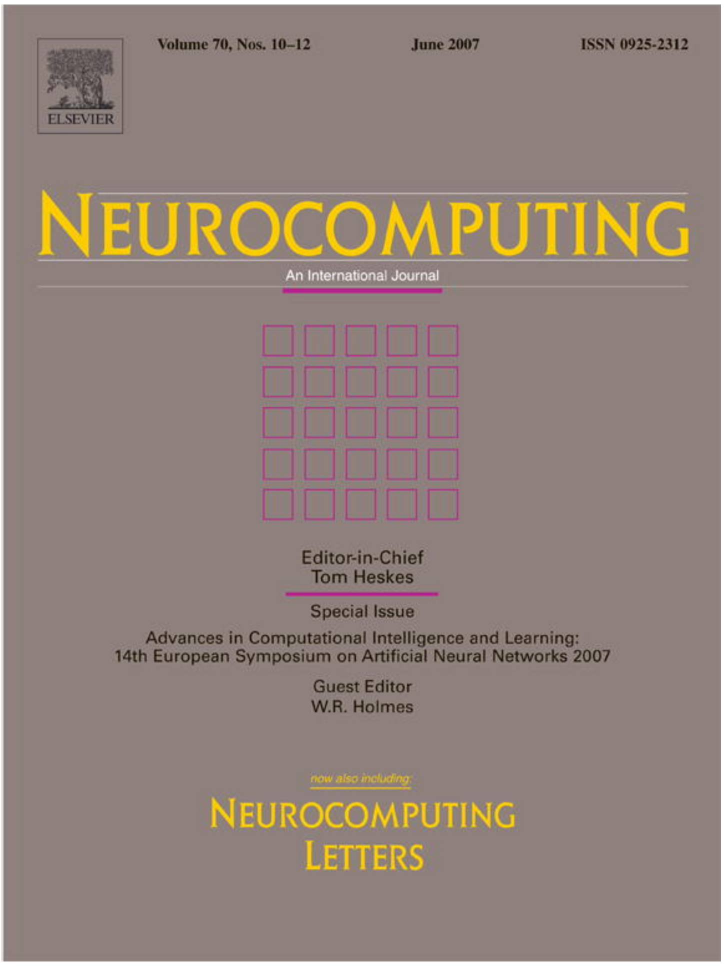

Volume 70, Nos. 10–12 June 2007 ISSN 0925-2312

ELSEVIER

NEUROCOMPUTING

An International Journal

Editor-in-Chief
Tom Heskes

Special Issue

Advances in Computational Intelligence and Learning:
14th European Symposium on Artificial Neural Networks 2007

Guest Editor
W.R. Holmes

now also including:

NEUROCOMPUTING
LETTERS

Available online at www.sciencedirect.com
ScienceDirect

Neurocomputing 70 (2007) 1892–1896

NEUROCOMPUTING

www.elsevier.com/locate/neucom

A means to an end: Validating models by fitting experimental data

M.D. Humphries*, K. Gurney

Adaptive Behaviour Research Group, Department of Psychology, University of Sheffield, UK

Available online 6 November 2006

Abstract

Validation of a computational model is often based on accurate replication of experimental data. Therefore, it is essential that modelers grasp the interpretations of that data, so that models are not incorrectly rejected or accepted. We discuss some model validation problems, and argue that consideration of the experimental design leading to the data is essential in guiding the design of the simulations of a given model. We advocate a "models-as-animals" protocol in which the number of animals and cells sampled in the original experiment are matched by the number of models simulated and artificial cells sampled. Examples are given to explain the underlying logic of this approach.

Keywords: Model validation; Experimental design; Network models

1. Introduction

Retrospective data-fitting is an effective diagnostic tool. If a computational model does *not* fit existing experimental data, then the model is probably inadequate. It is therefore essential that the modeler has an excellent grasp of the possible interpretations of that data, so that models are not incorrectly rejected or accepted. Data-fitting can be attempted at four distinct levels of accuracy: (L1) Trends: matching directional changes in an experimental variable across experimental conditions; (L2) Means: matching mean values of an experimental variable across experimental conditions; (L3) Distribution: matching the distribution function of an experimental variable across experimental conditions; (L4) Exact values: matching experimental variable values across experimental conditions. Note that each level is a stronger validation of a model than its predecessors, and necessarily subsumes them: for example, if the model fits the mean data values, it must also have captured the trends in that data.

For a (biologically-constrained) network of spiking neurons, levels L1, L2, and L3 would all seem obtainable, given experimental data on changes in statistical properties at the network level (by which we mean, for example, changes in population firing rates and patterns, as opposed to single-neuron statistics such as membrane potential changes). This paper discusses the issues facing the modeler in trying to achieve these levels of validation/replication.

The goal of *quantitatively* matching values of experimental variables could be attempted as follows: a single instantiation of the model could be simulated, the appropriate model parameters altered to mimic the experimental manipulations, followed by analysis of the outputs of *all* the cells in the model. However, it is often forgotten that many statistical properties of experimental data are dependent both on the number n of data-points collected, and the manner in which the data-points are collected. For example, neuroscience studies routinely pool cell recordings taken from different animals. Thus, if we are to compare the model and the data we need to replicate the data-gathering process too.

We therefore advocate a "models-as-animals" approach to replicating neuroscience experiments: matching both the number of data-points collected, and the number of animals used. The model is assumed to have some stochastic component so that different instantiations of the model are, in general, different from each other. If this holds, then each instantiation of the model is considered to represent a single animal. For each experimental condition, we instantiate the same number of models M as there were animals, and aim to collect the same number of

*Corresponding author.
E-mail address: m.d.humphries@sheffield.ac.uk (M.D. Humphries).

0925-2312/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2006.10.061

experimental observations n_d (in this case, cells) from each neural structure in that condition by sampling $n_\mathrm{m} = n_\mathrm{d}$ randomly-chosen cells across all M models (we use subscript d to indicate experimental data, subscript m to indicate model simulation data). In general, n_d will not be divisible by M and, if we collect equal samples from each model, we have at best $n_\mathrm{m} \approx n_\mathrm{d}$. However, it will usually be the case that $n_\mathrm{m} - n_\mathrm{d} \ll n_\mathrm{d}$ and the error in the approximation will be small.

We consider here a particular experimental data set and corresponding model to demonstrate the pertinent issues in data-fitting, highlighting simple and subtle pitfalls for the modeler. In the process, we shall see why we advocate "models-as-animals" as an approach to avoid these pitfalls, providing initial answers to the questions: why match the number of data-points? And, why match the number of animals?

2. Methods

The basal ganglia (BG) are a set of inter-connected structures of the fore- and mid-brain that receive massive input from many regions of cortex, and output to thalamus and brainstem targets. Our BG model comprises large populations of modified leaky-integrate-and-fire model neurons in each structure (examples presented here are from BG models with a total of $N_\mathrm{m} = 192$ neurons per structure). The model neurons themselves incorporate many details of their biological counterparts (e.g. Ca^{2+} dynamics, spontaneous firing, morphologically defined axon terminal locations). For our present purposes the model details do not matter (for a complete description see [2]) but a few key points are relevant. First, all parameters that could have an experimentally derived value do so. Second, the model has two stochastic components: neuron parameters are distributed around the experimentally derived mean, and neural connectivity, while constrained by the known anatomy of the major pathways, is probabilistic at the individual neuronal level. Third, the simulated cortical inputs to the model are matched to the patterns and statistical properties of cortical neuron firing recorded under the appropriate conditions.

Our target data is taken from a study by Magill et al. [5] who recorded from two components of the BG—sub thalamic nucleus (STN) and globus pallidus (GP). Here, we describe and replicate only a sub-set of their data for simplicity. They recorded from rats (under urethane anesthesia), and found that STN cell outputs had a slow-wave oscillation, seemingly tracking the cortical slow-wave activity (~1 Hz), but GP cells did not. Following lesion of almost all cortex, STN cell outputs had a significantly reduced mean firing rate and no longer oscillated at low frequencies, but GP cell output was unchanged. These results are remarkable because the anatomy of the BG suggests that STN and GP are tightly coupled in a negative feedback loop, yet their respective outputs do not seem to mirror changes in the other.

We matched the experiment structure and manipulations exactly, as described above. The M, n_d values for Magill et al. [5] are given in Fig. 1a. Two sets of simulations were carried out, one with and one without slow-wave cortical input, mimicking the two experimental conditions.

3. Results

3.1. Comparing the correct statistics

We consider here the matches to the reported changes in firing rates. Typically, single variable data that is continuous—such as firing rates—are summarized as arithmetic means $\bar{x}_\mathrm{d}$ and standard deviations s_d, and presented in tables or bar-charts with positive-only deviation bars (the "dynamite-plunger" plot). The arithmetic mean of each condition is the one we wish to match using the model. If we directly compare reported data and simulation data, the means do not match exactly but the simulation means do fall within one standard deviation of the data mean. However, this is not a correct interpretation of the reported statistics.

The experimental data mean is itself a sample mean of a very large underlying population (in this case, each rat has $\sim 13,600$ STN and $\sim 46,000$ GP cells, and there are multiple rats in each condition). Hence, a straight comparison of reported data to equivalent simulation data does not tell us a great deal. We should, instead, assess the confidence intervals on the means themselves, which are given by the standard error (estimated from a single sample distribution by $\mathrm{SE}(\bar{x}) = s/\sqrt{n}$). These intervals give the bounds on the actual population mean, since there is a probability of $1-\alpha$ of the population mean falling within $\bar{x} \pm t_{(\alpha,n-1)}\mathrm{SE}(\bar{x})$, where t is the value of the t-distribution for the significance level α, given the number of samples n. Typically, we plot 95% confidence intervals ($\alpha = 0.05$) of the data and model means against each other (Fig. 1a), and we can see that they overlap considerably in all cases.

We can quantitatively assess this overlap to find out whether or not the experimental and simulation means are representative of the same underlying population mean: this constitutes validation at level L2. One method is to compute independent group t-tests between the experimental and simulation data for each nucleus in each condition. A non-significant t value would predict that the mean firing rates of both simulated and real STN (or GP) were drawn from the same underlying population in that experimental condition. We compute t for all the comparisons made in Fig. 1. (Note that we cannot assume equal variances of the experimental and simulation data, because we have no means of testing this assumption in the absence of the full experimental data set. Thus, we must use the modified degrees-of-freedom form [7]: $\mathrm{d.f.} = (s_\mathrm{d}^2/n_\mathrm{d} + s_\mathrm{m}^2/n_\mathrm{m})^2/\,[(s_\mathrm{d}^2/n_\mathrm{d})^2/(n_\mathrm{d}-1) + (s_\mathrm{m}^2/n_\mathrm{m})^2/(n_\mathrm{m}-1)]$.) We find no significant differences (at $p = 0.05$) between the simulation and experimentally-derived mean firing rates, and

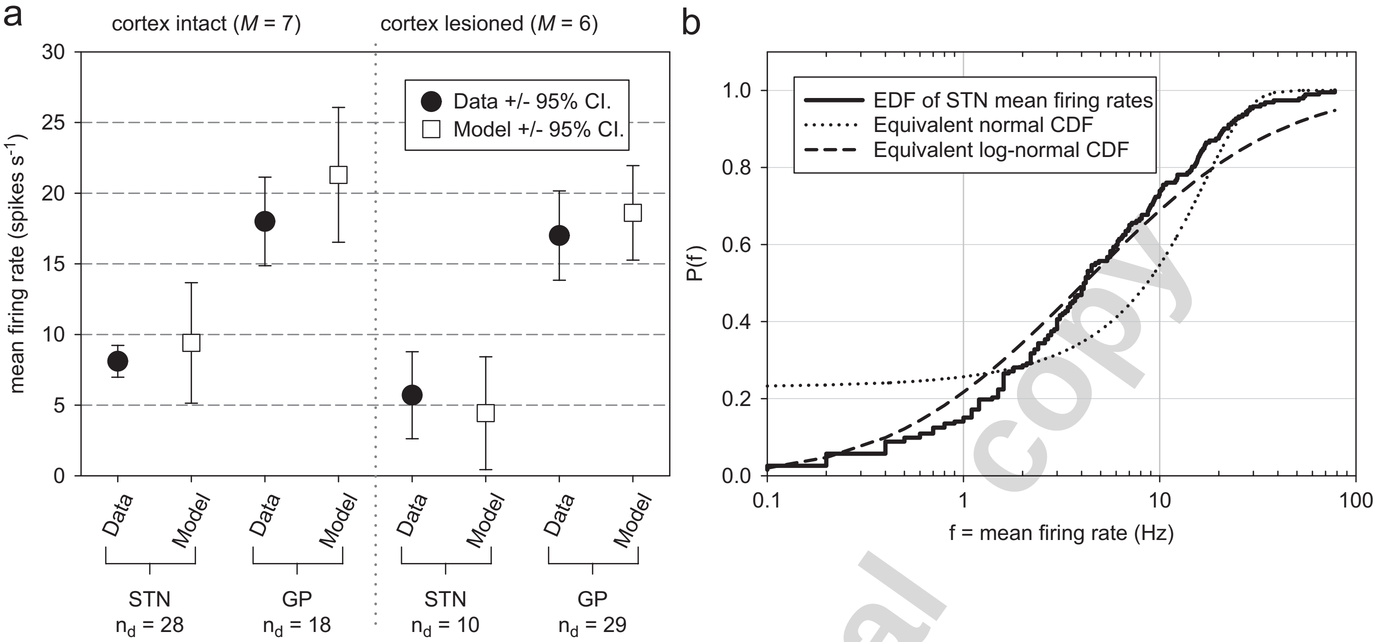

Fig. 1. Interpretation of mean firing rate data in the basal ganglia (specifically the STN and GP) from the study by Magill et al. [5]. (a) Experimental and model data showing means and 95% confidence intervals (CI) of the means: the overlapping confidence intervals are clearly shown. Experiment structure: M is the number of animals, n_d total number of cells sampled. (b) Assumptions of normality. Empirical distribution function (EDF) of the STN mean firing rates derived from all STN cells in a single model. The distribution is log-normal ($m = 1.41$ Hz, $S = 1.81$ Hz), rather than normal ($\bar{x} = 8.63$ Hz, $s = 11.69$ Hz).

could conclude that we have “fitted the data”, validating the model at L2.

Why match the number of data-points? Parametric tests of differences between populations—including the independent group *t*-test—are based on SE($\bar{x}$), which scales as a function of n. Therefore, reliable parametric comparisons of simulated and experimental data require the equivalent number of samples from model and data.

3.2. *Why assume normality?*

The underlying problem with the above approaches is that the use of arithmetic $\bar{x}$, s, and SE($\bar{x}$) assumes a normal distribution of the variable across the population as a whole. From the reported data of most neuroscience studies we are unable to test this assumption. This is important because only with access to the full experimental data set—here, the mean firing rate of every contributing STN and GP neuron—could we both test for normality and, if not present, do a non-parametric comparison between the simulation and experimental data (that is, a comparison that does not rely on an assumption that both underlying populations are normally-distributed, such as the Mann–Whitney U-test for independent groups).

Access to such detailed data can be difficult. But even if we could obtain it, what can it tell us about the assumption of normality? There are two issues here which have a bearing on why we advocate the “models-as-animals” approach: the underlying real population distribution, and the distribution of the experimental sample taken from it. We illustrate the first issue using the model: it turns out that, although the fits to the experimental mean firing rates are close (as they are for the firing patterns, data not shown), the underlying population distribution is not normal.

We look at the distribution of mean firing rates across all N_m cells from STN for a single model—a single representative “animal”. The chosen model is one in the first experimental condition (cortex-intact). We compute the empirical distribution function (EDF) by $P(x) =$ (number of observations$\leqslant x$)/(total number of observations). The equivalent normal cumulative distribution function (CDF) is calculated using the mean and standard deviation of the mean firing rate distributions. When plotted together, it is clear that the STN EDF deviates substantially from a normal distribution (Fig. 1b).

Numerous formal tests of normality (including Lilliefors—a Kolmogorov–Smirnov test specialized for normal distributions—chi-squared, Frossini) are available, but comparative studies have favored Anderson-Darling [6]. Using this test, we confirm the qualitative conclusion from the EDF–CDF comparison: the distribution of STN mean firing rates does significantly differ from a normal distribution ($A^2 = 17.49$, $p < 0.001$).

Biological data distributions are often log-normal [1,4], and power-law distributions are common place in real-world situations [3]. Indeed, the distribution of the log of the same STN mean firing rates does not significantly differ from a normal distribution ($A^2 = 0.59$, $p > 0.1$): it is, therefore, approximately log-normal (Fig. 1b). Thus, if the model STN accurately reflects the real STN, then we could not compute parametric statistics either on each of the experimental and simulated data set, or on comparisons between them.

3.3. Low power in neuroscience studies

So, turning to the second issue noted above, can we assess the experimental sample distribution? An inherent problem in neuroscience studies is the low n: the power in the sample is thus likely to be low, and therefore only large deviations from normality (or large magnitude changes due to experimental manipulations) are reliably detectable. The often low coefficient of variation ($\mathrm{CV} = s/\bar{x} < 0.2$) in distributions of biological data [1] implies that the magnitude of deviation from normality is small. Indeed, Gingerich [1] has shown that for randomly-generated log-normal data sets with $\mathrm{CV} = 0.1$, even the most powerful of normality tests requires a minimum of $n \simeq 1700$ to reliably reject them (at confidence levels of $\alpha = \beta = 0.05$). (For the single model data analyzed here, the CVs and n are comparatively high: STN $\mathrm{CV} = 1.36$, GP $\mathrm{CV} = 0.59$, and $n = N_\mathrm{m} = 192$, and hence the normality tests applied above are likely to be reliable.)

The low n of neuroscience studies thus means that the underlying distribution of the sample data cannot be confidently determined. In turn, this means that we cannot assess the fit between the simulated and experimental data distributions, as the experimental data distribution itself is unknown. Hence, validating network models by fitting distribution functions (L3) is difficult at the present time. (It awaits the advent of wide-spread multi-contact, multi-electrode, recording.)

3.4. The effect of pooling small samples from distributions

If we cannot assess the underlying distribution of the sample data, can we instead assess the expected deviation from the underlying distribution? That is, given both the small sample size and data-pooling across subjects in neuroscience studies, what can we say about the devia tion of a sample set's distribution from the underlying distribution?

The Central Limit Theorem tell us that, for a sufficiently large M, and assuming separate distributions with independent parameters in each of the M animals, then the sample distribution will converge on a normal distribution, regardless of the form of the contributing distributions. However, in most neuroscience studies M is not sufficiently large; and, for a given experiment, we cannot assume that the distributions are independent (as a trivial example, consider that a lesion would have similar effects in all animals, and hence distributions—of firing rates, etc.—in each animal would not be independent).

We carried out a series of Monte Carlo simulations to begin the assessment of the effect of pooling small samples from non-normal distributions. Following from the above results, we began by assuming a universal underlying log-normal firing rate distribution of $m^\mathrm{c} = 1.41\,\mathrm{Hz}$, and $S = 1.81\,\mathrm{Hz}$. For a given number of animals, we assess two cases: (1) that each animal has an identical underlying log-normal firing rate distribution of $L(m^\mathrm{c}, S)$; (2) that each animal has a distribution of $L(m^\mathrm{u}, S)$, with m^u sampled uniformly at random from the interval $[m^\mathrm{c} - m^\mathrm{c}/5, m^\mathrm{c} + m^\mathrm{c}/5]$.

For each case, we created a sample set by randomly sampling four cells (a typical value from neuroscience experiments) from the distributions of each animal. To assess the recovery of the underlying distribution, we computed the sample set's EDF, and calculated the root mean square error (RMSE) for the fit to the underlying log-normal distribution's CDF. To assess the recovery of the central tendency, we computed the maximum likelihood estimate of m from the sample set [4]. For a given number of animals, each case was repeated 1000 times.

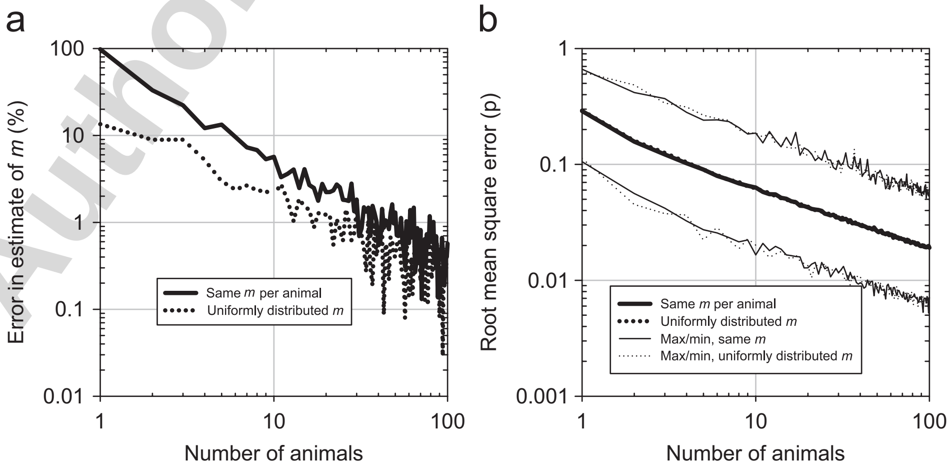

Fig. 2. The effect of experimental design on statistical properties of sampled data. (a) The error in recovering the log-normal location parameter m falls ~1/(number of animals), with error less than 1% as the number of animals exceeds ~40; (b) the mean error in the recovery of the whole log-normal distribution also falls ~1/(number of animals)—as did the maximum and minimum errors. The trend is identical for the two simulations.

Our results (Fig. 2) show that the mean error in recovery of both the overall distribution and central tendency fell monotonically, proportional to $\sim 1/(\text{number of animals})$. Moreover, the minimum and maximum distribution error fell in the same way. However, even with 100 animals, the maximum distribution error encountered was still ~6%. All these results were the same for both distribution cases we assessed. Thus, the error in the sample distribution systematically depends on the number of animals used, and is not negligible for typical values used in neuroscience experiments.

Why match the number of animals? The number of animals systematically affects the amount that the sample distribution differs from underlying distribution. Thus, if the model's underlying distribution is correct, then the same errors should result when following the same sampling protocol.

4. Discussion

We proposed that replicating experimental data requires a considered approach to simulation protocols that takes into account the number of observations n and animals M in the experiment(s): "models-as-animals". By taking the same n, we ensure that SE($\bar{x}$) is correctly scaled for the simulated data, if the model is accurate. Then, if para metric tests (*t*-tests, ANOVAs, and so on) can be applied, their results would not be distorted. By taking the same M, we ensure that the sampling process affects the experimental and simulation data in the same way: if the underlying population distribution is not normal, then the statistics of the resulting sampled populations will depend on the properties of that population distribution. Thus, we avoid the problem of incorrectly rejecting a model because the simulation data does not fit the experimental data due to distortions introduced in the sampling process (this does not mean that it is any easier to validate a model, just that validation is then made on grounds of accuracy of the model rather than spurious statistical effects). Moreover, we can consider each set of M simulations a single, virtual, experiment: repeating this experiment many times will thus show the limits of what the model can predict.

Consider the alternative approaches. With a single model, we have seen in Section 3.2 how analyzing every output from a structure could reveal distributions that are *not* reflected in the experimental data. Similarly, even if we sampled just n_d outputs from a single model, this is still an order-of-magnitude more than is actually sampled from a single animal, and the distributions are likely to differ for the same reasons outlined above. A further alternative, for either single or multiple models, is to match the relative percentage of sampled cells from each structure, rather than directly matching n_d. This has two problems. First, the outlined pitfalls remain: the distributions of simulated data may well differ from the experimental data simply as a result of a different sampling process, as $n_\mathrm{m} \neq n_\mathrm{d}$. Second, there is a methodological difficulty. Even for small neural structures, such as those (STN and GP) considered here, the percentage of sampled cells per animal is on the order of $10^{-4}-10^{-5}$% of the total number of cells in that structure. Thus, if relative percentages are matched, for even a single cell to be sampled from a model it must have on the order of 10^4-10^5 cells per structure.

Validating a computational neuroscience model by data-fitting is somewhat hampered by the structure, reporting, and low n of neuroscience experiments. The technical difficulties of such experiments make many of the problems inevitable, so the onus is on the modeler to be aware of the pitfalls and consider the solutions.

Acknowledgments

Funded by the EPSRC (GR/R95722/01) and the EU Framework 6 ICEA project.

References

[1] P.D. Gingerich, Statistical power of EDF tests of normality and the sample size required to distinguish geometric-normal (log-normal) from arithmetic-normal distributions of low variability, J. Theor. Biol. 173 (2) (1995) 125–136.

[2] M.D. Humphries, R.D. Stewart, K.N. Gurney, A physiologically plausible model of action selection and oscillatory activity in the basal ganglia. J. Neurosci. in press.

[3] E.F. Keller, Revisiting scale-free networks, Bioessays 27 (10) (2005) 1060–1068.

[4] E. Limpert, W.A. Stahel, M. Abbt, Log-normal distributions across the sciences: keys and clues, Bioscience 51 (2001) 341–352.

[5] P.J. Magill, J.P. Bolam, M.D. Bevan, Dopamine regulates the impact of the cerebral cortex on the subthalamic nucleus-globus pallidus network, Neurosci. 106 (2001) 313–330.

[6] M.A. Stephens, EDF statistics for goodness of fit and some comparisons, J. Am. Stat. Assoc. 69 (1974) 730–737.

[7] T.D.V. Swinscow, M.J. Campbell, Statistics at Square One, BMJ Publishing Group, 1997, ⟨http://bmj.bmjjournals.com/collections/statsbk/index.shtml⟩.

Mark Humphries received his Ph.D. in computational neuroscience from the University of Sheffield, and is currently a Postdoctoral Researcher. His research interests include action selection, bio-inspired robotics, computational modeling of vertebrate brain structures, and applications of network theory to neuroscience.

Kevin Gurney has degrees in mathematical physics, digital systems and a Ph.D. in neural networks. His postdoctoral work focused on visual psychophsyics and models of early visual processing. He is currently a Reader at the University of Sheffield where he is building computational models of selective processing for action and attention in the basal ganglia and associated neural circuits.