

# Hierarchical Bayesian Modelling of Variability and Uncertainty in Synthetic Action Potential Traces

Ross H Johnstone<sup>1</sup>, Rémi Bardenet<sup>2</sup>, David J Gavaghan<sup>1</sup>,  
Liudmila Polonchuk<sup>3</sup>, Mark R Davies<sup>3</sup>, Gary R Mirams<sup>1</sup>

<sup>1</sup> Computational Biology, Dept. of Computer Science, University of Oxford, Oxford, UK

<sup>2</sup> CNRS & CRISAL, Université de Lille, Lille, France

<sup>3</sup> Roche Innovation Center Basel, Switzerland

## Abstract

*There are many sources of uncertainty in the recording and modelling of membrane action potentials (APs) from cardiomyocytes. For example, there are measurement, parameter, and model uncertainties. There is also extrinsic variability between cells, and intrinsic beat-to-beat variability within a single cell. These combined uncertainties and variability make it very difficult to extrapolate predictions from these models, since current AP models have single parameter values and thus produce a single AP trace. We aim to re-parameterise existing AP models to fit experimental data, and to quantify uncertainty associated with ion current densities when measuring and modelling these APs. We then wish to propagate this uncertainty into model predictions, such as ion channel block effected by a pharmaceutical compound.*

*We perform an in silico study using synthetic data generated from different sets of parameters. We then ‘forget’ these parameter values and re-infer their distributions using hierarchical Markov chain Monte Carlo methods. We find that we can successfully infer the ‘correct’ distributions for most ion current densities for each AP trace, along with an approximation to the higher-level distribution from which these parameter values were sampled. There is, however, some level of unidentifiability amongst some of the current densities.*

## 1. Introduction

Mathematical models of single-cell electrophysiology are used in cardiac safety to predict effects of pharmaceutical compounds that block multiple ion channels. Data from screening with mathematical models of cardiac action potentials can help us to understand what combined effect a drug will have on overall cellular electrophysiology [1]. Moreover, a single cell expresses beat-to-beat variability when measuring action potentials (APs), which

adds another layer of variability onto making predictions.

In particular, stem cell-derived cardiomyocytes [2] have been proposed for use in pharmaceutical safety testing, but different batches of these cells show different ion channel expression, which will lead to different responses to compounds. The first mathematical model of these cells was developed by Paci *et al.* [3]. Tailoring models to individual cells or batches of cells should allow greater predictive power when considering compound effects on ion channels. Treating model parameters as random variables and using a hierarchical Bayesian approach to infer probability distributions instead of single best-fit point-estimates, the uncertainties associated with beat-to-beat variability, model fitting, and model predictions can be quantified and propagated to obtain ranges of possible predicted outcomes.

A general equation for an AP model is

$$C_m \frac{dV}{dt} = - \left( \sum_i G_i O_i (V - E_i) + I_{stim} \right), \quad (1)$$

where  $V$  is the membrane voltage,  $C_m$  is the membrane capacitance, the  $G_i$  are the maximal ion current conductances, the  $O_i$  are the open probabilities for the ion channels, the  $E_i$  are the ionic reversal potentials, and  $I_{stim}$  is an externally applied stimulus current.  $i$  indexes the various ion currents that are included in the model.

These ion channels are generally the targets or off-targets of drugs that are of concern in terms of cardiac safety. We therefore treat the  $G_i$  as the parameters of interest. We can model drug effects by scaling the  $G_i$ , since decreasing a  $G_i$  represents a blocking of that ion channel, so that less of that ion current is flowing across the cell membrane.

## 2. Methods

### 2.1. Statistical model

We work in a Bayesian framework, where we treat the parameters of interest as random variables. Then, given some data, we infer probability distributions, called the *posterior distribution* for these parameters. We are effectively finding all sets of parameter values which allow us to fit our model to the data, while assigning these sets different probability densities describing how good a fit they provide.

Suppose that the voltage trace produced by a model given by Equation (1) is  $\mathbf{y} = f(\mathbf{G})$ , where  $\mathbf{G}$  is a vector of maximal current densities parameters, and  $\mathbf{y}$  is the computed voltage trace. We model one dataset as being an observation of the true system behaviour, described by  $f(\mathbf{G})$ , plus some Normal random noise with standard deviation  $\sigma$ .

However, we wish to capture inter-experiment or beat-to-beat variability. We therefore treat each dataset as a realisation of the underlying model, but solved with different sets of parameter values,  $\mathbf{G}_i$ , where  $i = 1, \dots, N_e$  and  $N_e$  is the number of datasets, or experiments. We use a hierarchical statistical model to describe these sets of parameters as being related to each other through some overarching Normal distribution with mean  $\hat{\mathbf{G}}$  and independent standard deviations  $\hat{\sigma}$ .

We therefore have

$$\mathbf{y}_i \sim \mathcal{N}(f(\mathbf{G}_i), \sigma^2), \quad (2)$$

$$\mathbf{G}_i \sim \mathcal{N}(\hat{\mathbf{G}}, \hat{\sigma}^2), \quad (3)$$

for  $i = 1, \dots, N_e$ . A schematic of this hierarchical model is given in Figure 1.

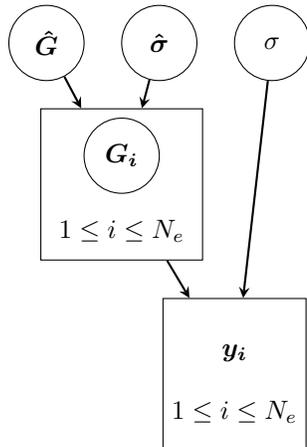


Figure 1. All circled variables are parameters for which we wish to infer probability distributions.

To quantify the observational uncertainty and inter-experiment variability, we want to infer probability distributions for all  $\mathbf{G}_i$ ,  $\hat{\mathbf{G}}$ ,  $\hat{\sigma}$ , and  $\sigma$ . Given these distributions, we can construct posterior predictive distributions for  $\mathbf{G}_i$  which will allow us to make predictions for how a cell will behave in a future experiment.

### 2.2. Synthetic data

We use synthetic data as a way to eliminate ‘model uncertainty’. That is, we work in a framework in which we have a ‘perfect’, or ‘correct’, model. This is a best-case scenario which we will never have when using real experimental data, but it provides a starting-point from which to develop our analysis.

We use synthetic data generated from the Paci et al. model from the CellML repository [4], which has 12 ion currents. We attempt to infer distributions for the maximal densities for all 12 of these currents. We define the ‘top-level’ distribution, *i.e.* the distribution from which the  $\mathbf{G}_i$  will be drawn, as a Normal distribution whose mean is the published set of maximal conductance values, and standard deviations of  $0.1 \times$  those values. The synthetic data is generated by solving the model equations in Chaste [5], given a set of maximal conductance parameter values drawn from the pre-defined distribution, and then adding random observation noise to every data point. This noise was Normally distributed with standard deviation 0.25 mV. This value was chosen by computing the standard deviation of a relatively flat region of a real experimental voltage trace from a canine cardiomyocyte, as in [6]. These synthetic datasets are plotted for  $N_e = 5$  and  $N_e = 10$  in Figure 2.

### 2.3. Parameter inference

The posterior distribution,  $p(\boldsymbol{\theta}|\text{data})$ , is defined using Bayes’ Theorem:

$$p(\boldsymbol{\theta}|\text{data}) = \frac{p(\text{data}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\text{data}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}, \quad (4)$$

where  $p(\text{data}|\boldsymbol{\theta})$  is the likelihood of the observed data given parameters  $\boldsymbol{\theta}$  and  $p(\boldsymbol{\theta})$  is the *prior* distribution of the parameters. The prior distribution contains our prior knowledge or belief about the parameters before observing any data.

The integral in the denominator of Equation (4) is generally intractable, so we use Markov chain Monte Carlo (MCMC) methods [7] to approximate  $p(\boldsymbol{\theta}|\text{data})$ . MCMC methods only require that we know the posterior distribution up to a factor of a constant, so it is enough to have that

$$p(\boldsymbol{\theta}|\text{data}) \propto p(\text{data}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \quad (5)$$

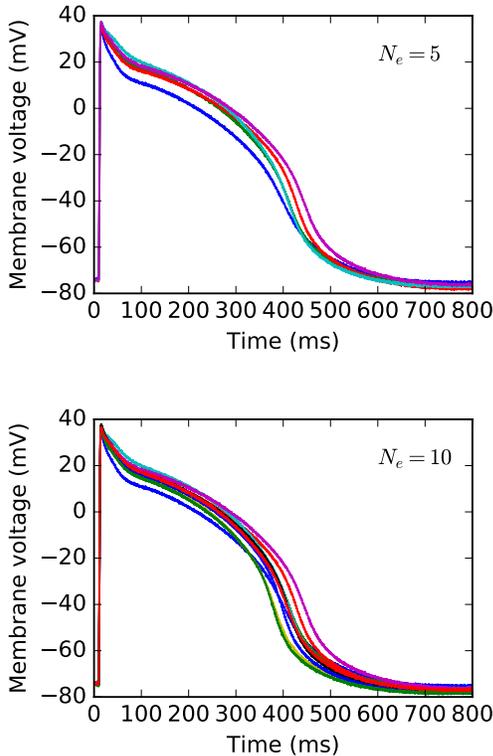


Figure 2. Synthetic voltage traces generated for  $N_e = 5$  (top) and  $N_e = 10$  (bottom).

to allow us to construct an approximation to the posterior distribution.

We ‘forget’ the parameter values used to generate the synthetic data and use an adaptive Metropolis-within-Gibbs [8] MCMC method to infer distributions for all  $G_i$ ,  $\hat{G}$ ,  $\hat{\sigma}$ , and  $\sigma$ . Suppose there are  $d$  parameters in the mathematical model, *i.e.*  $\mathbf{G} = (G_1, G_2, \dots, G_d)$ , then there are a total of  $(N_e + 2)d + 1$  parameters in the statistical model. We attempt to infer distributions for all  $(N_e + 2)d + 1$  parameters using MCMC. In this example, the Paci *et al.* model has 12 ion currents, so  $d = 12$  here. We first have to specify prior distributions over the ‘top-level’ parameters,  $\hat{G}$ ,  $\hat{\sigma}$ , and  $\sigma$ . We choose a uniform prior for  $\sigma$  which is wide enough to capture any plausible values of the observational noise standard deviation. We use a Normal-inverse-gamma distribution as the prior distribution over  $\hat{G}$  and  $\hat{\sigma}$ , which allows us to use *conjugate prior* [7] properties to speed up the MCMC sampling.

The aim is to identify how much information is present about the parameters in each individual trace as well as how they are related across experiments. This should allow us to quantify inter-experiment variability and provide ranges of predictions for future experiments along with associated probabilities.

### 3. Results

We use the mean absolute percentage error (MAPE) as an estimator for a ‘successful’ inference of a parameter. For the intra-experimental parameters  $G_i$ , we say that we have successfully inferred a parameter when its MAPE is less than 0.05, *i.e.* when, on average, the samples are within 5% of the value used to generate that voltage trace.

We are able to infer tight distributions for some parameters, but not for all of them. For example, in the case where  $N_e = 5$ , we infer a relatively tight distribution for  $G_{Na}$  around the values used to generate the synthetic traces. However, for other parameters, the inferred distributions do not capture the original value. Figure 3 shows inferred marginal distributions for  $G_{Na}$  and  $G_{Kr}$  for a single AP trace. In the bottom plot, the  $G_{Kr}$  distribution is to the left of the ‘true’ value since the ‘true’ values for the other 4 traces are also to the left and so the information shared across the hierarchical model places more probability mass to the left of the red line.

This tells us that that particular voltage trace contains much more information about  $G_{Na}$  than about  $G_{Kr}$ . We find that, of the 12 conductance parameters we wish to infer per experiment, we can consistently recover 5 of them, and can recover up to 10 for different experiments. This suggests that we use a more complicated experimental protocol to generate a dataset with more information about specific parameters.

We can construct the posterior predictive distributions by summing the Normal distributions defined by our samples of  $\hat{G}$  and  $\hat{\sigma}$  at every iteration in the MCMC and then normalising. This approximates the over-arching distribution which governs how the  $G_i$  are distributed. To make predictions of how a cell will behave in a future experiment, we would sample values of  $\mathbf{G}$  from this distribution many times to construct a range of predicted AP traces. The posterior predictive distributions for the cases where  $N_e = 5$  and  $N_e = 10$  are plotted in Figure 4, along with the ‘true’ distribution we are trying to approximate. Increasing the number of datasets being fit to increases the overall information content, so we can better approximate the true distribution.

### 4. Conclusions

Some of the ion current densities in AP models could be tailored to the individual cells upon which the experiments are performed. Using a Bayesian framework allows us to quantify and propagate uncertainties introduced through observational, residual, and input uncertainty [9]. In the case of multiple cells, or beat-to-beat variability, we can treat each AP as a separate dataset and using hierarchical MCMC methods to fit the model to the data simultaneously, while sharing information among the traces. In-

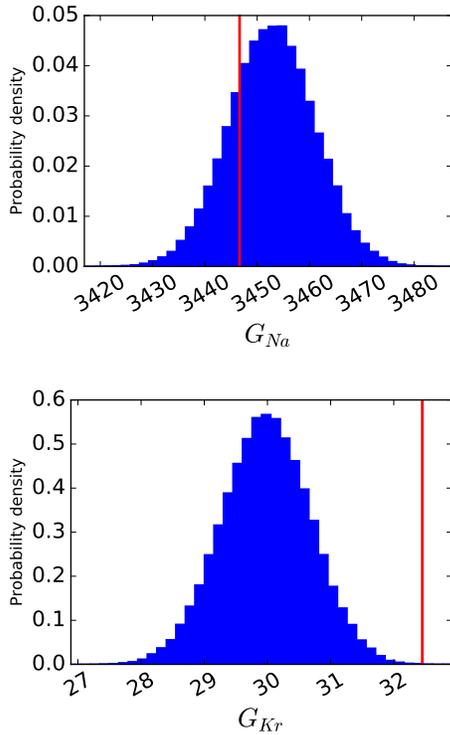


Figure 3. Normalised marginal histograms for  $G_{Na}$  and  $G_{Kr}$  inferred for a single AP trace when  $N_e = 5$ . The vertical red lines are the ‘true’ parameter values used to generate the synthetic data trace.

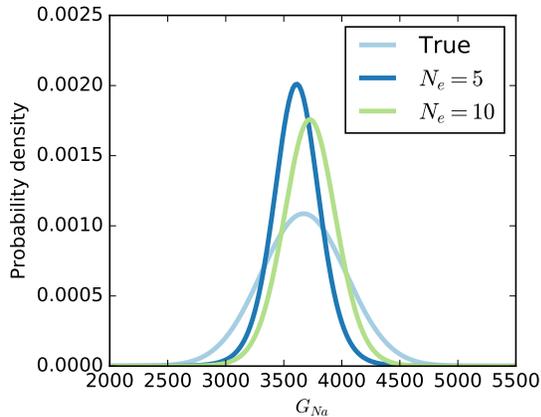


Figure 4. The ‘true’ distribution governing the  $G_i$  and the posterior predictive distributions for the  $G_i$  in the cases where  $N_e = 5$  and  $N_e = 10$ . Increasing the number of datasets fit to allows us to better approximate the true distribution.

creasing the number of datasets being fit to is important for increasing accuracy of predictions, but does increase computational cost.

## Acknowledgements

RHJ is supported by an EPSRC Systems Approaches to Biomedical Science Industrial Doctorate Centre studentship and by Roche.

GRM acknowledges the Royal Society and the Wellcome Trust for a Sir Henry Dale Fellowship.

## References

- [1] Mirams GR, Cui Y, Sher A, Fink M, Cooper J, Heath BM, McMahon NC, Gavaghan DJ, Noble D. Simulation of multiple ion channel block provides improved early prediction of compounds clinical torsadogenic risk. *Cardiovascular research* 2011;91(1):53–61.
- [2] Tanaka T, Tohyama S, Murata M, Nomura F, Kaneko T, Chen H, Hattori F, Egashira T, Seki T, Ohno Y, et al. In vitro pharmacologic testing using human induced pluripotent stem cell-derived cardiomyocytes. *Biochemical and biophysical research communications* 2009;385(4):497–502.
- [3] Paci M, Hyttinen J, Aalto-Setälä K, Severi S. Computational models of ventricular-and atrial-like human induced pluripotent stem cell derived cardiomyocytes. *Annals of biomedical engineering* 2013;41(11):2334–2348.
- [4] Lloyd CM, Lawson JR, Hunter PJ, Nielsen PF. The cellml model repository. *Bioinformatics* 2008;24(18):2122–2123.
- [5] Mirams GR, Arthurs CJ, Bernabeu MO, Bordas R, Cooper J, Corrias A, Davit Y, Dunn SJ, Fletcher AG, Harvey DG, et al. Chaste: an open source c++ library for computational physiology and biology. *PLoS Comput Biol* 2013;9(3):e1002970.
- [6] Johnstone RH, Chang ET, Bardenet R, De Boer TP, Gavaghan DJ, Pathmanathan P, Clayton RH, Mirams GR. Uncertainty and variability in models of the cardiac action potential: Can we build trustworthy models? *Journal of molecular and cellular cardiology* 2016;96:49–62.
- [7] Robert C, Casella G. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.
- [8] Gilks WR, Best N, Tan K. Adaptive rejection metropolis sampling within gibbs sampling. *Applied Statistics* 1995; 455–472.
- [9] Mirams GR, Pathmanathan P, Gray RA, Challenor P, Clayton RH. White paper: Uncertainty and variability in computational and mathematical models of cardiac physiology. *The Journal of physiology* 2016;.

Address for correspondence:

Ross H. Johnstone  
 Computational Biology, Department of Computer Science,  
 University of Oxford, Wolfson Building, Parks Road, Oxford,  
 OX1 3QD, UK  
 ross.johnstone@cs.ox.ac.uk