−5	5.2E−3±4E−6
	p_8	3.2E−2±9E−6	3.2E−2±2E−5	3.2E−2±8E−6	3.2E−2±3E−5	3.2E−2±1E−5
2.00	p_1	2.3E−4±2E−7	2.1E−4±3E−7	2.3E−4±7E−7	2.1E−4±3E−7	3.4E−4±6E−7
	p_2	6.8E−2±3E−5	6.1E−2±7E−5	7.0E−2±6E−5	5.8E−2±5E−5	6.0E−2±4E−5
	p_3	3.0E−5±1E−8	3.0E−5±4E−8	2.5E−5±4E−8	2.8E−5±5E−8	3.4E−5±2E−8
	p_4	5.9E−2±5E−6	5.8E−2±1E−5	6.0E−2±1E−5	5.8E−2±2E−5	5.8E−2±6E−6
	p_5	9.0E−2±7E−5	8.9E−2±8E−5	9.5E−2±7E−5	9.5E−2±3E−4	9.0E−2±5E−5
	p_6	1.0E−2±6E−6	9.6E−3±2E−5	9.0E−3±1E−5	8.6E−3±4E−5	9.7E−3±8E−6
	p_7	2.6E−3±3E−6	2.9E−3±2E−6	2.9E−3±2E−6	2.8E−3±1E−5	2.7E−3±2E−6
	p_8	2.7E−2±8E−6	2.7E−2±2E−5	2.6E−2±6E−6	2.8E−2±3E−5	2.7E−2±8E−6

Table 2 continued

λ		d_1	d_2	d_3	d_4	d_5
4.00	p_1	2.3E−4±2E−7	2.0E−4±3E−7	2.3E−4±8E−7	2.1E−4±2E−7	3.8E−4±5E−7
	p_2	6.6E−2±3E−5	5.7E−2±6E−5	7.0E−2±6E−5	5.5E−2±4E−5	5.7E−2±3E−5
	p_3	2.9E−5±1E−8	2.8E−5±4E−8	2.3E−5±4E−8	2.6E−5±5E−8	3.3E−5±2E−8
	p_4	6.1E−2±5E−6	6.0E−2±1E−5	6.2E−2±1E−5	6.0E−2±2E−5	5.9E−2±6E−6
	p_5	9.4E−2±7E−5	9.1E−2±9E−5	9.7E−2±7E−5	9.9E−2±4E−4	9.6E−2±5E−5
	p_6	1.1E−2±5E−6	9.7E−3±2E−5	9.1E−3±1E−5	8.6E−3±3E−5	1.0E−2±6E−6
	p_7	1.4E−3±1E−6	1.6E−3±1E−6	1.5E−3±1E−6	1.5E−3±5E−6	1.5E−3±1E−6
	p_8	2.5E−2±8E−6	2.5E−2±1E−5	2.4E−2±5E−6	2.6E−2±3E−5	2.5E−2±7E−6

These were obtained from each training protocol (d_1–d_5) for multiple repeats of synthetically generated data

Table 3 A summary of showing how the spread-of-predictions interval (Eq. 31) behaves under Case I

λ	Mean interval width (nA)	DGP in interval (%)	Midpoint RMSE (nA)
0.25	7.4E−2 ± 1.1E−4	34 ± 2.3E−2	1.6E−1 ± 8.9E−5
0.30	6.6E−2 ± 6.5E−5	37 ± 5.0E−2	1.4E−1 ± 8.9E−5
0.35	6.1E−2 ± 5.8E−5	42 ± 4.3E−2	1.3E−1 ± 9.1E−5
0.42	5.6E−2 ± 6.4E−5	50 ± 4.2E−2	1.1E−1 ± 9.2E−5
0.50	4.9E−2 ±7.1E−5	51 ± 5.1E−2	8.5E−2 ± 9.4E−5
0.59	4.0E−2 ± 8.4E−5	53 ± 6.7E−2	6.5E−2 ± 9.7E−5
0.71	2.8E−2 ± 9.1E−5	55 ± 8.8E−2	4.7E−2 ± 9.0E−5
0.84	1.4E−2 ± 9.3E−5	55 ± 0.17	3.4E−2 ± 8.4E−5
1.00	2.1E−4 ± 3.7E−5	94 ± 9.2	3.0E−2 ± 6.7E−5
1.19	1.2E−2 ± 8.9E−5	56 ± 0.69	3.3E−2 ± 5.1E−5
1.41	2.1E−2 ± 8.5E−5	57 ± 0.67	3.7E−2 ± 5.3E−5
1.68	2.8E−2 ± 8.5E−5	57 ± 1.7	4.2E−2 ± 6.2E−5
2.00	3.3E−2 ± 8.6E−5	56 ± 1.7	4.7E−2 ± 6.9E−5
2.38	3.6E−2 ± 8.6E−5	55 ± 1.4	5.1E−2 ± 7.3E−5
2.83	3.9E−2 ± 8.4E−5	54 ± 1.4	5.4E−2 ± 7.5E−5
3.36	4.1E−2 ± 8.6E−5	54 ± 0.17	5.7E−2 ± 7.6E−5
4.00	4.3E−2 ± 8.4E−5	53 ± 0.15	5.9E−2 ± 7.8E−5

Here we show: the mean width of the interval (averaged over each observation time); the proportion of observations for which the the underlying DGP (minus noise) lies within the interval; the RMSE between the data and the midpoint prediction (Eq. 32). By considering ten randomly sampled datasets (each containing a repeat each protocol d_1–d_5), we show the mean and standard deviation of these values

the DGP and so, the Wang model is an example of a correctly specified model, whereas the Beattie model is a discrepant model. This is reflected by the parameter estimates which show that when the Wang model is fitted to the data, we obtain similar parameter estimates from each protocol, whereas our parameter estimates for the Beattie model are dependent on the protocol used for training.

Table 4 The parameter estimates obtained for Case II (Sect. 3.1) when using the Beattie model to fit data generated by the Wang model

	d_1	d_2	d_3	d_4	d_5
g	1.5E−1±3E−5	1.5E−1±5E−5	1.6E−1±3E−5	1.5E−1±5E−5	1.6E−1±2E−5
p_1	1.6E−3±7E−7	1.6E−3±8E−7	1.7E−3±1E−6	1.7E−3±1E−6	2.0E−3±5E−7
p_2	7.3E−2±2E−5	7.9E−2±2E−5	3.7E−2±2E−5	8.4E−2±4E−5	5.4E−2±2E−5
p_3	1.9E−5±2E−8	2.2E−5±2E−8	4.3E−5±6E−8	2.2E−5±2E−8	3.0E−5±2E−8
p_4	5.2E−2±7E−6	5.1E−2±1E−5	4.6E−2±1E−5	5.1E−2±1E−5	4.9E−2±6E−6
p_5	1.1E−1±7E−5	9.6E−2±2E−5	9.3E−2±2E−5	1.2E−1±2E−4	9.6E−2±6E−5
p_6	2.3E−2±5E−6	2.4E−2±8E−6	2.2E−2±5E−6	2.7E−2±2E−5	2.3E−2±7E−6
p_7	8.9E−3±5E−6	7.1E−3±3E−6	7.4E−3±2E−6	8.9E−3±1E−5	6.8E−3±5E−6
p_8	2.9E−2±8E−6	3.1E−2±1E−5	3.0E−2±7E−6	2.9E−2±2E−5	3.1E−2±7E−6

Table 5 The parameter estimates obtained for Case II when using the correct Wang model to fit its synthetic data

	d_1	d_2	d_3	d_4	d_5
g	1.5E−1±3E−5	1.5E−1±4E−5	1.5E−1±2E−5	1.5E−1±4E−5	1.5E−1±2E−5
k_b	3.7E−2±4E−4	3.6E−2±1E−3	3.7E−2±2E−4	3.6E−2±2E−3	3.7E−2±3E−4
k_f	2.4E−2±9E−5	2.4E−2±4E−4	2.4E−2±9E−5	2.4E−2±6E−4	2.4E−2±7E−5
q_1	9.1E−2±5E−5	9.1E−2±7E−5	9.1E−2±2E−5	9.1E−2±1E−4	9.1E−2±6E−5
q_2	2.3E−2±6E−6	2.3E−2±1E−5	2.3E−2±5E−6	2.3E−2±1E−5	2.3E−2±8E−6
q_3	2.2E−2±5E−4	2.3E−2±7E−4	2.2E−2±3E−4	2.3E−2±8E−4	2.2E−2±3E−4
q_4	1.2E−2±4E−4	1.1E−2±7E−4	1.2E−2±2E−4	1.1E−2±8E−4	1.2E−2±2E−4
q_5	1.4E−2±2E−4	1.4E−2±4E−4	1.4E−2±1E−4	1.4E−2±7E−4	1.4E−2±1E−4
q_6	3.8E−2±2E−4	3.8E−2±3E−4	3.8E−2±1E−4	3.8E−2±6E−4	3.8E−2±2E−4
q_7	6.9E−5±7E−8	6.9E−5±3E−7	6.9E−5±1E−7	6.9E−5±4E−7	6.9E−5±1E−7
q_8	4.2E−2±8E−6	4.2E−2±4E−5	4.2E−2±1E−5	4.2E−2±5E−5	4.2E−2±1E−5
q_9	6.5E−3±4E−6	6.5E−3±5E−6	6.5E−3±2E−6	6.5E−3±8E−6	6.5E−3±7E−6
q_{10}	3.3E−2±7E−6	3.3E−2±2E−5	3.3E−2±9E−6	3.3E−2±2E−5	3.3E−2±8E−6
q_{11}	4.7E−2±1E−3	4.9E−2±3E−3	4.7E−2±4E−4	4.9E−2±3E−3	4.7E−2±6E−4
q_{12}	6.3E−2±4E−4	6.3E−2±6E−4	6.3E−2±4E−4	6.3E−2±6E−4	6.3E−2±2E−4

Table 6 shows the behaviour of our spread-of-prediction intervals (Eq. 31) for both the Wang model and the Beattie model as described in Sect. 3.2. Here, we see that the average width of this interval (averaged over the length of the protocol) is much larger for the Beattie model (a discrepant model) when compared with the Wang model (the same model used for data generation).

Table 6 A summary showing how the spread-of-predictions interval (Eq. 31) behaves under Case II (Sect. 3.2), for both the Beattie model (a discrepant model) and the Wang model (a correctly specified model)

Model	Mean interval width (nA)	DGP in interval (%)	Midpoint RMSE (nA)
Beattie	7.5E−2±9E−5	34±7E−2	1.1E−01±8E−5
Wang	7.0E−4±2E−4	87±20	3.0E−2±2E−5

The columns show: the mean width of the interval (averaged over each observation time); the proportion of observations for which the the underlying DGP (minus noise) lies within the interval; the RMSE between the data and the midpoint prediction (Eq. 32). By considering ten randomly sampled datasets (each containing a repeat each protocol d_1–d_5), we show the mean and standard deviation of these values

References

Akaike H (1998) Information theory and an extension of the maximum likelihood principle. In: Parzen E, Tanabe K, Kitagawa G (eds) Selected papers of Hirotugu Akaike. Springer series in statistics. Springer, New York, pp 199–213. https://doi.org/10.1007/978-1-4612-1694-0_15

Anderson EW, Ghysels E, Juergens JL (2009) The impact of risk and uncertainty on expected returns. J Financ Econ 94(2):233–263. https://doi.org/10.1016/j.jfineco.2008.11.001

Bates DM (1988) Nonlinear regression analysis and its applications. In: Bates DM, Watts DG (eds). Wiley, New York

Beattie K (2015) Mathematical modelling of drug-ion channel interactions for cardiac safety assessment. PhD thesis, University of Oxford, https://ora.ox.ac.uk/objects/uuid:b6da189b-9495-4efb-be97-548fde5b1a79

Beattie KA, Hill AP, Bardenet R et al (2018) Sinusoidal voltage protocols for rapid characterisation of ion channel kinetics. J Physiol 596(10):1813–1828. https://doi.org/10.1113/JP275733

Beven K (2006) A manifesto for the equifinality thesis. J Hydrol 320(1):18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007

Chen S, Shojaie A, Witten DM (2017) Network reconstruction from high-dimensional ordinary differential equations. J Am Stat Assoc 112(520):1697–1707. https://doi.org/10.1080/01621459.2016.1229197

Clerx M, Beattie KA, Gavaghan DJ et al (2019a) Four ways to fit an ion channel model. Biophys J 117(12):2420–2437. https://doi.org/10.1016/j.bpj.2019.08.001

Clerx M, Robinson M, Lambert B et al (2019b) Probabilistic inference on noisy time series (PINTS). J Open Res Softw 7(1):23. https://doi.org/10.5334/jors.252

Creswell R, Lambert B, Lei CL et al (2020) Using flexible noise models to avoid noise model misspecification in inference of differential equation time series models. arXiv preprint arXiv:2011.04854

Creswell R, Robinson M, Gavaghan D et al (2023) A Bayesian nonparametric method for detecting rapid changes in disease transmission. J Theor Biol 558(111):351. https://doi.org/10.1016/j.jtbi.2022.111351

Fink M, Noble D (2009) Markov models for ion channels: versatility versus identifiability and speed. Philos Trans R Soc A: Math Phys Eng Sci 367(1896):2161–2179. https://doi.org/10.1098/rsta.2008.0301

Frazier DT, Robert CP, Rousseau J (2020) Model misspecification in approximate Bayesian computation: consequences and diagnostics. J R Stat Soc Ser B (Stat Methodol) 82(2):421–444. https://doi.org/10.1111/rssb.12356

Gelman A, Carlin JB, Stern HS et al (2013) Bayesian data analysis. CRC Press, Boca Raton

Goldwyn JH, Imennov NS, Famulare M et al (2011) Stochastic differential equation models for ion channel noise in Hodgkin-Huxley neurons. Phys Rev E 83(041):908. https://doi.org/10.1103/PhysRevE.83.041908

Guan J, Wei Y, Zhao Y et al (2020) Modeling the transmission dynamics of COVID-19 epidemic: a systematic review. J Biomed Res 34(6):422–430. https://doi.org/10.7555/JBR.34.20200119

Hansen N (2006) The CMA evolution strategy: a comparing review. In: Lozano JA, Larrañaga P, Inza I et al (eds) Towards a new evolutionary computation: advances in the estimation of distribution algorithms. Springer, Heidelberg, pp 75–102. https://doi.org/10.1007/3-540-32494-1_4

Harris CR, Millman KJ, Van Der Walt SJ et al (2020) Array programming with numpy. Nature 585(7825):357–362. https://doi.org/10.1038/s41586-020-2649-2

Hille B (2001) Ion channels of excitable membranes, 3rd edn. Sinauer, Sunderland

Johnstone RH, Chang ETY, Bardenet R et al (2016) Uncertainty and variability in models of the cardiac action potential: can we build trustworthy models? J Mol Cell Cardiol 96:49–62. https://doi.org/10.1016/j.yjmcc.2015.11.018

Keizer J (1972) On the solutions and the steady states of a master equation. J Stat Phys 6:67–72

Kemp JM, Whittaker DG, Venkateshappa R et al (2021) Electrophysiological characterization of the hERG R56Q LQTS variant and targeted rescue by the activator RPR260243. J Gen Physiol 153(10):e202112,923. https://doi.org/10.1085/jgp.202112923

Kennedy MC, O'Hagan A (2001) Bayesian calibration of computer models. J R Stat Soc Ser B (Stat Methodol) 63(3):425–464. https://doi.org/10.1111/1467-9868.00294

Klimovskaia A, Ganscha S, Claassen M (2016) Sparse Regression Based Structure Learning of Stochastic Reaction Networks from Single Cell Snapshot Time Series. PLoS Comput Biol 12(12):e1005,234. https://doi.org/10.1371/journal.pcbi.1005234

Lambert B, Lei CL, Robinson M et al (2022) Autocorrelated measurement processes and inference for ordinary differential equation models of biological systems. arXiv preprint arXiv:2210.01592 https://doi.org/10.48550/arXiv.2210.01592

Lei CL, Mirams GR (2021) Neural network differential equations for ion channel modelling. Front Physiol 12:1166. https://doi.org/10.3389/fphys.2021.708944

Lei CL, Clerx M, Beattie KA et al (2019a) Rapid characterization of hERG channel kinetics II: temperature dependence. Biophys J 117(12):2455–2470. https://doi.org/10.1016/j.bpj.2019.07.030

Lei CL, Clerx M, Gavaghan DJ et al (2019b) Rapid characterization of hERG channel kinetics I: using an automated high-throughput system. Biophys J 117(12):2438–2454. https://doi.org/10.1016/j.bpj.2019.07.029

Lei CL, Clerx M, Whittaker DG et al (2020a) Accounting for variability in ion current recordings using a mathematical model of artefacts in voltage-clamp experiments. Philos Trans R Soc A: Math Phys Eng Sci 378(2173):20190,348. https://doi.org/10.1098/rsta.2019.0348

Lei CL, Ghosh S, Whittaker DG et al (2020b) Considering discrepancy when calibrating a mechanistic electrophysiology model. Philos Trans R Soc A: Math Phys Eng Sci 378(2173):20190,349. https://doi.org/10.1098/rsta.2019.0349

Lei CL, Clerx M, Gavaghan DJ et al (2022) Model-driven optimal experimental design for calibrating cardiac electrophysiology models. bioRxiv https://doi.org/10.1101/2022.11.01.514669

Li Z, Dutta S, Sheng J et al (2017) Improving the In Silico Assessment of Proarrhythmia Risk by Combining hERG (Human Ether-à-go-go-Related Gene) Channel-Drug Binding Kinetics and Multichannel Pharmacology. Circ Arrhythm Electrophysiol 10(2):e004,628. https://doi.org/10.1161/CIRCEP.116.004628

Mangold KE, Wang W, Johnson EK et al (2021) Identification of structures for ion channel kinetic models. PLoS Comput Biol 17(8):e1008,932. https://doi.org/10.1371/journal.pcbi.1008932

Mirams GR, Pathmanathan P, Gray RA et al (2016) Uncertainty and variability in computational and mathematical models of cardiac physiology. J Physiol 594(23):6833–6847. https://doi.org/10.1113/JP271671

Rudy Y, Silva JR (2006) Computational biology in the study of cardiac ion channels and cell electrophysiology. Q Rev Biophys 39(1):57–116. https://doi.org/10.1017/S0033583506004227

Seber G, Wild C (2005) Nonlinear regression. Wiley series in probability and statistics. Wiley, New York

Smith RC (2013) Uncertainty quantification: theory, implementation, and applications, vol 12. SIAM, Philadelphia

Sung CL, Barber BD, Walker BJ (2020) Calibration of inexact computer models with heteroscedastic errors. arXiv:1910.11518

ten Tusscher KHWJ, Noble D, Noble PJ et al (2004) A model for human ventricular tissue. Am J Physiol Heart Circul Physiol 286(4):H1573–H1589. https://doi.org/10.1152/ajpheart.00794.2003

Wang S, Liu S, Morales MJ et al (1997) A quantitative analysis of the activation and inactivation kinetics of HERG expressed in Xenopus oocytes. J Physiol 502(1):45–60. https://doi.org/10.1111/j.1469-7793.1997.045bl.x

Whittaker DG, Clerx M, Lei CL et al (2020) Calibration of ionic and cellular cardiac electrophysiology models. WIREs Syst Biol Med 12(4):e1482. https://doi.org/10.1002/wsbm.1482

Wieland FG, Hauber AL, Rosenblatt M et al (2021) On structural and practical identifiability. Curr Opin Syst Biol 25:60–69. https://doi.org/10.1016/j.coisb.2021.03.005
Willmott CJ, Ackleson SG, Davis RE et al (1985) Statistics for the evaluation and comparison of models. J Geophys Res Oceans 90(C5):8995–9005. https://doi.org/10.1029/JC090iC05p08995