Check for updates

SOFTWARE TOOL ARTICLE

REVISED Hierarchical Bayesian inference for ion channel screening dose-response data [version 2; referees: 2 approved]

Ross H Johnstone[1], Rémi Bardenet[2], David J Gavaghan[1], Gary R Mirams[1,3]

[1]Computational Biology, Department of Computer Science, University of Oxford, Oxford, UK
[2]CNRS & CRIStAL, Université de Lille, Lille, France
[3]Centre for Mathematical Medicine & Biology, School of Mathematical Sciences, University of Nottingham, Nottingham, UK

Abstract

Dose-response (or 'concentration-effect') relationships commonly occur in biological and pharmacological systems and are well characterised by Hill curves. These curves are described by an equation with two parameters: the inhibitory concentration 50% (IC50); and the Hill coefficient. Typically just the 'best fit' parameter values are reported in the literature. Here we introduce a Python-based software tool, **_PyHillFit_** , and describe the underlying Bayesian inference methods that it uses, to infer probability distributions for these parameters as well as the level of experimental observation noise. The tool also allows for hierarchical fitting, characterising the effect of inter-experiment variability. We demonstrate the use of the tool on a recently published dataset on multiple ion channel inhibition by multiple drug compounds. We compare the maximum likelihood, Bayesian and hierarchical Bayesian approaches. We then show how uncertainty in dose-response inputs can be characterised and propagated into a cardiac action potential simulation to give a probability distribution on model outputs.

Open Peer Review

Referee Status: ✔ ✔

	Invited Referees	
	1	**2**
REVISED **version 2** published 13 Mar 2017		
version 1 published 15 Nov 2016	✔ report	✔ report

1 **Jonathan Stott**, AstraZeneca UK, **Stanley Lazic**, AstraZeneca UK

2 **Derek Leishman**, Eli Lilly and Company USA

Discuss this article

Comments (0)

Competing interests: No competing interests were disclosed.

> **REVISED** **Amendments from Version 1**
>
> Version 2 of the manuscript contains four improved figures, and a completely new version of Figure 15. Text amended for clarity in various places thanks to comments from reviewers.
>
> **See referee reports**

1 Introduction

In this article, we describe the approach our software tool takes to inferring parameters of dose-response curves from experimental data. This introduction addresses the problem, standard approach, and our motivation for developing an approach that also characterises uncertainty in dose-response parameters, due to variability in the data.

1.1 Dose-response curves

'Dose-response' (or 'concentration-effect') curve-fitting is generally performed to describe how increasing compound concentration provokes a response in a process. 'Dose-response' generally relates to *in-vivo* experiments where a drug dose is administered but the concentration at the relevant site is not precisely known; whereas 'concentration-effect' generally relates to *in-vitro* experiments where the concentration is accurately applied. We will refer to both as 'dose-response' in this article for simplicity. In the case study we will pursue, the 'response' is binding and blocking of ion channels, measured via inhibition of ion currents. Dose-response curves are summarised by two parameters: an inhibitory concentration 50% (IC50) value, that is the concentration of the compound that gives 50% of the maximum effect; and a Hill coefficient, which sets the 'steepness' of the curve as it passes the IC50. Examples of dose-response data and a fitted curve are given in Figure 1.
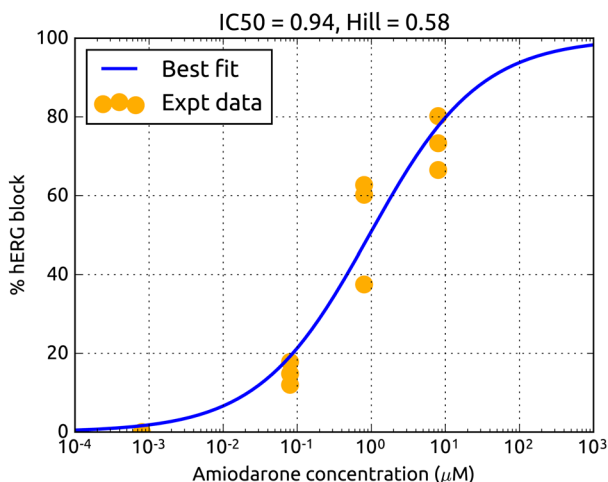

Figure 1. A dose-response curve fitted to experimental data for hERG block by amiodarone from the Crumb *et al.* (2016) dataset. A dose-response curve is shown fitted to all data points at once with a least-square-differences algorithm, with the resulting parameters shown at the top of the graph.

The equation for a dose-response curve was proposed by Hill (1910), and subsequently another name for the curve is a *Hill curve* (Weiss, 1997). If we let x be the concentration of a compound, we describe the effect of the compound by

$$\text{response} = f_{DR}(x; IC50, Hill) := \frac{100}{1 + \left(\dfrac{IC50}{x}\right)^{Hill}}, \qquad (1)$$

where IC50 and *Hill* are parameters that take positive values. In our motivating example this response will be "% block" of a particular type of ion channel.

1.2 Standard fitting procedure

A Hill curve is often fitted to all data points simultaneously, to obtain 'average' IC50 values and Hill coefficients for a particular curve. This gives the most likely set of parameter values. For example, Crumb *et al.* (2016) recently published dose-response screening data for 30 compounds on 7 different ion channels, along with best-fit IC50 values and Hill coefficients. But taking this approach, there is no associated probability given to these IC50 and Hill values; different possible ranges for these parameter values are not considered. The usual fitting procedure can also give rise to models which differ in behaviour from each individual experiment, as shown by Pathmanathan *et al.* (2015) in the case of inactivation of the fast sodium current in action potential models, and as we will show in the case studies below.

1.3 Variability and uncertainty

Real-world experiments exhibit *intrinsic* and *extrinsic* variability. The characterisation of this variability is becoming of greater importance as we move to quantitatively predictive models, particularly as part of the global cardiac modelling effort (Johnstone *et al.*, 2016a; Mirams *et al.*, 2016; Pathmanathan *et al.*, 2015). Intrinsic variability describes fluctuations that may be due to inherent randomness, and extrinsic variability describes differences between individuals (in this case cells/experiments). Variability contributes to uncertainty, but there are other sources of uncertainty when modelling and performing experiments (Vernon *et al.* (2010) provide a good introduction). There will also be observation error, which arises from imperfect measurements, thus introducing more uncertainty.

If we are going to use a model to predict future behaviour, or infer some underlying behaviour, we want to study the impact of uncertainty and give probabilistic predictions. Here, if there is a distribution of parameter values that could have given rise to the experimental dose-response data, we would like to capture this distribution (*uncertainty characterisation*). When using these data as inputs to further simulations (as discussed below in Section 6) we would then construct a distribution of possible outputs corresponding to the distribution of inputs, a process known as *uncertainty propagation*. The whole process is known as *uncertainty quantification*, or *UQ* (US National Research Council, 2012), and is part of a framework for ensuring safety-critical simulations are reliable which is known as validation, verification and uncertainty quantification, or *VVUQ* (Pathmanathan & Gray, 2013).

1.4 Motivation

As part of the *Comprehensive in vitro Pro-arrhythmia Assay* initiative (CiPA, Fermini *et al.*, 2016; Sager *et al.*, 2014), it is proposed that pharmaceutical compounds be tested on up to 7 ion channels that both strongly influence ventricular repolarisation and are frequently blocked by pharmaceutical compounds. The proposal is for ion channel screening data to be passed into an *in silico* human ventricular action potential model to see if the compound produces pro-arrhythmic behaviour, or indicators of such behaviour, at the whole-cell level.

In this article we present a software tool to infer distributions of possible dose-response curves from experimental data. When making predictions of block at a given concentration, these distributions of possible dose-response curves provide us with a probability distribution for the level of block at that concentration. To illustrate the consequences of this, we use the distributions of block (for multiple ion channels) inferred from data provided in Crumb *et al.* (2016) as inputs into an *in silico* action potential model. We then run forward simulations to predict a distribution of outputs — in this case action potential durations resulting from application of a compound at a particular concentration. We show that given the limited number of repeats of ion-channel experiments, there are wide ranges of predicted action potentials, with overlapping results for different compounds with different associated pro-arrhythmic risks.

1.5 Our approach

To explore and characterise the uncertainty in the dose-response measurements published by Crumb *et al.*, and to propagate these uncertainties into model predictions, we use a Bayesian statistical framework to explore different possible dose-response curves that may have produced these data. Each dose-response curve is assigned a likelihood score, which, roughly speaking, describes how well the curve fits the data. Instead of computing point-estimates for the IC50 value and Hill coefficient, we infer probability distributions of these parameters, as well as a distribution for the possible observational noise. This provides us with a method for propagating uncertainty in experimental data into simulations by drawing parameters from these inferred distributions, and using these samples as simulation inputs.

We describe two different types of Bayesian statistical models: one where all data points are treated equally (as though they were obtained from the same experiment); and another where we believe that each repeat of an experiment has distinct properties (through some source of inter-experiment variability) and therefore its own set of parameters to infer. The first case we will refer to as 'single-level', and the second case as 'hierarchical'. The single-level case does not consider extrinsic variability, since we are assuming that all data points are generated by the same behaviour. The hierarchical case does consider extrinsic variability, which we model by assuming that each experimental dataset was generated according to its own IC50 value and Hill coefficient, which may vary across experiments.

2 Bayesian statistical modelling approach

We use a Bayesian framework to quantify the uncertainty present in the ion channel screening data. The tool reads in doses in μM,

but instead of working with IC50 in μM, we work with pIC50, where

$$pIC50\left[\log(\text{Molar})\right] = 6 - \log_{10}\left(IC50\left[\mu\text{Molar}\right]\right), \quad (2)$$

with square brackets indicating units. This transformation makes it much easier for fitting algorithms to explore the parameter space, as linear variation in IC50s does not result in linear changes to dose-response curves, which are commonly plotted on log scales. The dose-response model we therefore work with in practice is

$$\% \text{ channel block} = f\left(x; pIC50, Hill\right) := f_{DR}\left(x; 10^{(6-pIC50)}, Hill\right). \quad (3)$$

So we assume that the underlying behaviour is described by the dose-response model, *f*, given by Equation (3). We assume that an experimental observation is Normally distributed around some underlying behaviour with some standard deviation, σ (that has the same units as the measured response). That is, given an applied compound concentration, *x*, our statistical model is that the response, *y*, is a Normally-distributed random variable with mean $f(x; pIC50, Hill)$ and standard deviation σ, that is:

$$y \sim \mathcal{N}\left(f\left(x; pIC50, Hill\right), \sigma^2\right). \quad (4)$$

When we have noisy data, different sets of parameters might allow us to fit the equation to the experimental data equally well. In our Bayesian framework, we treat these model parameters as random variables, in part due to the uncertainty introduced through observational error and any parameter identifiability problems (see e.g. Daly *et al.* (2015); Raue *et al.* (2009); Siekmann *et al.* (2012) for a discussion of how identifiability relates to inferred probability distributions). We therefore want to infer a probability distribution, instead of point-estimates, for the parameters *pIC*50, *Hill* and σ. This probability distribution, $p(\theta|data)$, is the *posterior* distribution of the parameters, θ, given the observed experimental data.

The posterior distribution is defined using Bayes' Theorem:

$$p(\theta \mid data) = \frac{p(data \mid \theta) p(\theta)}{\int_\theta p(data \mid \theta) p(\theta) \, d\theta}, \quad (5)$$

where $p(data|\theta)$ is the likelihood of the parameters θ under our model given the observed data **y**, and $p(\theta)$ is the *prior* distribution of the parameters θ. The prior distribution contains our prior knowledge or belief about the parameters before observing any data.

The integral in the denominator of Equation (5) is generally intractable, so we use Markov chain Monte Carlo (MCMC) methods to approximate $p(\theta|data)$. MCMC methods only require that we can evaluate the posterior distribution, pointwise, up to a factor of a constant, so it is enough to have that

$$p(\theta \mid data) \propto p(data \mid \theta) p(\theta), \quad (6)$$

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

3 Single-level model

In our example, each experiment consisted of applying one or more concentrations of a compound to a cell and measuring the degree of

block of an ion current. There were multiple recordings, leading to multiple response data-points at each concentration.

3.1 Methods

In this statistical model, we will assume that there is no inter-experiment variability, so all data points are (effectively) from one experiment, and all the data points are generated using the same set of parameter values. Under this model, the data for hERG block by amiodarone (Crumb *et al.*, 2016), for example, is generated by the process schematic shown in Figure 2.

Under our statistical model shown in Figure 2, all data points $y^{(j)}$ from a single drug and channel combination are identically independently distributed according to

$$y^{(j)} \sim \mathcal{N}\left(f\left(x^{(j)}; pIC50, Hill\right), \sigma^2\right), \tag{7}$$

where $x^{(j)}$ is the applied compound concentration, and $j = 1, \ldots, K$, where K is the total number of data points. The likelihood of the parameters $pIC50$, $Hill$, and σ, given a single data point $y^{(j)}$ is then

$$p\left(y^{(j)} \mid pIC50, Hill, \sigma\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y^{(j)} - f\left(x^{(j)}; pIC50, Hill\right)\right)^2}{2\sigma^2}\right). \tag{8}$$

In practice, we work with the log of the target distribution, and therefore the log of the likelihood given in Equation 8. For more details on the implementation, see Johnstone *et al.*, (2016a, Supplement B2).

In the data published by Crumb *et al.*, any observations below 0% or above 100% were capped to 0% or 100%, respectively (W. Crumb, personal communication), because these extreme values are assumed to be due to observational error. Accordingly, we truncate the Normal distribution in Equation (7) at 0 and 100, since these are imposed bounds on the data for % channel block (it would be better not to filter the data in this way, as, before making this adjustment, we observed repeated zero entries leading to the erroneous conclusion that there was almost no noise σ on the data).

Since $pIC50$, $Hill$, and σ are parameters that we infer, we need to specify a prior distribution across them, corresponding to $p(\theta)$ in Equation (5). We choose independent uniform distributions for each parameter, but we could have chosen a more informative prior based on previous ion channel screening data, where available. We allow $Hill$ to take values in (0,10). The Hill coefficient must be positive and describes the steepness of the dose-response curve, so after a certain point, increasing the Hill coefficient does not make a noticeable difference to the curve, so we choose 10 as a generous upper bound, above any biologically-plausible drug-binding we are aware of. Similarly, a compound that has no measurable effect could be thought of as having a very large IC50, and it makes no difference practically to model it as having an even larger IC50. This corresponds to a negative pIC50 value, and so we choose to

allow $pIC50$ to take values in (-1,15) as values outside this interval will not have much effect (see Figure 3). We let σ, which is a standard deviation parameter and therefore also positive, take values in (0,50), where 50 is a generous upper bound for observational error, which we expect to be closer to 5–10% in practice.

As described in our previous publication (Johnstone *et al.*, 2016a), we first perform a covariance matrix adaptation evolution strategy optimisation (CMA-ES, Hansen *et al.*, 2003) to find an optimal starting point for exploring possible parameter sets. We then use an adaptive Metropolis-Hastings MCMC algorithm (Haario *et al.*, 2001) to infer $p(\theta|data)$, where $\theta = \{pIC50, Hill, \sigma\}$. Briefly, we want to construct a sequence (Markov chain)
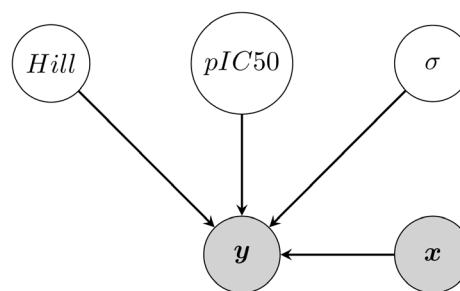

Figure 2. Statistical model for the generation of dose-response data *y*. All non-shaded variables are parameters for which we wish to infer probability distributions. **y** = $\{y^{(1)}, \ldots, y^{(K)}\}$ is the vector of all experimentally-recorded response data points.
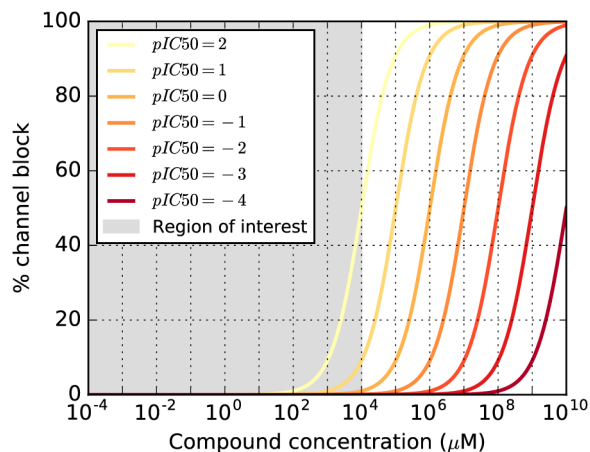

Figure 3. Large IC50 values are indistinguishable when they are orders of magnitude above the relevant concentration range. Here we show the effect of decreasing pIC50 (increasing IC50), while maintaining *Hill* = 1. The shaded "region of interest" covers the minimum and maximum concentrations in the data published by Crumb *et al.*. As pIC50 decreases, there is no significant change to the dose-response curve across the relevant range of concentrations — all predictions are close to zero response. As a result, we somewhat artificially 'cap' pIC50 priors to exclude pIC50 < −1, otherwise (for datasets with no response signal) convergence of minimisation and MCMC algorithms is difficult if not impossible.

of parameter-value sets that approximate samples from the left-hand-side of Equation (5). A set of parameter values is proposed by sampling from a multivariate Normal distribution centred on the most recent iteration's parameter values. The posterior distribution (Equation (6)) at these newly-proposed parameter values is then computed. The set of these proposed parameter values is accepted into the chain with a probability computed as the ratio of posterior distribution values between the current parameters and the proposed parameters. If the proposed set of parameter values is accepted, it is appended to the history of the chain and the next iteration will be taken from these new parameter values. As the MCMC algorithm runs, the covariance matrix of the proposal distribution skews in the directions where more sets of parameters are being accepted into the chain. After a large number of iterations, we discard the first quarter (or any suitably large fraction) of samples, known as 'burn-in' when the MCMC algorithm was still finding the best regions of parameter space. Then we plot normalised histograms of the remaining samples to approximate the posterior distribution.

The posterior distribution in the single-level model is given by (up to a factor of a constant)

$$p\left(pIC50, Hill, \sigma \mid \mathbf{y}\right) \propto \prod_{j=1}^{K} p\left(y^{(j)} \mid pIC50, Hill, \sigma\right) p\left(pIC50\right) p\left(Hill\right) p\left(\sigma\right), \quad (9)$$

where the first term on the right-hand-side is given by Equation (8).

3.2 Results

In Figure 4 we plot normalised marginal and pairwise histograms for the values of the parameters for each sample of the MCMC algorithm output, which are approximate projections of the posterior distribution across these parameters. The spread in each distribution corresponds to the uncertainty in that parameter; if a parameter's marginal posterior distribution is narrower, we are more certain about its value from the observed data.

Before propagating these uncertainties, we first draw ($pIC50$, $Hill$) samples from the MCMC output, and plot dose-response curves with these parameter values. Examples are given in Figure 5, where amiodarone has a measurable blocking effect on hERG, but no measurable effect on Kir2.1. As we take more samples, the plotted curves build up a distribution of possible dose-response curves given the experimental data. For each compound concentration, we then have a range of possible responses with their relative probability densities being given by the density of dose-response curves at that concentration.

4 Hierarchical (multi-level/mixture) model

When we plot the ion channel screening data and group the data points according to their respective experimental repeats, instead of treating them as data points from one experiment as in Section 3, we see that data points from the same experiment generally keep their relative position. That is, we often see that
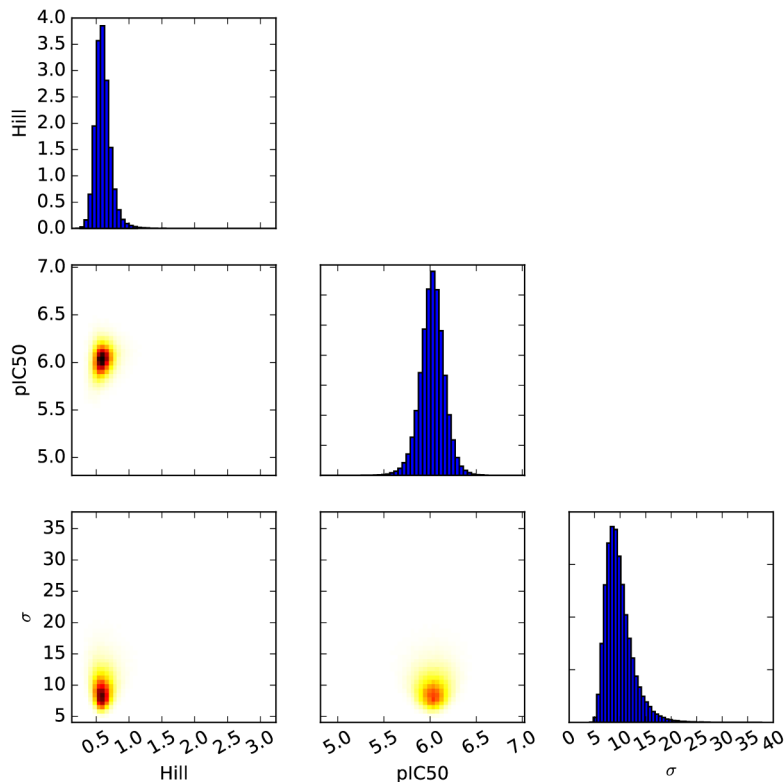

Figure 4. Matrix plot of normalised marginal and pairwise marginal histograms of the MCMC algorithm output samples for each parameter, in the amiodarone and hERG example. The well-defined narrow distributions, with lack of cross-correlation, suggest each parameter is being successfully inferred from the data.
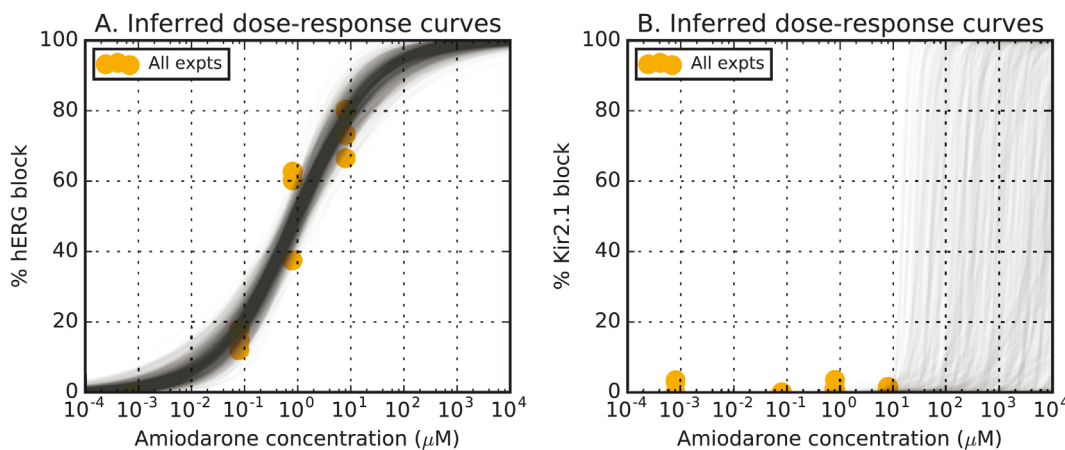
Figure 5. Inferred dose-response curves from MCMC (*pIC*50, *Hill*) samples. A: amiodarone with measurable effect on hERG. **B**: amiodarone with no measurable effect on Kir2.1. Outside of the range of measured concentrations, the MCMC algorithm was unable to find narrow ranges for possible parameter values, because the experimental data does not contain enough information.

the highest value at each concentration was observed during the same experiment, as shown for the amiodarone and hERG case in Figure 6.

4.1 Methods

The intra-experiment correlation seen in Figure 6 suggests that each experiment has its own distinct properties. We model this as inter-experiment variability in the pIC50 value and Hill coefficient. That is, we treat each experiment as having its own *pIC50* and *Hill*, which are drawn from distributions which are shared across experiments. We are assuming that two experiments have different pIC50 values, say, but that these values are mutually informative. We let N_e be the number of experiments performed. The vector of data points obtained from experiment i is $\mathbf{y}_i$, where $i = 1, \ldots, N_e$.

We take a hierarchical approach and assume that there is some 'higher-level' distribution that governs how these parameters vary across experiments (see Congdon, 2010, for an introduction to this approach).

Hill coefficient and pIC50 distributions for ion channel screening datasets of up to $N_e > 12,000$ repeats were published in Elkins *et al.* (2013), there they were found to fit independent log-logistic and logistic distributions, respectively. We assume the same type of distributions would occur here (if the experiments were repeated enough): that is, each experiment's *Hill$_i$* is drawn from a log-logistic distribution with parameters α and β; and each experiment's *pIC50$_i$* is drawn from a logistic distribution with parameters μ and s. We have assumed that the observational errors are drawn from the same Normal distribution across all N_e repeats, so we infer just a single noise standard deviation parameter σ. A schematic of this hierarchical statistical model is given in Figure 7, with the 'mid-level'

parameters and 'bottom-level' data points being independently distributed according to

$$Hill_i \sim \text{log-logistic}(\alpha, \beta), \tag{10}$$

$$pIC50_i \sim \text{logistic}(\mu, s), \tag{11}$$

$$y_i^{(j)} \sim \mathcal{N}\left(f\left(x_i^{(j)}; pIC50_i, Hill_i\right), \sigma^2\right), \tag{12}$$

where $y_i^{(j)}$ is the jth concentration entry in experiment i's responses $\mathbf{y}_i$. We suppose that every experiment i has K_i data points (to generalise to cases where different experiments tests different numbers of concentrations), so $j = 1, \ldots, K_i$ and $i = 1, \ldots, N_e$.

We now need to specify prior distributions over the 'top-level' parameters (in Figure 7): α, β, σ, μ, and s. Prior distributions are chosen to contain any prior information or beliefs we have about the parameters before observing the data. We can therefore inform our choice of prior distributions by considering previously-published ion channel screening data. We use gamma distributions for all of these, since gamma distributions, in general, only put probability mass on positive values, and because these parameters are all positive, with the exception of μ. μ, however, can take any value and represents the centre of the logistic distribution. For μ, we used a gamma distribution which is shifted along the x-axis down to -4, so there is little probability mass below -2. We choose this because as the pIC50 value becomes lower and lower, the IC50 becomes larger, and eventually any possible compound effects occur well above the experimental concentrations. This was illustrated above in Figure 3, where dose-response curves (with *Hill* = 1) have been plotted for varying values of *pIC*50. The shaded "region of interest" covers the minimum and maximum concentrations in the data
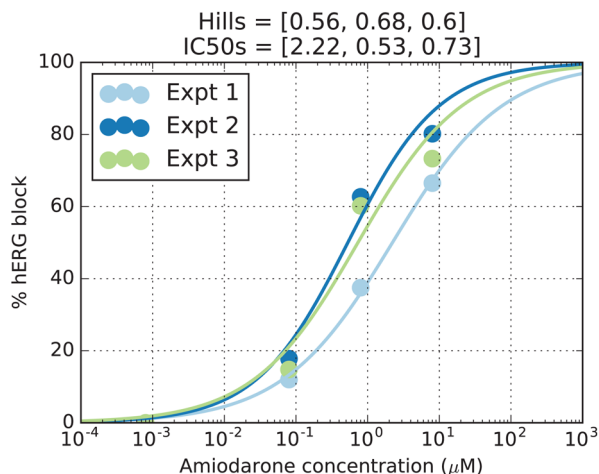
Figure 6. Data for hERG block by amiodarone suggests inter-experiment variability. Different whole-cell patch-clamp experiments are plotted with different colours. In this case, the responses are consistently in the same ordering relative to the other experiments at each compound concentration (*e.g.* Experiment 2 always shows the largest response, and Experiment 1 always the smallest). This suggests inter-experiment variability that is distinct from observational error σ.
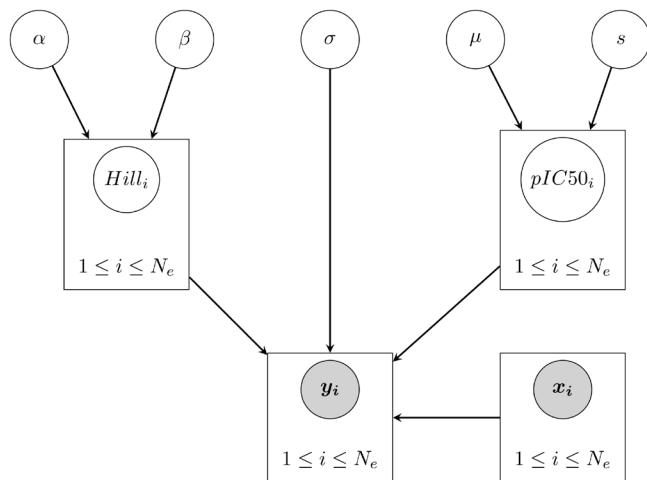

Figure 7. Hierarchical statistical model for dose-response data. *i* indexes the individual experiments. All non-shaded variables are parameters for which we wish to infer probability distributions.

published by Crumb *et al.* Ion channel screening is generally performed at concentrations that range to well-above therapeutic concentrations, and so we do not want to infer how a compound will behave at even higher concentrations. These gamma prior distributions were tuned to cover values provided by Elkins *et al.* (2013), but also allow more room for variation. Plots of the prior distributions for α, β, μ, s, and σ are given in Figure 8.

In addition to covering the values published by Elkins *et al.*, we restricted β to be greater than 2, so that all log-logistic

distributions generated would have no probability mass at 0, and also the gradient of the probability density function would be zero. This prevents Hill coefficients equal to 0 from being sampled by the MCMC algorithm. We also enforce that σ be greater than 10^{-3}, since we believe there is always the possibility of observation error, and hence there must be a positive standard deviation, we also run into division-by-zero numerical problems with the evaluation of the target distribution if we sample $\sigma = 0$ (see Equation (8)).

The choice of prior distribution will have an effect on the posterior distributions (via Equation (6)). However, the more information that is contained in our data, the less effect we expect our prior distribution to have on our posterior distribution. For example, Figure 9 shows how the marginal posterior distribution for the 'top-level' parameters correspond to their respective prior distributions in the case of synthetic data, where we fit to different numbers of experimental datasets. This synthetic data were generated by sampling pIC50 values from a logistic distribution with $\mu = 6$ and $s = 0.1$, and Hill coefficients were sampled from a log-logistic distribution with $\alpha = 1$ and $\beta = 5$. Normally-distributed observation noise with standard deviation $\sigma = 1$ was added to the dose-response model at every concentration.

We want to infer the posterior probability distribution for α, β, μ, s, σ, $Hill_i$ and $pIC50_i$, for $i = 1, \ldots, N_e$, giving a total of $5 + 2N_e$ parameters. Using Bayes' Theorem, the posterior distribution in Equation (6) is now given by

$$p\left(\alpha, \beta, \mu, s, \sigma, \{Hill_i\}_{i=1}^{N_e}, \{pIC50_i\}_{i=1}^{N_e} \mid \{\mathbf{y}_i\}_{i=1}^{N_e}\right)$$
$$\propto \left(\prod_{i=1}^{N_e} p(\mathbf{y}_i \mid \sigma, Hill_i, pIC50_i) p(Hill_i \mid \alpha, \beta) p(pIC50_i \mid \mu, s)\right) p(\alpha) p(\beta) p(\mu) p(s) p(\sigma). \ (13)$$

We use the same adaptive Metropolis-Hastings MCMC algorithm as in Section 3.1 to infer a posterior distribution from the experimental data. Since we have many more parameters than in the non-hierarchical case, we expect to have to run our MCMC algorithm for more iterations to adequately approximate the posterior distribution. For most cases in this dataset, we have $N_e = 3$ or $N_e = 4$, which is not too demanding since our mathematical model is a simple analytic expression, and does not require solving differential equations as our previous work did (Johnstone *et al.*, 2016a). However, if N_e became very large, we may have to use alternative MCMC techniques.

4.2 Implementation
Our tool *PyHillFit* (Johnstone *et al.*, 2016b) takes a CSV file of dose-response points as its input. The file should be comma separated values (.CSV) in the following format for each line:

compound name, channel name, experiment number, dose (μM), response (% inhibition)

The Monte Carlo algorithms were written in Python using **NumPy** 1.11.0 for numerical linear algebra (van der Walt *et al.*, 2011), and functions from the **SciPy** 0.15.1 library (Jones *et al.*, 2001). **Pandas** 0.17.1 was used to read the input data csv files
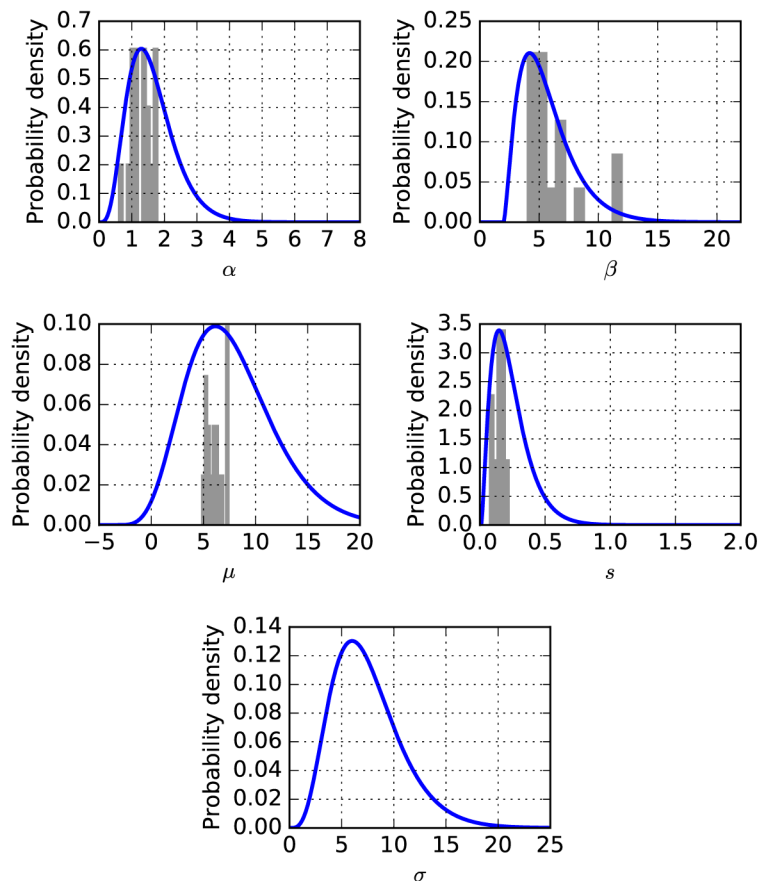
Figure 8. Prior distributions. Blue: gamma distributions used as the prior distributions over α, β, μ, s and σ. Grey: histograms of parameter estimates for different strongly-blocking control compounds, with large numbers of repeats, as previously published in Elkins *et al.* (2013). However, we want to be able to fit to all compounds, including ineffective ones that elicit no response, hence the increased width of the prior distributions on certain parameters (μ particularly).

(McKinney, 2010). **cma** 1.1.6 was used for initial optimisation to find best-fit parameter values (Hansen *et al.*, 2003), which act as starting positions for the MCMC algorithms. All figures were plotted in **matplotlib** 1.5.1 (Hunter, 2007) and **seaborn** 0.7.1 (http://seaborn.pydata.org/).

PyHillFit output takes the form of files listing samples of the posterior distributions for the dose-response curve parameters, together with some visualizations of these (as shown throughout this article).

4.3 Results
As before, we plot normalised marginal histograms to approximate the marginal posterior distributions for each parameter. Such histograms are plotted for α, β, μ, s, and σ in Figure 10 for the amiodarone and hERG case. We can compare these to the prior distributions shown in Figure 8, and we see that in most cases we

have much narrower marginal posterior distributions than prior distributions. This tells us that the data contains enough information about those parameters to constrain them to narrower intervals.

We can also superimpose the normalised histograms for each $Hill_i$ and $pIC50_i$ to give us an idea of how much inter-experiment variability is present in these parameters. These superimposed histograms are plotted in Figure 11 for the amiodarone and hERG case.

To make predictions about how a particular compound and channel will interact if we perform another experiment, we consider the *posterior predictive* distributions for $Hill_i$ and $pIC50_i$. That is, what are $p(Hill_{N_{e+1}}|\text{data})$ and $p(pIC50_{N_{e+1}}|\text{data})$? Since the $Hill_i$ and $pIC50_i$ are modelled as being drawn from log-logistic and logistic distributions, respectively, we sum the log-logistic and logistic
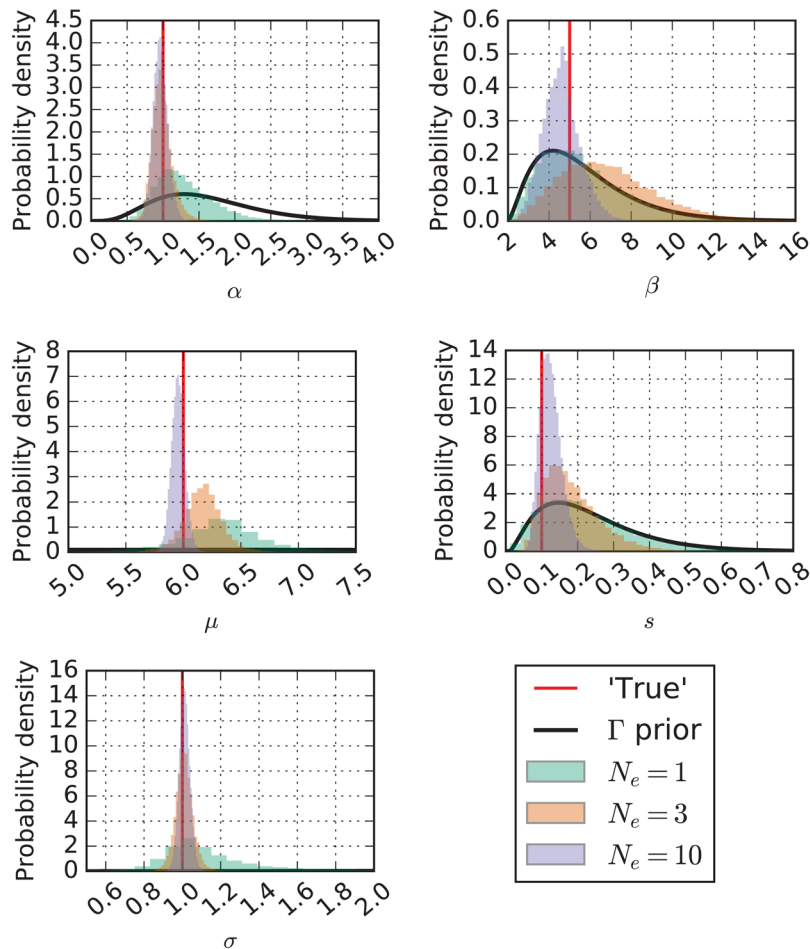
Figure 9. A comparison between marginal posterior distributions for 'top-level' parameters in the hierarchical model, with their respective gamma (Γ) prior distributions. The number of (synthetic) experimental datasets, N_e, being fitted was increased. As we fit to more experiments, the prior distributions have a smaller effect on the posterior distributions. Where the black line for the prior distributions looks as if it lies along the x-axis (for μ and σ, the prior distribution was much wider than the marginal posterior distribution; there is a lot of information on these parameters with just one experiment).

distributions generated from the 'top-level' parameters at every iteration of our MCMC algorithm output, then normalise them to obtain two new probability distributions.

$$p\left(Hill_{N_e+1} \mid \text{data}\right) \approx \frac{1}{T}\sum_{t=1}^{T} \text{log-logistic}\left(Hill_t;\alpha_t,\beta_t\right), \quad (14)$$

$$p\left(pIC50_{N_e+1} \mid \text{data}\right) \approx \frac{1}{T}\sum_{t=1}^{T} \text{logistic}\left(pIC50_t;\mu_t,s_t\right), \quad (15)$$

where t indexes the samples in our Markov chain, after having discarded a number of initial samples as a burn-in.

These are not necessarily distributions which can be sampled from directly, but we can approximately sample from them using the inverse-cumulative distribution function (CDF) method. We sum and then normalise the individual log-logistic and logistic CDFs. After sampling from these new distributions, we plot similar dose-response curves as in Section 3.2. A plot of predicted dose-response curves for a future experiment, following the hierarchical model

MCMC, is given in plot A of Figure 12. To make a prediction of what %-block will be induced by that compound at a particular concentration, we take a vertical cross-section through these dose-response curves and plot a normalised histogram of these levels of block to approximate a probability distribution, as shown in plot B of Figure 12.

Note that the hierarchical model allows us to make two sets of predictions. Firstly, using the posterior predictive distribution given by Equation (14) & Equation (15) as shown in Figure 12 (panel A). This distribution includes inter-experiment variability, and can therefore be considered a distribution that predicts where data points from future experiments may lie. Secondly, we can examine the variability in the underlying properties of the compound; the 'average' effect, before it is altered by inter-experiment variability (panel B). We generated this plot by taking samples of α and μ to use as Hill and pIC50 values. We would expect panel B to be more directly comparable with the single-level approach
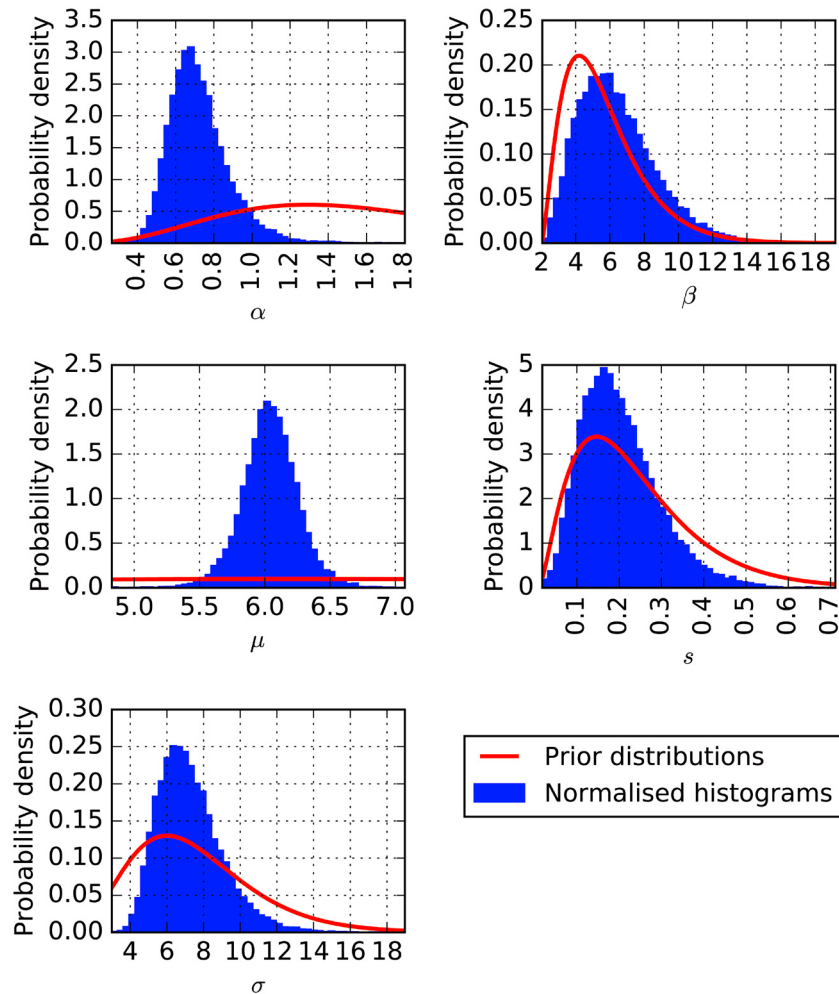
Figure 10. Normalised marginal histograms for the 'top-level' parameters, α, β, *μ, s,* and σ after running the fitting the hierarchical model to the amiodarone and hERG dataset using the MCMC algorithm. The red lines indicate the respective prior distributions. Most of these distributions are narrower than their respective prior distributions in Figure 8, with the exception of β. We therefore conclude that the experimental data does not contain much information about β, in line with the synthetic data study shown in Figure 9.

(which fits 'average' data points), which is shown in panel C for comparison.

Which of the two distributions (illustrated in panels A or B in Figure 12) one may wish to use for predictions is subtle. If we consider that the source of variability between experiments is also present in the system that we are making predictions for, then the first case (panel A) would be the best to use. If however we consider that there is a single underlying effect, and the act of measuring it introduced inter-experiment variability that is not present in the real system, then the second distribution (panel B) would be more appropriate. Most biological experiments implicitly assume the second case is true — that by taking repeated measurements

and then taking the average, a more accurate assessment of the underlying system is made.

5 A comparison of single-level and hierarchical models
There are advantages and disadvantages to choosing either the single-level statistical model, or the hierarchical statistical model. The main benefit of the single-level model is that we are only fitting three parameters, meaning that the parameter space of interest is relatively easy to explore. This means that we need to run our MCMC algorithm for fewer iterations to obtain an acceptable approximation of the posterior distribution than if we had a larger number of parameters, reducing overall computation time. However, a model with an analytic solution such as the one
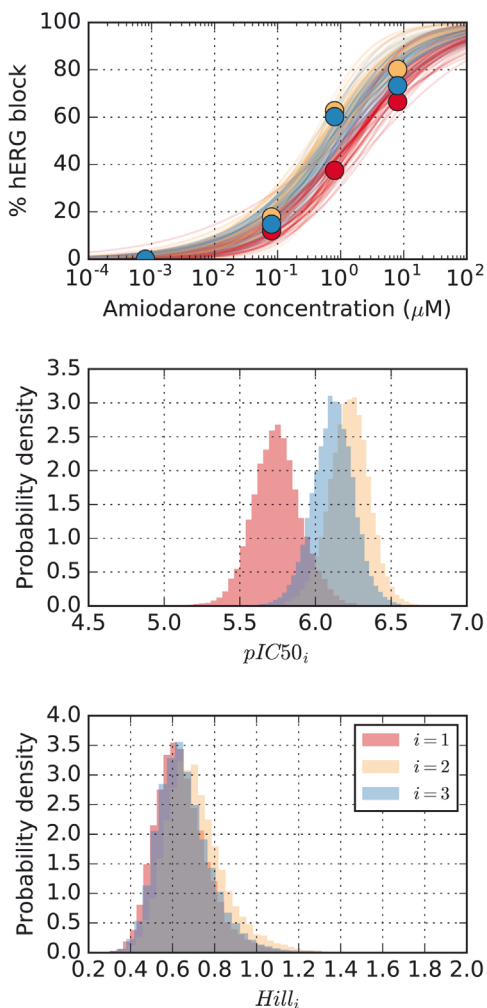
Figure 11. Inferred parameters for individual experiments. Top: dose-response curves plotted using the 'mid-level' $pIC50_i$ and $Hill_i$ samples from our MCMC algorithm output from the amiodarone and hERG dataset. Middle & bottom: superimposed normalised histograms for $pIC50_i$ and $Hill_i$, after fitting our hierarchical model to the amiodarone and hERG dataset using the MCMC algorithm. We find that the Hill coefficient does not vary much between experiments, however there is variability within the pIC50 value.

we have here (Equation (3)) can be solved very quickly, and so computation time is generally not a problem for even the hierarchical version of the model, for this application. There is also little sensitivity to the prior distributions, as there is a lot of information about all three parameters in even one experiment.

In Figure 12 we can compare the results of the two types of inference for real data on amiodarone block of the hERG current. There is a small difference in the predicted curves.

In the single-level case, by fitting to all data points at once, the inference can misinterpret inter-experiment variability and assign it to the 'wrong' parameter(s). To demonstrate this, we generated two

sets of synthetic data corresponding to these fictitious compounds, with fictitious inter-experiment variability properties:

1. For shamiodarone, *Hill* was fixed as 1 across all experiments, and the $pIC50_i$ were drawn from a logistic distribution, with $\mu = 6$ and $s = 0.2$.

2. For shamitriptyline, $pIC50$ was fixed as 6 across all experiments, and the $Hill_i$ were drawn from a log-logistic distribution with $\alpha = 1$ and $\beta = 2.5$.

In both cases, we simulated 5 experiments where each experiment consists of measuring % channel block at 4 different compound concentrations. We added Normal observation noise with standard deviation $\sigma = 0.5$ to each point.

Case 1: fixed Hill and varying pIC50
A sample of the inferred curves for this case are given in Figure 13. The plots in panels A & D represent $pIC50$ and *Hill* parameter values being drawn from their respective posterior predictive distributions; this gives predictions of how we believe the observations from a future experiment would behave. The plots in B & E are based on α and μ samples — what we believe to be the underlying 'average' behaviour of the compound interacting with an ion channel, when experimentally-introduced variability is discounted. The hierarchical model was able to identify consistency within the $Hill_i$, and the MCMC algorithm generally only infers that $pIC50_i$ varied between experiments.

The histograms in Figure 13 (panels D–F) are a cross-section of the dose-response curves at different concentrations, and represent the probability density of % block at that compound concentration. Note that each curve in panel B has approximately the same slope, corresponding to a consistent Hill coefficient, whereas in panel C we see that there is a greater range of slopes, corresponding to (slightly more) variability in the Hill coefficient. Comparing plot E with plot F, we see that at 0.05 μM concentration, the non-hierarchical model is less certain about the % channel block than the hierarchical model, because the former has incorrectly inferred there is more variation in the Hill coefficient. So the single-level model tends to compensate for the varying parameter values by fitting curves that fit through an 'average' of the points. The algorithm does this by varying both *Hill* and *pIC50* to obtain curves that could fit the data reasonably well, even when the synthetic data were generated by holding one parameter fixed and varying the other.

Case 2: fixed pIC50 and varying Hill
A sample of the inferred curves for this case are given in Figure 14. The single-level model in panel C does show small variability in the Hill coefficient, as well as small variability in the IC50 (and hence pIC50). This leads to a reasonably spread prediction of ion channel block at both a concentration near pIC50 and at a higher concentration (panel F). But we know that the underlying data had the same pIC50, and so variability near the IC50 should be minimal, and indeed the spread of predictions at a higher concentration should be larger. The hierarchical model captures the desired underlying variability better (compare panels E and F). The Hill coefficient varies (panel B) while also
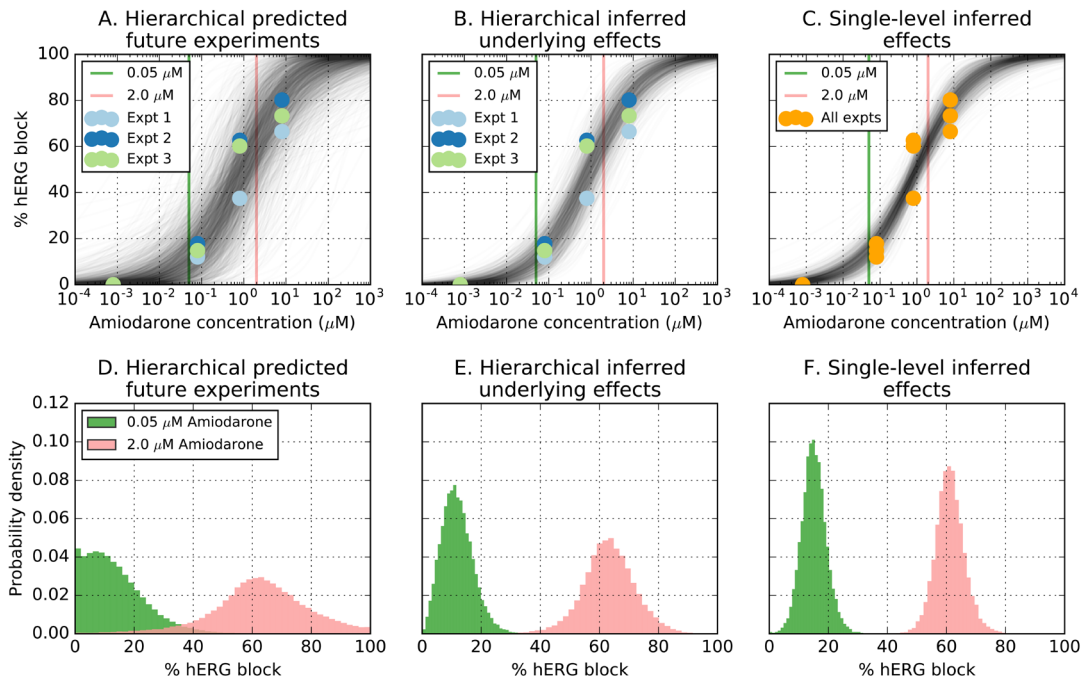
Figure 12. Predicted dose-response curves and associated probability distributions for levels of block at example concentrations when fitting to the amiodarone and hERG dataset. A: Hierarchical model — predictions for how a future experiment will behave, with samples taken from the posterior predictive distributions. **B**: Hierarchical model — inferred distribution for the underlying behaviour of the system, plotted by using μ and α samples from the MCMC algorithm output directly as values for $pIC50$ and $Hill$, respectively. **C**: Single-level — inferred distribution, plotted by taking samples from the MCMC algorithm output. **D,E,F**: Histograms of the intersections between the vertical lines and dose-response curves in **A**,**B**,**C**, respectively, at two different concentrations of amiodarone.
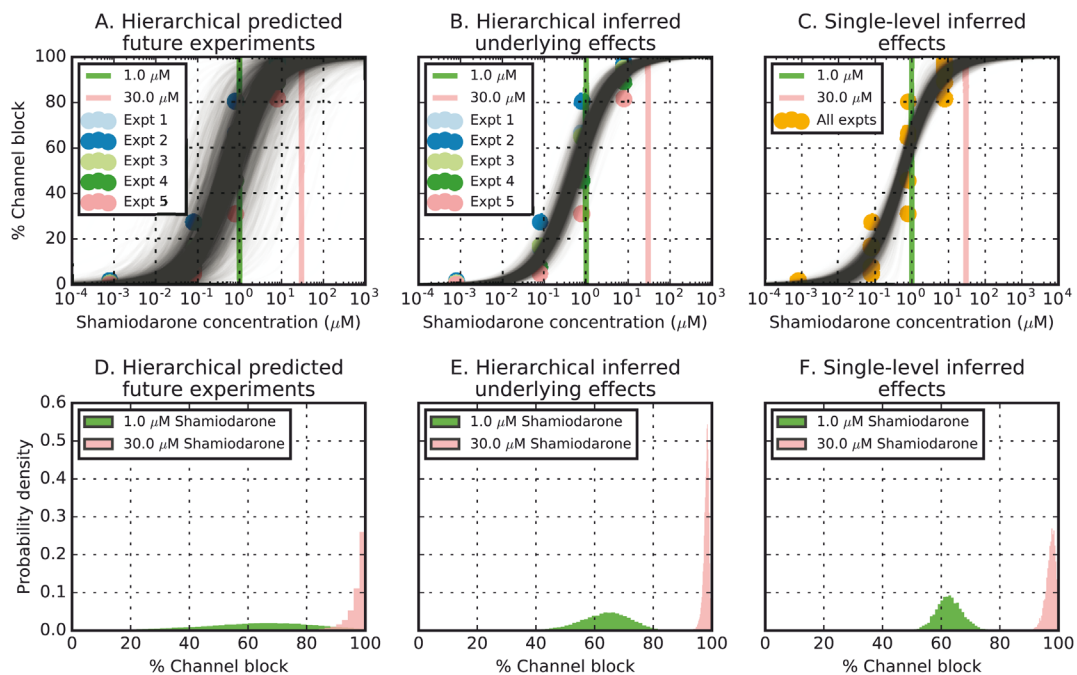

Figure 13. Inference on synthetic data generated by fixing *Hill* = 1 and varying *pIC*50. A: Predicted dose-response curves, with $pIC50$ and $Hill$ sampled from their respective posterior predictive distributions, taking inter-experiment variability into account. **B**: Inferred underlying behaviour of the compound-ion channel interaction, with inter-experiment variability discounted. **C**: Inferred dose-response curves from single-level inference. **D–F**: Normalised histograms of the cross sections plotted with vertical lines in plots **A**–**C**. These histograms represent probability density functions of % block at a particular concentration, given the (synthetic) experimental data.
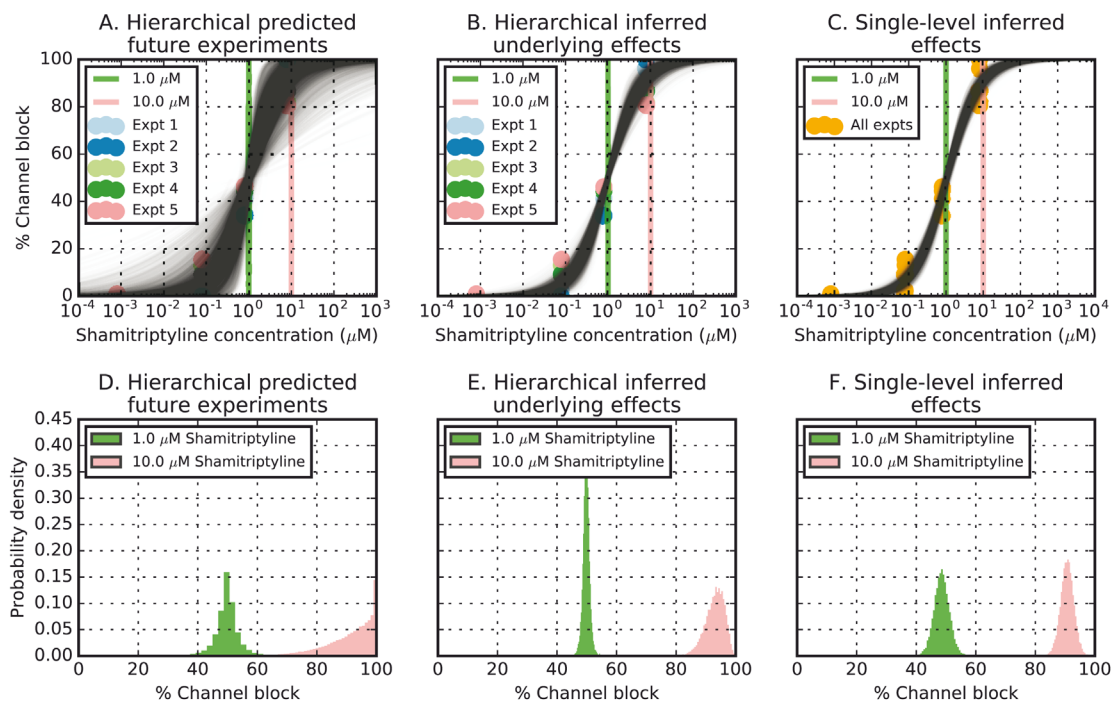
Figure 14. Inference on synthetic data generated by fixing *pIC*50 and varying *Hill*. A: Predicted dose-response curves, with *pIC*50 and *Hill* sampled from their respective posterior predictive distributions, taking inter-experiment variability into account. **B**: Inferred underlying behaviour of the compound-ion channel interaction, with inter-experiment variability discounted. **C**: Inferred dose-response curves from single-level inference. **D–F**: Normalised histograms of the cross sections plotted with vertical lines in plots **A–C**. These histograms represent probability density functions of % block at a particular concentration, given the (synthetic) experimental data.

capturing the low variability in the pIC50 value (there is still some variability in the inferred pIC50 distribution due to observational noise and a low number of repeat experiments). As a result, the predicted percentage blocks in panels E and F are different. As for Case 1, either too-much or too-little variability is predicted by the single-level inference, depending on the concentration.

6 Propagating dose-response uncertainty

The proposed Comprehensive *in-vitro* Pro-arrhythmia Assay (CiPA) recommends the use of computational action potential models in the drug safety process. Ion channel screening will be performed, and the IC50 values and Hill coefficients obtained from these experiments are to be used in action potential models to predict whether or not a compound is likely to be pro-arrhythmic. One simple proposed measure of pro-arrhythmia is action potential duration prolongation (Mirams *et al.*, 2011), directly related to prolongation of the QT-interval, which can be a precursor to potentially fatal arrhythmias such as Torsade de Pointes.

Using best-fit IC50 values and Hill coefficients obtained from ion channel screening data, we can compute a predicted level of block of each of the ion currents in an action potential model, at a particular compound concentration. We then simulate an action potential and measure the action potential duration prolongation relative to the control case (see Beattie *et al.* (2013) for an example of this

approach). However, when using best-fit IC50 values and Hill coefficients, we obtain a single predicted action potential after simulating a particular compound concentration.

Instead, we use our tool to infer probability distributions for each experiment's IC50 value and Hill coefficient. Each sample is 'equally likely', but there will be more samples around the regions of greater probability density. We can then randomly take samples from these distributions and compute an action potential duration for each sample. Again, each output is 'equally likely', but there will be more outputs close to each other where there is the greatest probability density. This allows us to construct a probability distribution for predicted action potential durations based on the original ion channel screening data (uncertainty propagation).

To illustrate the proposed uncertainty propagation we use our tool to fit hierarchical dose-response parameters to thirty drug compounds for seven ion currents each using the Crumb *et al.* (2016) dataset supplied with the code associated with this article. We then take 500 samples from the MCMC output for each drug and channel combination, based on the 'underlying effects' curves from hierarchical fits (see Figure 12, Figure 13 & Figure 14), and simulate action potential durations after applying seven ion channel block using the approach outlined in Mirams *et al.* (2011). We plot the resulting predicted action potential duration distributions as violin plots in Figure 15. A violin plot for each of the thirty compounds
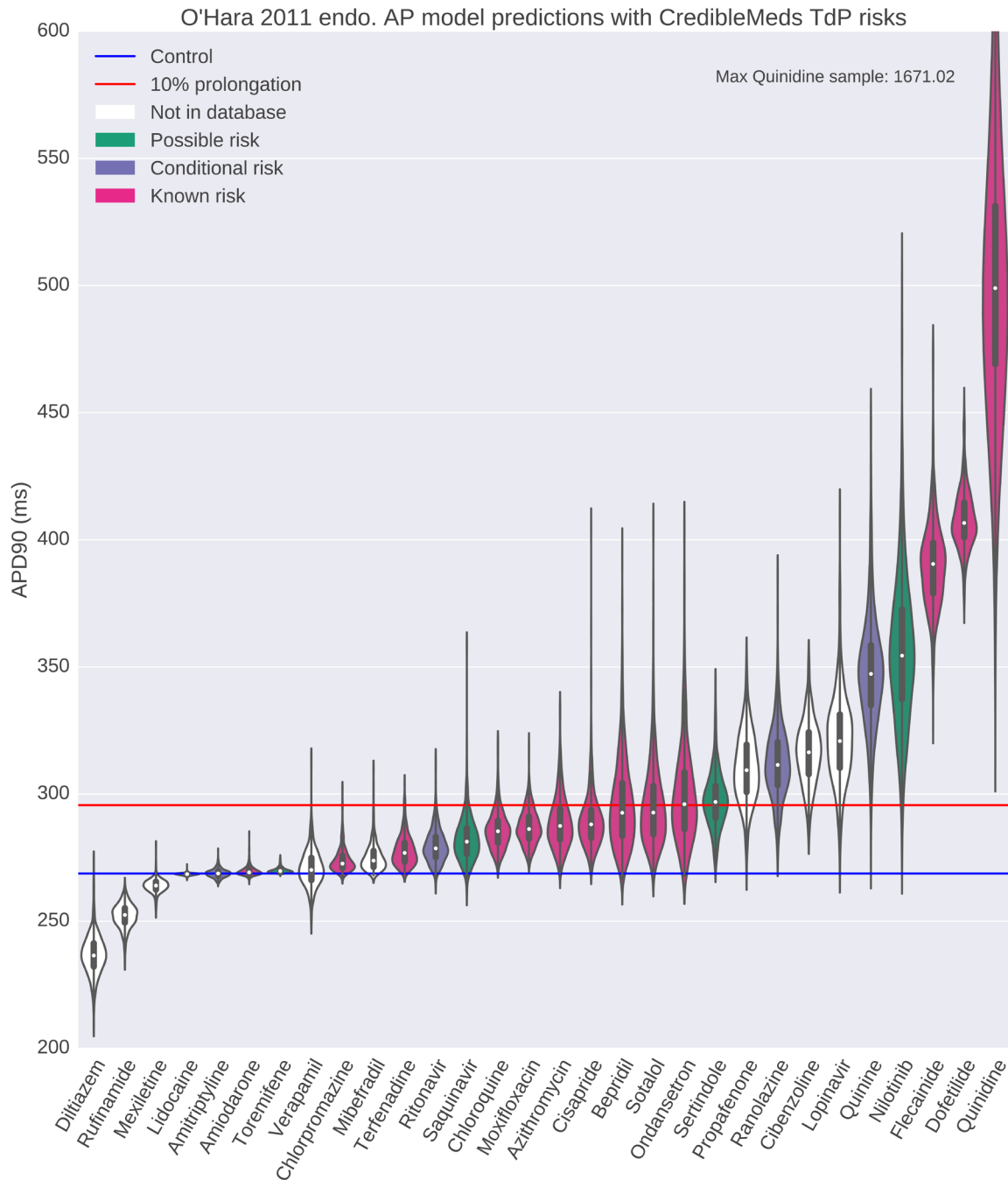
Figure 15. Violin plots of prediction action potential duration (APD90, time taken for the cell to return to 90% repolarisation after depolarisation) using 500 samples from the iterations of our MCMC algorithm under the hierarchical statistical model. Simulations were run using the O'Hara *et al.* (2011) human ventricular cardiomyocyte action potential model, and APD90s were computed. We used 'AP-predict' (Williams & Mirams, 2015), a bolt-on project for the Chaste open-source computational biology C++ library (Mirams *et al.*, 2013). Violin distribution plots are shown for each of the 30 compounds discussed by Crumb *et al.* In one case, several samples led to very long action potentials and so the y-axis is cut off early for clarity.

is coloured according to the Torsade/QT risk categorisation in the CredibleMeds database (https://crediblemeds.org/). We also plot the control action potential duration in the O'Hara model along with a 10% action potential duration prolongation.

We see that the vast majority of the compounds have overlapping probability distributions for predicted APD at maximum free therapeutic plasma concentration, suggesting that at least this previously-proposed measure will be insufficient to distinguish compounds in terms of risk based on data such as these.

This suggests that either: action potential prolongation is not a good enough marker of pro-arrhythmia; or, there was too much uncertainty associated with the experimental data to constrain these distributions to narrow distinct ranges. To counter the latter point, we suggest that more experimental repeats be performed. In either case, it is imperative to realise that the data being used have a level of uncertainty which means it is not possible to rank the majority of these drugs in terms of their predicted APD.

7 Discussion

A Bayesian framework is a useful tool to address uncertainty characterisation in ion channel screening data. When no uncertainty characterisation is performed, one can obtain best-fit parameter values for the data presented, but there is no associated probability in terms of the behaviour that generated the data, or in predictions informed by the data. The single-level Bayesian inference model can provide ranges of possible dose-response curves (and underlying parameters) that fit ion channel screening data. But parameter-specific inter-experiment variability can be missed when using a 'single-level' statistical model, as the algorithm treats all points equally and so varies the parameters without considering the inter-experiment correlations. This leads to an 'averaging' effect, where the dose-response model is fitting to an average of the experimental data points, but may not reflect the behaviour of any individual experiment. However, this single-level inference is quick to run as it only requires fitting 3 parameters, and provides a better approximation of probability distributions than a single best-fit.

A hierarchical statistical model can capture inter-experiment variability within certain dose-response parameters, as demonstrated in the synthetic cases discussed in Section 5. The hierarchical model can therefore be used to infer inter-experiment behaviour, and hence predict how a future experiment might behave. By taking samples for the 'top-level' parameters from our MCMC output, we can build distributions of how we believe the compound is interacting with the ion channel. At a given compound concentration, we then have a probability distribution for possible levels of ion channel block. The hierarchical model is able to determine what variability is being introduced at the experimental level, and allows us to make probabilistic statements about the underlying behaviour.

Our hierarchical model is similar to *nonlinear mixed effects* (NLME) modelling, but we operate in a Bayesian framework.

NLME assumes a similar structure to that shown in Figure 7, but infers best-fit values for the 'top-level' parameters, and a distribution from which the 'mid-level' parameters are sampled. While it does capture inter-experiment variability and would allow us to make predictions about how a future experiment might behave, it only provides a point-estimate for underlying behaviour, rather than different possibilities with relative probabilities.

A possible limitation of the hierarchical model is that computation time increases with the number of experimental datasets being fit to at once. This is not a problem for up to 4 or 5 experiments, but the number of parameters quickly becomes intractable for the adaptive-Metropolis MCMC algorithm that we have been using. In general, with MCMC methods, we want to run our algorithm for as long as possible, to best approximate samples from the posterior distribution. There is therefore no upper limit for how long this method takes, although for these examples we have run our algorithm for 500,000 iterations, which takes approximately 12 minutes for the amiodarone-hERG case which has 3 experimental datasets of 4 concentrations each.

Another possible limitation of the hierarchical model is the dependence on the prior distributions for the 'top-level' parameters. As shown in Figure 9, when there is not much data, the posterior is heavily influenced by the prior. However, we chose our priors based on data published by Elkins *et al.* which was based on 12,000 ion channel screening experiments, and we therefore have some confidence in their shapes (Figure 8). In a Bayesian framework, should new data become available, we can compute new posterior distributions for the parameters according to Equation (5), by using a previous posterior distribution as the new prior distribution.

A benefit of both inference techniques is that we introduce the observation noise as a parameter to be fitted, along with the pIC50 values and Hill coefficients. Instead of estimating the observation noise and then fitting dose-response curves based on our estimate, we allow the MCMC algorithm to find likely levels of noise, while also quantifying the uncertainty in those estimates. Since all of these parameters are being fit at the same time by the MCMC algorithm, we can extract how much noise on dose-response parameters is introduced by inter-experiment variability, and how much noise is due to observation error.

7.1 Conclusions

Single best-fit parameter estimates from ion channel screening data can give a *most likely* set of dose-response curve parameters. However, this approach does not provide us with a measure of uncertainty around these parameters. The software tool we present can quantify some of the uncertainty associated with dose-response curves, with its default priors set for ion channel screening data, for the purposes of propagating this uncertainty into further quantitative studies.

Data and software availability

Latest source code and datasets used in the publication: https://github.com/mirams/PyHillFit

Archived source code and datasets as at the time of publication: https://doi.org/10.5281/zenodo.237643 (Johnstone *et al.*, 2016b)

License: BSD 3-Clause

The code contains the experimental input data required to reproduce the examples shown here in comma separated value (CSV) format in the file `data/crumb_data.csv`. Installation instructions for the tool and its dependencies can be found in the README file, in the main folder at the above links.

Author contributions

RHJ, GRM, RB and DJG conceived the study. GRM did some boring data extraction. RB designed the algorithms. RHJ wrote the software tool. RHJ and GRM ran the simulations and prepared the manuscript. All authors were involved in the revision of the draft manuscript and have agreed to the final content.

Competing interests

No competing interests were disclosed.

References

Beattie KA, Luscombe C, Williams G, *et al.*: **Evaluation of an *in silico* cardiac safety assay: using ion channel screening data to predict QT interval changes in the rabbit ventricular wedge.** *J Pharmacol Toxicol Methods.* 2013; **68**(1): 88–96.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Congdon PD: **Applied Bayesian hierarchical methods.** 2010.
Publisher Full Text

Crumb WJ Jr, Vicente J, Johannesen L, *et al.*: **An evaluation of 30 clinical drugs against the comprehensive *in vitro* proarrhythmia assay (CiPA) proposed ion channel panel.** *J Pharmacol Toxicol Methods.* 2016; **81**: 251–262.
PubMed Abstract | **Publisher Full Text**

Daly AC, Gavaghan DJ, Holmes C, *et al.*: **Hodgkin–Huxley revisited: reparametrization and identifiability analysis of the classic action potential model with approximate Bayesian methods.** *R Soc Open Sci.* 2015; **2**(12): 150499.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Elkins RC, Davies MR, Brough SJ, *et al.*: **Variability in high-throughput ion-channel screening data and consequences for cardiac safety assessment.** *J Pharmacol Toxicol Methods.* 2013; **68**(1): 112–122.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Fermini B, Hancox JC, Abi-Gerges N, *et al.*: **A New Perspective in the Field of Cardiac Safety Testing through the Comprehensive *In Vitro* Proarrhythmia Assay Paradigm.** *J Biomol Screen.* 2016; **21**(1): 1–11.
PubMed Abstract | **Publisher Full Text**

Haario H, Saksman E, Tamminen J: **An adaptive Metropolis algorithm.** *Bernoulli.* 2001; **7**(2): 223–242.
Publisher Full Text

Hansen N, Müller S, Koumoutsakos P: **Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES).** *Evol Comput.* 2003; **11**(1): 1–18.
PubMed Abstract | **Publisher Full Text**

Hill AV: **Proceedings supplement: The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves.** *Journal Physiol.* 1910; **40**: iv–vii.
Reference Source

Hunter JD: **Matplotlib: A 2D Graphics Environment.** *Comput Sci Eng.* 2007; **9**(3): 90–95.
Publisher Full Text

Johnstone RH, Chang ET, Bardenet R, *et al.*: **Uncertainty and variability in models of the cardiac action potential: Can we build trustworthy models?** *J Mol Cell Cardiol.* 2016a; **96**: 49–62.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Johnstone RH, Bardenet R, Gavaghan DJ, *et al.*: **PyHillFit: code to fit dose-response Hill curve parameters in a Bayesian inference framework [Data set].** *Zenodo.*

2016b.
Data Source

Jones E, Oliphant E, Peterson P, *et al.*: **SciPy: Open Source Scientific Tools for Python.** 2001.
Reference Source

McKinney W: **Data Structures for Statistical Computing in Python.** *Proc Of The 9th Python In Science Conf.* 2010; 51–56.
Reference Source

Mirams GR, Arthurs CJ, Bernabeu MO, *et al.*: **Chaste: an open source C++ library for computational physiology and biology.** *PLoS Comput Biol.* 2013; **9**(3): e1002970.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Mirams GR, Cui Y, Sher A, *et al.*: **Simulation of multiple ion channel block provides improved early prediction of compounds' clinical torsadogenic risk.** *Cardiovasc Res.* 2011; **91**(1): 53–61.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Mirams GR, Pathmanathan P, Gray RA, *et al.*: **White paper: Uncertainty and variability in computational and mathematical models of cardiac physiology.** *J Physiol.* 2016.
PubMed Abstract | **Publisher Full Text**

O'Hara T, Virág L, Varró A, *et al.*: **Simulation of the undiseased human cardiac ventricular action potential: model formulation and experimental validation.** *PLoS Comput Biol.* 2011; **7**(5): e1002061.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Pathmanathan P, Gray RA: **Ensuring reliability of safety-critical clinical applications of computational cardiac models.** *Front Physiol.* 2013; **4**: 358.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Pathmanathan P, Shotwell MS, Gavaghan DJ, *et al.*: **Uncertainty quantification of fast sodium current steady-state inactivation for multi-scale models of cardiac electrophysiology.** *Prog Biophys Mol Biol.* 2015; **117**(1): 4–18.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Raue A, Kreutz C, Maiwald T, *et al.*: **Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood.** *Bioinformatics.* 2009; **25**(15): 1923–1929.
PubMed Abstract | **Publisher Full Text**

Sager PT, Gintant G, Turner JR: **Rechanneling the cardiac proarrhythmia safety paradigm: a meeting report from the Cardiac Safety Research Consortium.** *Am Heart J.* 2014; **167**(3): 292–300.
PubMed Abstract | **Publisher Full Text**

Siekmann I, Sneyd J, Crampin EJ: **MCMC can detect nonidentifiable models.** *Biophys J.* 2012; **103**(11): 2275–2286.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

US National Research Council: **Assessing the reliability of complex models: mathematical and statistical foundations of verification, validation, and uncertainty quantification**. National Academies Press; 2012.
Publisher Full Text

van der Walt S, Colbert SC, Varoquaux G: **The NumPy Array: A Structure for Efficient Numerical Computation.** *Comput Sci Eng.* 2011; **13**(2): 22–30.
Publisher Full Text

Vernon I, Goldstein M, Bower RG: **Galaxy formation: a Bayesian uncertainty analysis.** *Bayesian Anal.* 2010; **5**(4): 619–669.
Publisher Full Text

Weiss JN: **The Hill equation revisited: uses and misuses.** *FASEB J.* 1997; **11**(11): 835–841.
PubMed Abstract

Williams G, Mirams GR: **A web portal for *in-silico* action potential predictions.** *J Pharmacol Toxicol Methods.* 2015; **75**: 10–16.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Open Peer Review

Current Referee Status: ✔ ✔

Version 1

Referee Report 30 November 2016

✔ **Derek Leishman**
Eli Lilly and Company, Indianapolis, IN, USA

This is an interesting and stimulating manuscript. The structure of the manuscript in title, abstract, problem statement (captured in the introduction and motivation), methodology, results and conclusions is appropriately clear and all sections are well written. The conclusions are well founded. The motivation in particular for the software tool is clear as is the potential impact of the tool.

The intended reader who might benefit most from the manuscript and tool - the experimental electrophysiologist generating ion channel screening data to predict potential safety issues - may find the methodology and underlying math daunting. Nonetheless there are principles described which need to be appreciated by those generating those data, especially when that data is further propagated in *in silico* models of cardiac tissues. The authors have written and formatted the article very effectively to make the manuscript understandable and approachable. Given the availability of computational tools such as Python it is now also much more practical to perform these analyses and appropriately quantify the uncertainty. The practice of reporting 'best-fit' parameter values may have been the best which could practically be achieved previously. It may also have been a model sufficient for many questions but the developing need to use these data in computational models of tissues for safety purposes requires a more robust treatment of uncertainty. Simply carrying forward best-fit values and generating point estimates of effects on tissues and point estimates of margins between therapeutic drug concentrations and concentrations impacting cardiac tissues may not adequately serve drug development and regulatory assessment.

In the practice of electrophysiology screening experiments some compromises are made in order to make the experiments both quick enough and inexpensive enough to meet the drug discovery process. This will often mean that a limited range of concentrations are tested and a minimum number of individual experimental repeats are made. When calculating 'best-fit' parameters some values are 'fixed' to make calculating the key parameter IC50 possible e.g. fixing minimum current inhibition to 0%, maximum inhibition to 100% and the Hill Slope to 1. The manuscript describes some of the potential issues which may result from such compromises. There are also hidden impacts of experimental design. As an example the experiment may call for following compound effect for as long as necessary to reach a steady state effect. As the kinetics of channel block are concentration dependent this means that steady state will occur more slowly at low concentrations. In addition to drug effect there is also often current 'run-down' or 'run-up' evident in these experiments. The 'run-down' effect is the more common. This means that at low concentrations of compound the observed effect may have a larger contribution of 'run-down' than at larger concentrations. This serves to somewhat flatten the Hill Slope and left shift the IC50 from 'truth'. Alternatively, fixed durations of observation may be applied at each concentration. As

the norm is to try and minimize the overall duration of the recording this may mean that compound effect is under estimated at low concentrations leading to the Hill slope becoming steeper and the IC50 being slightly right shifted. Some electrophysiologists may choose to have uniform long compound application periods or use run-down correction techniques to minimize these experimental impacts on parameter estimates. The significant value of the current manuscript is that it allows the data from the experiments to be used in their most raw form without the compromises described. It leverages priors based on previous experimental information. It gives the key information as a probability distribution rather than a single value. It also serves to illustrate the uncertainty and give the opportunity to generate more data for those compounds which show promise and where more certainty is desirable.

There is one key outstanding element which the current manuscript cannot address. In this experimental context it is the assessment of the kinetics of block. When relating the effect of a compound to how it may impact a tissue when the underlying kinetics of the ion channel experiment and that of the ionic current in a tissue under the normal heart rates differ cannot be captured in an IC50 value or Hill slope even taking into account the uncertainty. This will likely require different experimental voltage-clamp paradigms and the estimation of at least one further parameter. That said that parameter estimate then becomes accessible to the types of calculation described here.

Overall this is a very well thought out piece of work, well written and with a potentially useful piece of software described.

This review is based on my personal professional opinion.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

✔ **Jonathan Stott**, **Stanley Lazic**
   Quantitative Biology, IMED Discovery Sciences, AstraZeneca, Cambridge, UK

Summary of the Paper

In this manuscript Johnstone *et al* propose a new tool (PyHillFit) for fitting dose response curves from cardiac ion channels. This tool uses Bayesian inference methods to determine the distributions of the parameters. They investigate two alternative models, one that ignores inter-experimental variability and a second hierarchical model that characterises variability. Using a recently published dataset of inhibition of multiple ion channels they demonstrate the use of the tool and how experimental variability propagates through to the fitted dose response curves. They then extend these to simulations using a cardiac action potential model to show the distribution of model outputs.

The benefits of a hierarchical Bayesian approach are clearly demonstrated. The paper describes the tool and its underlying model and assumptions well, but could benefit from further development of the arguments around the cardiac action potential distribution.

Main Comments

The most interesting part of the paper was propagating the uncertainty though the action potential simulations. The authors conclude that this measure is insufficient to distinguish compounds in terms of risk (Fig 15), but if the distributions were plotted so that they are not on top of each other it may indeed be possible to distinguish the compounds. For example, "caterpillar plots", boxplots, or violin plots (e.g. Fig 2 in http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0096431) would better show the result.

The authors also do not attempt to classify the compounds on the basis of their distribution (e.g. >50% of the distribution over a 10% prolongation of APD(90) to indicate risk of TdP) and so demonstrate how poor (or good) the distributions are as a predictive tool. The manuscript would be strengthened with this additional information.

Minor Comments

It is unclear why the PyHillFit package was developed as these models can be easily fit with existing MCMC software such as JAGS or Stan. Furthermore, using existing MCMC software allows other models to be fit (e.g. 4 parameter logistic models -- when the upper and lower asymptote are not rescaled to 100% and 0%), and the sampling may be more efficient (especially with Stan, which handles hierarchical models well). The benefits of PyHillFit over these other tools should be highlighted.

The data are analysed with two models but the verbal description of the models is ambiguous. There are three models that could be run:
1. Parameters (IC50 and Hill) are common for all experiments (complete pooling, corresponding to the first model)

2. Parameters are different for all experiments; each experiment has its own parameter estimated independently from the other experiments (no pooling; not used in this paper)

3. Parameters are different for all experiments, but shared across experiments and thus mutually informative (partial pooling; the second model)

It is clear from the equations that their second model is number 3 above, but the verbal description is that of number 2. Please clarify the verbal description.

Other minor suggestions and corrections
- Figures 2 and 7 should have a shaded "x" node pointing into "y", as y depends not just on parameters but on other observed data.

- Figure 2, and Eq 7. Is this across all compounds and channels or separately for each compound and channel combination?

- Figure 10 might benefit from a line showing the prior distributions to allow easy comparison of prior and posterior.

- Figures 12 (A-C), 13 (A-C), 14 (A-C) would be clearer if the data points were plotted above the distribution curves

- Not clear how many cardiac action potential simulations are used to generate the distributions for Figure 15. Surely not the 375,000 samples from the posterior?

- Author contributions: RJH should be RHJ?

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Author Response 27 Feb 2017

Gary Mirams,

We would like to thank the reviewers for their thoughtful, insightful and careful review, we have uploaded a second version of the manuscript which we hope is improved in light of their comments.

> *The most interesting part of the paper was propagating the uncertainty though the action potential simulations. The authors conclude that this measure is insufficient to distinguish compounds in terms of risk (Fig 15), but if the distributions were plotted so that they are not on top of each other it may indeed be possible to distinguish the compounds. For example, "caterpillar plots", boxplots, or violin plots (e.g. Fig 2 in http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0096431) would better show the result.*
  - Thank you for the suggestion, we have updated Figure 15 to use a violin plot. This does allow a better inspection of the distributions and their overlap, as we can now see more clearly, there is still substantial overlap.

> *The authors also do not attempt to classify the compounds on the basis of their distribution (e.g. >50% of the distribution over a 10% prolongation of APD(90) to indicate risk of TdP) and so demonstrate how poor (or good) the distributions are as a predictive tool. The manuscript would be strengthened with this additional information.*
  - This is something we specifically tried to avoid, partly because it becomes a large exercise that detracts from the main software tool, and also not least because the risk of TdP is not well-defined or agreed upon for some of these compounds. Our point isn't that the distributions will help/hinder predictions per-se, rather you could easily be anywhere 'within' a distribution, so wherever they overlap there will necessarily be no way to reliably distinguish between compounds. If the 50% value happened to be a good classifier for these data, we would still not expect this to hold for new compounds, as the ordering could easily reverse when new data become available wherever the distributions overlap to this extent. Nevertheless, we annotated Figure 15 with risk classifications from the CredibleMeds database (a more credible data source than its name suggests!). As one might expect, only the clear shorteners and prolongers correspond well to 'no known risk' or 'known TdP risk'; the majority of compounds have overlapping distributions of moderate APD prolongation, that cannot really be ranked reliably.

> *It is unclear why the PyHillFit package was developed as these models can be easily fit with existing MCMC software such as JAGS or Stan. Furthermore, using existing MCMC software allows other models to be fit (e.g. 4 parameter logistic models -- when the upper and lower*

asymptote are not rescaled to 100% and 0%), and the sampling may be more efficient (especially with Stan, which handles hierarchical models well). The benefits of PyHillFit over these other tools should be highlighted.

- PyHillFit is intended for the specific case of ion channel screening data. We do appreciate that this could have been coded in Stan, but we wanted to have more 'control' over what was happening below the surface, and in how the results were outputted and analysed, so we wrote the algorithms ourselves in Python. Perhaps as much of our 'work' here was in deriving the statistical model and appropriate prior distributions, as it was in writing the software itself. If a 4 parameter agonist compound was identified then an extension in Stan to cover this might be appropriate.

> *The data are analysed with two models but the verbal description of the models is ambiguous...it is clear from the equations that their second model is number 3 above, but the verbal description is that of number 2. Please clarify the verbal description.*

- Text amended in section 4.1.

> *Other minor suggestions and corrections*

- We appreciate these being pointed out and have made the suggested amendments:

> *Figures 2 and 7 should have a shaded "x" node pointing into "y", as y depends not just on parameters but on other observed data.*

- Updated, thanks for the suggestion.

> *Figure 2, and Eq 7. Is this across all compounds and channels or separately for each compound and channel combination?*

- Separately, text updated.

> *Figure 10 might benefit from a line showing the prior distributions to allow easy comparison of prior and posterior.*

- Updated, thanks for the suggestion.

> *Figures 12 (A-C), 13 (A-C), 14 (A-C) would be clearer if the data points were plotted above the distribution curves*

- Updated, thanks for the suggestion.

> *Not clear how many cardiac action potential simulations are used to generate the distributions for Figure 15. Surely not the 375,000 samples from the posterior?*

- Now mentioned in the caption.

> *Author contributions: RJH should be RHJ?*

- Updated(!), well spotted!

Competing Interests: No competing interests were disclosed.